In this thesis we center the design and development of natural language technology around humans. We are motivated from two angles, roughly summarized as: (i) who are the users of these systems, and what do they want?, and (ii) how can we use our knowledge of human language processing and acquisition? We argue that a human-centered approach to NLP is essential to help us understand model behavior and capabilities, identify where and how modeling can be improved, and make sure models are in line with users' needs. Each of the chapters in this thesis is driven by one or more of these aspects.

New Directions in Human-Centered Language Technology

Maartje ter Hoeve

New Directions in Human-Centered Language Technology

Understanding and Improving NLP Models

Maartje ter Hoeve

New Directions in Human-Centered Language Technology

Understanding and Improving NLP Models

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties
ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 17 mei 2023, te 11:00 uur

door

Maartje Anne ter Hoeve

geboren te Apeldoorn

PROMOTIECOMMISSIE

Promotor:

prof. dr. M. de Rijke	Universiteit van Amsterdam

Copromotor:

dr. Y. Kiseleva	Microsoft Research

Overige leden:

prof. dr. R. Fernandez Rovira	Universiteit van Amsterdam
prof. dr. E. Kanoulas	Universiteit van Amsterdam
prof. dr. C. Monz	Universiteit van Amsterdam
prof. dr. B. Plank	Ludwig Maximilian University of Munich
dr. R. Sim	Microsoft Research

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The work described in this thesis has been primarily carried out at the Information Retrieval Lab of the University of Amsterdam and in part during internships at Microsoft Research, Facebook AI Research, and Apple Machine Learning Research. The research carried out at the University of Amsterdam was funded by the Amsterdamse Politie.

Printed by Ridderprint, The Netherlands.

ISBN: 978-94-6483-007-1

Copyright © 2023 by M.A. ter Hoeve, Amsterdam, The Netherlands.

ACKNOWLEDGEMENTS

That was it! I vividly remember my first day as a PhD student. I went to my office, opened my laptop, and wondered 'so, how does one do that, a PhD?'. Luckily, it turned out I did not have to do it all by myself, and I am thankful to everyone who played such an important role along the way.

First, of course, Maarten. Thank you for your supervision. You taught me how to become an independent researcher, and I am deeply thankful for that opportunity. I have experienced ILPS, and later IRLab, as a wonderful place to do research, and I am amazed by how you built the group. I will miss it.

Julia, thank you for being my co-supervisor and my friend. I admire your way of giving honest, but warm feedback. You are a great researcher, and I am grateful that I could learn from you.

Zeynep, thank you for working together during the first stages of my PhD. You provided me with a valuable different perspective, and I consider myself lucky that I was able to learn from you.

Barbara, Christof, Evangelos, Raquel and Rob, thank you for agreeing to be part of my PhD committee, and for your valuable time to read and discuss my thesis.

I am thankful for the unique experience I had being part of TROI. In particular, I would like to thank Adri, Arthur, Bob, Dominique, Gwennyn, Hilde, and Peter. The situation was exciting and new for all of us, and I truly appreciate your help to make it work so well.

I would also like to thank everyone who was part of ILPS or IRLab at some point along the way: Alessio, Ali, Ali, Amir, Ana, Andrew, Anna, Antonis, Arezoo, Artem, Barrie, Bob, Chang, Christophe, Chuan, Clara, Clemencia, Dan, David, Gabriel, Georgios, Hamid, Harrie, Hinda, Hongyu, Hosein, Ilias, Ilya, Jiahuan, Jie, Jin, Jingfen, Julia, Julien, Ke, Maarten M, Mahsa, Maria, Mariya, Marlies, Marzieh, Maurits, Ming, Mohammad, Mostafa, Mounia, Mozhdeh, Negin, Nikos, Olivier, Pablo, Peilei, Pengjie, Philipp, Pooya, Praveen, Rolf, Romain, Ruben, Ruqing, Sam, Sami, Sebastian, Shaojie (Thank you, buddy!), Shashank, Shubha, Spyretta, Svitlana, Thilina, Thong, Tom, Trond, Vaishali, Vera, Wanyu, Weija, Xinyi, Yangjun, Yifei, Yuanna (I'm so happy you joined our office!), Yuanxing, Zihan, and Ziming. Thank you for making the lab such a great place, whether it was through discussions about research or by sharing countless bitterballen platters during the Friday drinks.

Of course, I am especially thankful to Petra. Thank you for all your hard work for the group, the nice conversations, and moments of encouragement.

Ke and Marzieh, thank you for adopting me in the Tardis during the first months of my PhD. Ke, you told me that I should try to prove my ideas wrong as quickly as possible. Marzieh, on my first day, when I wondered whether I could go home now, you told me it was fine. I have tried.

Ana, I am grateful that we got to share all the ups, and the downs, of doing a PhD together. And even managed to write a paper together!

Maurits, we started in the master's and now we're here. It's been great to be able to continue to collaborate, share an office, and discuss life in general.

Mariya, thank you for all the fun and serious conversations, the ice skating, dinners and brunches, and for agreeing to be my paranymph!

Gabriel, Hamid, Harrie, Jasmijn, Mahsa, Mostafa, Nikos, Rolf, and Sarah, thank you for being so important for making my PhD experience more than just research. I cherish all the nice memories. And Rolf, thank you for being such an excellent Slurm master, and helping (saving) me all these times.

When we started with Inclusive AI, I could not have imagined what an impact it would have on my own journey as well. Thank you Andy, Daniel, Elise, Karen, Marco, Maxi, Melika, Pascal, Putri, Rianne, Sadaf, Tim, Urja, and all the other mentors and mentees.

I am also grateful for having had people in my surroundings who have fulfilled mentor roles for me as well. In particular, I would like to thank Anne and Titia. You have both been extremely important in my decision to do a PhD, and I am thankful for your insights and patient conversations.

I would also like to thank the master students who I supervised: David, David, Dennis, Henning, Joris, Justine, Michael, Mink, and Rahel, thank you for teaching me more than I ever taught you.

I am grateful that I was able to do a summer internship at Microsoft Research in Redmond in 2019. This experience continued to play a very important role during the rest of my PhD. Thank you Adam, Alex, Bahar, Elnaz, Julia, Michael, Rob, Ryen, and all the others in KTX.

In 2021 I did a summer internship at Facebook AI Research. It was a special year, in the midst of the pandemic. Despite everything being remote, from onboarding to offboarding, I still felt part of the team. Thank you Dieuwke, Emmanuel and Evgeny.

In 2022 I spent a summer in Paris, at Apple Machine Learning Research. Thank you Borja, Charlotte, David, Ersy, James, Marco, Michal, Natalie and Samy. I am excited for what comes next.

Of course, I would also like to thank my friends outside of PhD life: Bas, Djamila, Ferry, Floris, Iris, Laura, Liam, Lisanne, Jonas, Joost, Jörg, Julian, Marije, Marjon, Nathalie, Thijs, Ties, thanks for all the moments of fun, relaxation, and distraction.

Finally, I would like to thank my family for all their love and support.

I am deeply thankful for having had my grandfather as an early example of what it is like to do a PhD. Unfortunately, my defense came just too late.

I am thankful to my parents for giving me the confidence to learn new things. It turned out to be an invaluable lesson. Jaco, knowing that there is someone who doesn't need many words to understand is a feeling hard to put in words. Thank you.

Thomas, thank you for being my paranymph. During my defense, and at every step of the way.

Maartje
Amsterdam, January 2023

CONTENTS

1 INTRODUCTION

Imagine you were to start writing this introduction. How would you go about it? Probably, you would wonder about your readers. They should be able to easily follow along with the argumentation in this chapter, and next, with the argumentation in this thesis in general. Who are they? What storyline would work best for them? And do we know anything about how people generally read texts?

In that sense, writing this introduction is not much different from how we design and develop natural language technology in this thesis. We are motivated from two angles, roughly summarized as: (i) who are the users of these systems, and what do they want?, and (ii) how can we use our knowledge of human language processing and acquisition?

Both questions cover different aspects of human-centered natural language processing (NLP) — the main topic of this thesis. That is, throughout this thesis, we center the design and development of natural language technology around humans. We argue that a human-centered approach to NLP is essential to help us understand model behavior and capabilities, identify where and how modeling can be improved, and make sure models are in line with users' needs. Each of the chapters in this thesis is driven by one or more of these aspects.

We visit a diverse set of tasks: digital assistance and question-answering (QA), automatic text summarization, part-of-speech (POS) tagging, machine translation (MT), and language modeling. As we proceed, we find many new ways in how we can approach these tasks such that a wider range of users is taken into account. We also find that there are still many opportunities to more adequately model these approaches, despite significant progress in recent years (e.g., Vaswani et al., 2017; Devlin et al., 2019; Lewis et al., 2020; Brown et al., 2020; Ouyang et al., 2022). Hence, this thesis is the start of a variety of new research directions in human-centered NLP — we propose new tasks, data, and (evaluation) methodologies.

1.1 SCOPE AND RESEARCH QUESTIONS

We scope our investigations around five research questions, each of which we will answer in one of the chapters of this thesis. Here, we give a brief overview.

We start in the space of digital assistance and question-answering, and we specifically focus on a scenario in which users are writing and consuming documents. We call this type of assistance *document-centered assistance*. This is a new scenario, which intuitively differs from other types of question-answering, such as factoid QA (e.g., Rajpurkar et al., 2016; Rajpurkar et al., 2018). We expect users' information needs to be different. However, as this is a new scenario, we do not exactly know what users expect from this type of assistance. Therefore, we formulate our first research question as follows:

Research Question 1: *What does document-centered assistance look like, and how can we model it?*

To answer this research question, we first conduct a survey to explore the space of questions that people might pose in a document-centered scenario. Once we have a good understanding of the type of assistance that people would like to receive, we proceed to a larger data collection phase. We collect a human-labeled, English dataset with questions and answers in the context of document-centered assistance. Next, we proceed to a modeling step in which we aim to align models with the needs that users identified for the document-centered scenario. We show that earlier state-of-the-art models for question-answering obtain promising results in the document-centered scenario, but we also find that the gap compared to their performance on more standard question-answering tasks is still substantial. As such, this work also helps us understand the capabilities of question-answering models, and identify where these models can still be improved.

We now continue our investigation in the space of automatic text summarization. We are motivated by the observation that automatic summarization methods often optimize for automatic metrics like Rouge (Lin, 2004) and human evaluation metrics such as informativeness, fluency, succinctness and factuality (e.g., Lin, 2004; Nenkova and Passonneau, 2004; Paulus et al., 2018; Narayan et al., 2018b; Goodrich et al., 2019; Wang et al., 2020; Xie et al., 2021). Often, the users of summaries are not explicitly incorporated in the design process of automatic summarization methods, making it hard to judge whether these summaries are fully in line with users' needs. This motivates our next research question:

Research Question 2: *What makes a good and useful summary for users of automatically generated summaries?*

To answer this research question, we propose a survey methodology to investigate the needs of users of *pre-made* summaries, i.e., summaries that are written by someone else — which is also the category that automatically generated summaries belong to. Our survey can be used to identify users' wishes before designing and developing automatic summarization methods. Next, it is important to evaluate whether an implemented method indeed aligns well with users' needs. Therefore, we also propose an evaluation methodology to evaluate the *usefulness* of automatically generated summaries for users.

Our survey is easily adaptable to different user groups, and we choose university students as our first target group. We find that current automatic summarization methods are not always in line with participants' wishes for premade summaries. The majority of these methods aim to generate a summary of a few sentences long, in raw text format (e.g., See et al., 2017; Narayan et al., 2018b; Liu and Lapata, 2019; Lewis et al., 2020). However, a purely raw text summary is rather unpopular with participants in our survey. Instead, participants indicate a need for summaries with a variety of graphical elements, e.g., arrows or colored text. This finding inspires our next research question:

Research Question 3: *How can we fulfill users' request for summaries that include graphical elements?*

In answering this research question we are also motivated by our knowledge about human text understanding, summarized by the *given-new strategy* (Clark and Haviland, 1974; Haviland and Clark, 1974; Clark and Haviland, 1977). According to this strategy, humans read a text while attaching *new* information to already known, i.e., *given* information, when building a mental model of the text. We propose a task to build the summaries with graphical elements according to the given-new strategy. We use our evaluation methodology from the previous question to confirm that a critical mass of people finds our proposed summaries useful. Encouraged by these positive findings, we collect a human-labeled dataset to support research into the task, which we call GraphelSums. This dataset contains summaries with graphical elements for English news documents. Next, we propose baseline methods for the task of summarization with graphical elements, which show that the task is feasible, yet also challenging. That is, just like for our first research question, these experiments help us understand the challenges that our proposed solutions still face.

So far, our efforts have solely focused on English as a language, which limits our user-centered approach. We now shift our focus to languages that do not have as many easily accessible written resources available as there are for English. Research on these lower resource languages is often grounded in high-resource scenarios, potentially biasing the results on the low-resource languages. Inspired by this observation, we formulate our next research question as follows:

Research Question 4: *How are low-resource investigations in NLP biased by high-resource approaches?*

A prominent approach to study a low-resource scenario is by downsampling from a high-resource dataset to simulate a low-resource dataset. For this research question we investigate the validity of this approach, as we hypothesize that the obtained downsample can be a poor proxy of an actual low-resource dataset. Empirically, we focus on two well-known NLP tasks that are also popular in the low-resource domain: part-of-speech tagging and machine translation. We find that random downsampling indeed results in a biased view of how well systems for these tasks work in a low-resource scenario. The reason is twofold. On the one hand, high-resource datasets are typically higher in quality than low-resource datasets, for example in terms of vocabulary size. This positively affects the quality of the downsample, and the performance of models trained on these datasets. On the other hand, high-resource datasets are often less carefully created than low-resource datasets, and thus they can contain more noise. This negatively affects the performance of models trained on a downsampled version of these datasets.

For our final research question we take a less user-focused approach as we are fully inspired by human language acquisition, specifically in the context of language modeling. Although large language models perform increasingly well (e.g., Devlin et al., 2019; Brown et al., 2020; Rae et al., 2021), they are trained on large amounts of data and their training regime appears unnatural from the perspective of human language acquisition — humans clearly do not learn language while reading large amounts of text while predicting the next, or even masked words. Instead, human language learning is much more interactive in nature. This motivates us to explore ways in which interaction can play a role in artificial language modeling, and thus we formulate our next research question as follows:

Research Question 5: *How can we make artificial language modeling more human-like by taking a more interactive approach?*

We also refer to this interactive approach to language modeling as *interactive language modeling*. This research question is exploratory in nature. We first define the objective of interactive language modeling in more detail, after which we propose a road map to achieve this objective. We then take the first steps on this road map, showing the initial feasibility of the approach, and paving the way for taking the next steps on the road map in future work.

This concludes the overview of our research questions. In the next section we summarize the main contributions of this thesis.

1.2 MAIN CONTRIBUTIONS

We divide the contributions in this thesis into theoretical, empirical, and data contributions. Together, these contributions help us (i) understand model behavior or capabilities, (ii) identify where and how modeling can be improved, and (iii) ensure that models are in line with users' needs.

Theoretical Contributions

- We propose a re-usable survey design to investigate the needs of users' of automatically generated summaries ((ter Hoeve et al., 2022d); Chapter 3).
- We propose an evaluation methodology to evaluate the usefulness of automatically generated summaries for users in a feasible and comprehensive manner ((ter Hoeve et al., 2022d); Chapter 3).
- We propose a new task, *summarization with graphical elements* ((ter Hoeve et al., 2022c); Chapter 4).
- We define the objective of *interactive language modeling* ((ter Hoeve et al., 2021); Chapter 6).
- We propose a road map towards interactive language modeling ((ter Hoeve et al., 2021); Chapter 6).

Empirical Contributions

- We develop an understanding for what kinds of assistance people would like to receive in a document-centered scenario ((ter Hoeve et al., 2020); Chapter 2).
- We show that passage ranking and question-answering baselines perform promising in the document-centered scenario, but not yet as well as in other scenarios ((ter Hoeve et al., 2020); Chapter 2).
- We develop a thorough understanding of how automatic summarization methods can benefit users in the educational domain ((ter Hoeve et al., 2022d); Chapter 3).
- We show that baselines based on abstractive summarization and information extraction methods perform promising on the task of *summarization*

with graphical elements, but that there is still a lot of progress to make ((ter Hoeve et al., 2022c); Chapter 4).

- We show that downsampling from a high-resource dataset to simulate a low-resource dataset results in a biased view of how well systems trained on these datasets work, for two well-known NLP tasks: part-of-speech-tagging and machine translation ((ter Hoeve et al., 2022a); Chapter 5).

- We take the first steps on the road map towards interactive language modeling ((ter Hoeve et al., 2021); Chapter 6).

Data Contributions

- We provide a detailed exploration of a human-labeled dataset that contains (i) a collection of work-related documents, (ii) questions that people might pose about these documents, (iii) answers to these questions, and (iv) metadata indicating properties of the questions ((ter Hoeve et al., 2020); Chapter 2).

- We collect a human-labeled dataset that we call GraphelSums, to support research into the task of *summarization with graphical elements* ((ter Hoeve et al., 2022c); Chapter 4).

1.3 THESIS OVERVIEW

In this section we give an overview of the thesis, and provide some recommendations for reading directions. This thesis consists of seven chapters, of which you are currently reading the first. The next five chapters discuss each of the research questions that we discussed in Section 1.1 one by one. Each chapter is based on one paper (see Section 1.4 below), and can therefore be read independently. However, Chapter 4 is a direct follow-up to Chapter 3. We therefore advise reading these two chapters together. We conclude this thesis and outline future research directions in Chapter 7.

1.4 ORIGINS

The chapters in this thesis are based on the following papers:

- **Chapter 2**
  Maartje ter Hoeve, Robert Sim, Elnaz Nouri, Adam Fourney, Maarten de Rijke, and Ryen W. White. 2020. Conversations with Documents. An Exploration of Document-Centered Assistance. In *CHIIR '20: Conference on Human Information Interaction and Retrieval.*

  This work was done during an internship at Microsoft Research AI in 2019. RS and RW proposed the initial idea. MtH scoped the final idea, based on discussions with RS, EN, AF, and RW. MtH designed the survey and the data collection pipeline, and ran the experiments. RS further helped with the survey and data collection design and running experiments. RS, EN, AF, and RW had important advisory roles. All authors contributed to the writing. MtH did most of the writing.

- **Chapter 3**
  Maartje ter Hoeve, Julia Kiseleva, and Maarten de Rijke. 2022. What Makes a Good and Useful Summary? Incorporating Users in Automatic Summarization Research. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

  MtH proposed the idea, and designed, ran and analyzed the survey. JK and MdR had important advisory roles. All authors contributed to the writing. MtH did most of the writing.

- **Chapter 4**
  Maartje ter Hoeve, Julia Kiseleva, and Maarten de Rijke. 2022. Automatic Summarization with Graphical Elements. Incorporating User Preferences in Automatic Summarization Research. *Under Submission.*

  MtH proposed the idea, designed the human evaluation, the data collection, and ran the experiments. JK and MdR had important advisory roles. All authors contributed to the writing. MtH did most of the writing.

- **Chapter 5**
  Maartje ter Hoeve, David Grangier, and Natalie Schluter. 2022. High-Resource Methodological Bias for Low-Resource Investigations. *Under Submission*.

  This work was done during an internship at Apple Machine Learning Research in 2022. NS proposed the initial idea. MtH scoped the final idea, based on discussions with NS and DG. MtH ran the experiments. DG helped with the experimental design of the MT experiments. All authors contributed to the writing. MtH did most of the writing.

- **Chapter 6**
  Maartje ter Hoeve, Evgeny Kharitonov, Dieuwke Hupkes, and Emmanuel Dupoux. 2022. Towards Interactive Language Modeling. In *ACL, Workshop on Semiparametric Methods in NLP & NeurIPS, Second Workshop on Interactive Learning for Natural Language Processing*.

  This work was done during an internship at Facebook AI Research in 2021. ED proposed the initial idea. MtH scoped the final idea, based on discussions with ED, DH and EK. MtH ran the experiments. EK helped with the experimental design. ED, DH and EK had important advisory roles. All authors contributed to the writing. MtH did most of the writing.

The writing of this thesis also benefited from work on the following publications:

- Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, **Maartje ter Hoeve**, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, Arthur Szlam, Yuxuan Sun, Katja Hofmann, Marc-Alexandre Côté, Ahmed Awadallah, Linar Abdrazakov, Igor Churin, Putra Manggala, Michiel van der Meer, and Taewoon Kim. 2022. Interactive Grounded Language Understanding in a Collaborative Environment: IGLU 2021. In *Proceedings of Machine Learning Research, NeurIPS 2021 Competitions and Demonstrations Track.*

- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, [...], **Maartje ter Hoeve**, [...], Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *Under Submission*.

- Julia Kiseleva, Alexey Skrynnik, Artem Zholus, Shrestha Mohanty, Negar Arabzadeh, Marc-Alexandre Côté, Mohammad Aliannejadi, Milagro Teruel, Ziming Li, Mikhail Burtsev, **Maartje ter Hoeve**, Zoya Volovikova, Aleksandr Panov, Yuxuan Sun, Kavya Srinet, Arthur Szlam, and Ahmed Awadallah. IGLU 2022: Interactive Grounded Language Understanding in a Collaborative Environment at NeurIPS 2022. In *NeurIPS, Competition Track*.

- Ana Lucic, **Maartje ter Hoeve**, Gabriele Tolemei, Maarten de Rijke, and Fabrizio Silvestri. 2022. CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks. In *International Conference on Artificial Intelligence and Statistics*.

- David Stap, Maurits Bleeker, Sarah Ibrahimi, and **Maartje ter Hoeve**. 2020. Conditional Image Generation and Manipulation for User-Specified Content. In *CVPR, AI for Content Creation Workshop*.

- Joris Baan, **Maartje ter Hoeve**, Marlies van der Wees, Anne Schuth, and Maarten de Rijke. 2019. Do Transformer Attention Heads Provide Transparency in Abstractive Summarization? In *SIGIR, Workshop FACTS-IR*.

2

CONVERSATIONS WITH DOCUMENTS

As the first step of our investigations we examine digital assistance and question-answering (QA). We are motivated by the observation that the role of conversational assistants has become more prevalent in helping people increase their productivity. Document-centered assistance, for example to help an individual quickly review a document, has seen less significant progress, even though it has the potential to tremendously increase a user's productivity. This type of document-centered assistance is the focus of this chapter.[1] An important goal of this thesis is to design NLP models such that they are in line with users' needs. This goal is explicitly part of this chapter, as we answer the first research question of this thesis:

Research Question 1: *What does document-centered assistance look like, and how can we model it?*

To answer this research question we first present a survey to understand the space of document-centered assistance and the capabilities people expect in this scenario. We also investigate the types of queries that users will pose while seeking assistance with documents, and show that document-centered questions form the majority of these queries. After a larger data collection phase, we present a set of initial machine learned models that show that (i) we can accurately detect document-centered questions, and (ii) we can build reasonably accurate models for answering such questions. These positive results

[1] This chapter is based on (ter Hoeve et al., 2020).

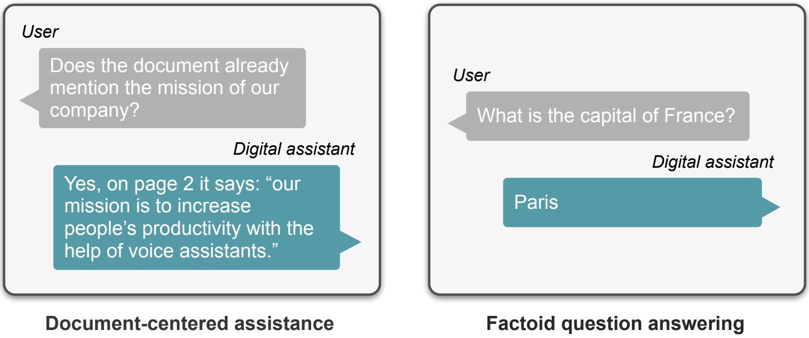

User
Does the document already mention the mission of our company?
Digital assistant
Yes, on page 2 it says: "our mission is to increase people's productivity with the help of voice assistants."
Document-centered assistance
User
What is the capital of France?
Digital assistant
Paris
Factoid question answering

Figure 2.1: An example of document-centered assistance (left) vs. factoid question-answering (right).

are encouraging, and suggest that even greater results may be attained with continued study of this interesting and new problem space.

2.1 INTRODUCTION

Digital assistants are used extensively to help people increase their productivity (Microsoft, 2019). A person can rely on their voice assistant, such as Amazon Alexa, Microsoft Cortana, or Google Assistant, to set an alarm while cooking, to play some music in the background, and to do a web search on a recipe's ingredients. Conversational interaction is also playing an increasingly important role in helping people to increase their productivity for work-related tasks (Tsai, 2018).

One area of interest that has not seen significant progress is document-centered assistance. Consider the following example: a person is driving to a crucial business meeting to prepare for a day with potential investors. The person is co-authoring a document about their company that will be provided to its investors, and it will be finalized in the upcoming business meeting. To be optimally prepared for the meeting, the individual wants to review what is already in the document. Since they are driving, they do not have direct access to the document, so they call their conversational assistant. The assistant has access to the document and can answer any query related to the document. The driver might pose queries such as *"does the document mention the mission of* *our company?"* or *"summarize what it says about our growth in the last two years."* — queries that help them understand what is already outlined in the document and what they still have to add to finalize the document. At the same time, the driver is unlikely to ask factoid questions, such as *"who is the CEO of our company?,"* given that they are already familiar with the organization.

In fact, previous work in the context of email and web search has shown that people's information needs are different when they are a co-owner of a document than when they are not (Ai et al., 2017). We hypothesize a similar difference in information needs in the context of document-assistance, motivated by the given example. This implies that document-centered assistance should critically differ from existing question-answering (QA) systems, which are mostly trained to give short answers to factoid questions (e.g., Rajpurkar et al., 2018; Reddy et al., 2019). Figure 2.1 gives an example of this difference. Document-centered assistance would also differ from non goal-oriented "chit-chat" scenarios (e.g., Sankar and Ravi, 2018; Yan and Zhao, 2018) — in our document-centered scenario, people have very clear information needs.

In this chapter, we investigate this space of document-centered assistance. This is an important task, since good document-centered assistance has the potential to significantly increase a person's productivity. We specifically focus on text consumption and document comprehension scenarios in a work context. We answer the first research question of this thesis by dividing it into three subquestions that we address one by one throughout this chapter:

Research Question 1.1: *What kinds of conversational assistance would people like to receive in a document consumption scenario?*

Research Question 1.2: *What kinds of queries might people use to receive this assistance when conversing with a document-aware assistant?*

Research Question 1.3: *How well do initial baseline models perform in a document-centered scenario?*

With this work we contribute:

- An understanding of assistant capabilities that are important to enable the document consumption scenario;

- Insights into the types of questions people may ask in the context of document-centered assistance;
- A detailed exploration of a human-annotated dataset with: (i) a collection of work-related documents, (ii) questions a person might ask about the documents, given some limited context, (iii) potential answers to the questions as represented by text spans in the document, (iv) additional metadata indicating some properties of the questions (for instance, it is a yes/no closed question, or the question is unanswerable given the document);
- Baseline experiments applied to the dataset exploring ways to handle document-centered questions.

Our research consists of three steps. We first perform a survey to answer Research Question 1.1 and Research Question 1.2 (Section 2.3), then we proceed with a data collection step, outlined in Section 2.4, and finally we answer Research Question 1.3 in Section 2.5.

2.2 RELATED WORK

The work in this chapter is related to two broad strands of research. In the first part of this section we look into voice controlled document narration and natural language interactions with productivity software, which is relevant to the first step of our research, the survey. Our initial modeling steps focus on single-turn conversations, and so we conclude this section with work on question-answering.

2.2.1 *Voice-Controlled Document Narration*

Document-centric assistance in the context of text consumption is related to prior work that explores adding voice interactions to screen readers. Screen readers are accessibility tools that narrate the contents of screens and documents to people who are blind, or who have low-vision. In this space, Ashok et al. (2015) implemented CaptiSpeak — a voice-enabled screen reader that maps utterances to screen reader commands and navigation modes (e.g., "read the next heading", "click the submit button"). More recently, Vtyurina et al. (2019) developed VERSE, a system that adds screen reader-like capabilities into a more contemporary virtual assistant. VERSE leverages a general knowledge-base to answer factoid questions (e.g., "what is the capital of Washington"), but then differentiates itself by allowing users to navigate documents through voice (e.g., "open the article and read the section headings"). An evaluation with 12 people who are blind found that VERSE meaningfully extended the capabilities of virtual assistants, but that the QA and document navigation capabilities were too disjoint — participants expressed a strong interest in being able to ask questions about the retrieved documents. This strongly motivates the research presented in this chapter.

2.2.2 *Interactions with Productivity Software*

There is an increasing interest in how people use different devices for their work-related tasks (e.g., Karlson et al., 2010; Jokela et al., 2015; Di Geronimo et al., 2016; Williams et al., 2019). Martelaro et al. (2019) show that in-car assistants can help users to be more productive while commuting, yet in easy, non-distracting traffic scenarios. While digital assistance in cars is a recent development (e.g., Lo and Green, 2013), natural-language interfaces have existed for much longer in more traditional work scenarios; for example the search box in products such as Microsoft Office and Adobe Photoshop. Bota et al. (2018) research search behavior in productivity software, specifically in Microsoft Office, and characterize the most used search commands. Fourney and Dumais (2016) investigate different types of queries users pose to a conversational assistant. Specifically, they focus on *semi implicit system queries* and *fully implicit system queries*. They show that different types of queries can be reliably detected and that forms of query alteration can boost retrieval performance.

2.2.3 *Question-Answering*

Question-answering is the task of finding an answer to a question, given some context. A lot of progress has been made in the area, driven by the successful application of deep learning architectures and the increase of large-scale datasets (e.g., Yang et al., 2015; Nguyen et al., 2016; Rajpurkar et al., 2016; Joshi et al., 2017; Dunn et al., 2017; Trischler et al., 2017; Kočiský et al., 2018;

Rajpurkar et al., 2018; Yang et al., 2018; Kwiatkowski et al., 2019). Although these datasets are all unique, they mostly contain factoid questions that can be answered by short answer spans of only a few words. In addition, none of them contain queries that reference the document directly as the subject of the query, a distinction that can cause existing QA models to yield irrelevant or confusing responses in the context of document-centered assistance.

Considerable research has targeted neural QA (e.g., Bordes et al., 2015; Chen et al., 2017; Dehghani et al., 2019; Gan and Ng, 2019; Kratzwald et al., 2019). Recently, Devlin et al. (2019) introduced BERT, or Bidirectional Encoder Representations from Transformers. BERT is a language representation model that is pre-trained to learn deep bidirectional representations from text. A pre-trained BERT model can be fine-tuned on a specific task by adding an additional output layer. BERT has made a tremendous impact in many NLP tasks, including QA. In this chapter, we use BERT for the baseline models.

Some QA work has focused specifically on the low-resource setting that we are also interested in in this work. Various approaches have been applied to augment small datasets to achieve good performance on language tasks (e.g., Yang et al., 2018; Daniel et al., 2019; Gan and Ng, 2019; Lewis et al., 2019). In order to accommodate our low-resource scenario, the data we have collected is supplemented with publicly available QA datasets.

All the work cited above plays a role in setting context for our scenario. With the possible exception of VERSE, none have specifically explored how people might want to receive conversation-based assistance with documents, and in particular documents that they have rich context about. In the next section, we explore what features and queries users are most likely to pose to their assistant when a document is the focus of the conversation.

2.3 STEP 1 – SURVEY

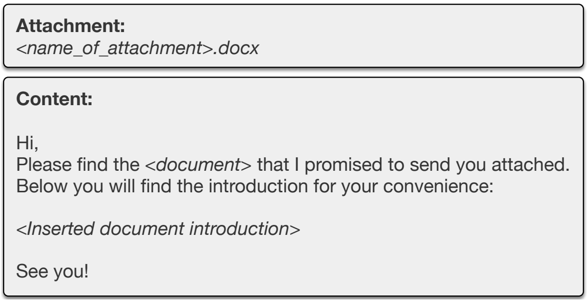
In the first step of our research, we aim to answer Research Question 1.1 (*What kinds of conversational assistance would people like to receive in a document consumption scenario?*), and Research Question 1.2 (*What kinds of queries might people use to receive this assistance when conversing with a document-aware assistant?*). To do so, we conduct a survey to explore the space of queries that people might pose when communicating with a voice assistant about a document, while not having full access to this document. We focus on a consumption scenario while on the go (i.e., limited primarily to voice and some touch input/output). Specifically, participants in our survey are presented with the following scenario: *"You are on your way to a business meeting. To help you prepare, your manager has sent you an email with a document attached. The objective of the meeting is to finalize this document, so that it can be shared with the rest of the organization. Your manager's email also includes the introduction of the document. You have been able to read this introduction, so you have an idea what to expect. You have not read the full document yet, but you can assume the document is approximately 6 pages long. On your way to the business meeting you do not have time to access the document, but you do have your smartphone equipped with a voice assistant like Alexa, Google Assistant, or Cortana. The voice assistant can help you navigate and understand what is written in the document, so that you will arrive prepared at your meeting. The voice assistant can answer your questions via audio or by displaying information on your smartphone screen."*

2.3.1 Survey Overview

Our survey consisted of two parts, corresponding to Research Question 1.1 and Research Question 1.2. In the first part, our primary goal was to explore three subquestions:

- Do users recognize the outlined scenario as relevant to their daily lives?
- Would users find voice assistance in the outlined scenario helpful?
- What range of features are important to users in a voice-first document consumption scenario?

Having identified the range of functionalities that a document-centered conversation might cover, in part two of our survey we aimed to gain a better understanding of the types of questions users might ask. Therefore, we collected questions that are grounded in specific documents. To this end, participants were primed with the same scenario as in the first part. The scenario is simulated by presenting them with an email that mimicked the email they received from their manager while on the go. The email contained the document introduction as a means to give them context about a specific document, to ensure that participants were able to ask informed questions, yet did not



Attachment:
<name_of_attachment>.docx

Content:

Hi,
Please find the *<document>* that I promised to send you attached.
Below you will find the introduction for your convenience:

<Inserted document introduction>

See you!

Figure 2.2: Sample email used to inform participants.

have full knowledge about what is written in the document. Figure 2.2 shows an example of an email provided to participants.

2.3.2 Participants

Our task was performed by 23 participants in a judging environment comparable to Amazon Mechanical Turk.[2] Participants were all English speaking and U.S.-based. Participants were paid at an hourly rate, removing the incentive to rush responses. We did set a maximum time of ten minutes per document.

Instructions given to participants

Before the task, participants were provided with detailed guidelines of the task and trained to follow them. In these guidelines, we explicitly encouraged participants to ask questions that were document-centered, i.e., to closely keep the outlined scenario in mind when asking questions. The participants were instructed to avoid questions that might be posed about any document, and answered using more mechanical solutions (e.g., *who is the author?*, *how many pages?*), and steered towards a scenario where they imagined having some familiarity with the document subject. Although we acknowledge that these more general questions are highly relevant, we argue that we do not need many sample questions of this type to fully understand the space of potentially relevant mechanical questions. Note that in the first part of the survey we investigated what participants would find the most and least important features in the outlined scenario, and this gives them the opportunity to select more

[2] `https://www.mturk.com/`

mechanical features. Participants were explicitly told to imagine their ideal voice assistant and to not limit themselves by any prior assumptions about the capabilities of currently existing voice assistants.

Participant training

Participants performed two training rounds, after which we provided them with feedback on their constructed questions in part two. This way we aimed to ensure that participants understood the task and devised high quality responses.

2.3.3 Document Selection

We selected 20 documents from a larger data set of 615 documents in Microsoft Word format. These documents were retrieved from a broad crawl of the web and meet the requirements that they are written in English and can be easily summarized. This last requirement, which was manually verified, ensures that we have a high quality dataset where noisy documents such as online forms are excluded. We selected the 20 documents from this set based on a number of requirements:

- The document should contain a clear introduction;
- The document should be between 3 and 10 pages long;
- The topic of the document should be understandable for non-experts on this topic and should not be offensive to anyone.

Table 2.1 gives more details on the nature of the selected documents. In addition to these 20 documents, we chose another two documents with which to train the participants. Although slightly deviating from the co-ownership scenario, providing users with documents ourselves allowed us to collect data in a more structured way, which we can use for the remaining research questions at a later stage. In the second part of the survey, the question collection round, each participant was asked to pose five questions about a given document. We required 20 judges per document. Since we have 20 documents we acquire 400 human intelligence tasks ("HITs"), resulting in 2000 questions.

Table 2.1: Categories of selected documents (20 in total) and their frequency in the survey distributed to participants.

Document category	Document count
Report	3
Job application	3
Description of a service	3
General description	2
Guidelines	3
Policy	3
Informative / Fact sheet	3

2.3.4 Survey Results

In this section, we provide the precise formulation of our survey questions, as well as the participants' responses to these questions.

Part 1 – Survey Questions

1. *Do you recognize the outlined scenario (i.e., needing to quickly catch up on a document while on the go) or some variation of it as something you experience in your daily life?*

   22 out of 23 participants indicated that they recognized the scenario.

2. *Do you expect to find it helpful if a voice assistant helps you to quickly familiarize yourself with the document in the outlined scenario?*

   22 out of 23 participants indicated that they would find this helpful.

3. *From the list below, choose three capabilities that you would find* ***most useful*** *in a voice-powered AI assistant to help prepare you for the meeting.*

   Participants could choose from the capabilities listed in Table 2.2. We randomized the order in which the features were presented, to avoid position bias. Note that the prompt specifically references the consumption scenario that participants are primed to consider. The results are given in Figure 2.3. Please refer to Table 2.2 to match the abbreviation on the x-axis with the feature description.

4. *From the list below, choose three capabilities that you would find* ***least useful*** *in a voice-powered AI assistant to help prepare you for the meeting.*

   Again, participants could choose from the capabilities in Table 2.2 and again this list is randomized for each participant. Figure 2.3b shows the results for this question. Comparing the results in Figure 2.3a and Figure 2.3b shows that participants are very consistent in the capabilities they find most and least useful.

5. *Can you think of any other features that you would like the voice assistant to be capable of? Please describe.*

   We divided the participants' answers into "mechanical" and "overview" features. A sample of the answers is presented below.

   Mechanical features:

   - "Voice recognition to unlock phone"
   - "Automatic spelling and grammar check"
   - "Remind me where I stopped when reading"
   - "The ability to link another app, such as maps or notes to the document directly"
   - "Bookmarking specific sections for future reference"
   - "Another useful feature would be the ability to add highlighted text to multiple programs simultaneously such as email notes and any other app"
   - "The Assistant should be able to turn tracked changes on and off and accept/reject changes and clean up a document and finalize"

   Overview features:

   - "Give bullet points of main topics"
   - "Give information about key points"
   - "Just highlight key points, summarize document"
   - "I would like for the voice assistant to be able to pick out the main points and read them out to me via voice output"
   - "If the assistant was able to give a synopsis then ask 1 or 2 questions to be sure the user understands the info"

Table 2.2: Assistant capabilities suggested to participants and judged for their utility. Abbreviations were never shown to users and are only used to map plots in this chapter to the corresponding capability.

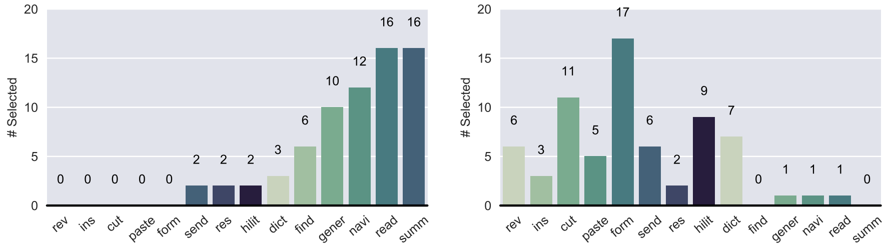
Abbr.	Capability
cut	Cut content from the document using voice
dict	Dictate input to the document
find	Find specific text in the document using voice input
form	Change text formatting using voice
gener	Respond to general questions about the document content, using voice input and output
hilit	Highlight text using voice
ins	Insert new comments into the document using voice
navi	Navigate to a specific section in the document using voice input
paste	Paste content from the device clipboard using voice
read	Read out the document, or parts of it, using voice output
res	Respond to existing comments in the document using voice
rev	Revise a section of text using voice input
send	Send or share a section of text using voice input
sum	Summarize the document, or parts of it, using voice output

Part 2 – Collecting Questions

Here we present the results of the second part of the survey, in which participants were prompted to generate questions about a document. Recall that the participants were only shown the document introduction or preamble and did not have visibility into the full document text.

6. *Please ask five questions to your voice assistant that would help you understand what is written in the document.*

We can divide participants' answers into a hierarchy of question categories. Note that the responses can be both questions and directives (e.g., *"go to Section X"*). Since the vast majority of the collected responses are questions, for brevity we refer to both of these response types as *questions*. Figure 2.4 shows the hierarchy. It was developed by sampling a set of participants' questions, which an expert studied and categorized. Three experts then reviewed all questions and categorized them according to the proposed taxonomy. By reviewing where



(a) Responses to question 3 – most useful assistant capabilities.

(b) Responses to question 4 – least useful assistant capabilities.

Figure 2.3: Most and least useful assistant capabilities; names explained in Table 2.2. On the y-axis: the number of times this particular capability was selected by participants (max = 23).

the experts disagreed, some minor adjustments were made to the hierarchy to arrive at the final one shown here. Level 1 of the hierarchy corresponds to how the question can be best responded to, or what kind of system or model would be suited best to handle the questions. Because document-centered questions are the main interest of our current research, we divide those into another set of categories, describing the intents of users on this level in more detail. This is level 2. We also subdivide the yes / no questions into the rest of the categories of level 2 and call this level 3. We do this because it is questionable whether a person would really be satisfied with a simple "*yes*" or "*no*" in response. We describe the question types in Table 2.3, and also provide verbatim examples sourced from the participants' responses. Figure 2.5a shows the distribution of question categorizations on level 1. Document-centered questions form the largest category of the questions. Recall that participants had to ask 5 questions per document; we investigated whether these questions differed in type. E.g., did participants ask mechanical questions first ("*bring me to Section 2.*") and then a document-centered question ("*what does it say there about X?*")? We did not find such a difference. We also investigated whether the type of document (Table 2.1) was an indication for the types of questions that were asked, but we found no difference between document types. The user was a strong indication for the type of question that was asked, indicating varying interpretations of the outlined scenario. Some users ask only factoid questions, some users only ask document-centered questions and only a few ask a mixture of all question types.

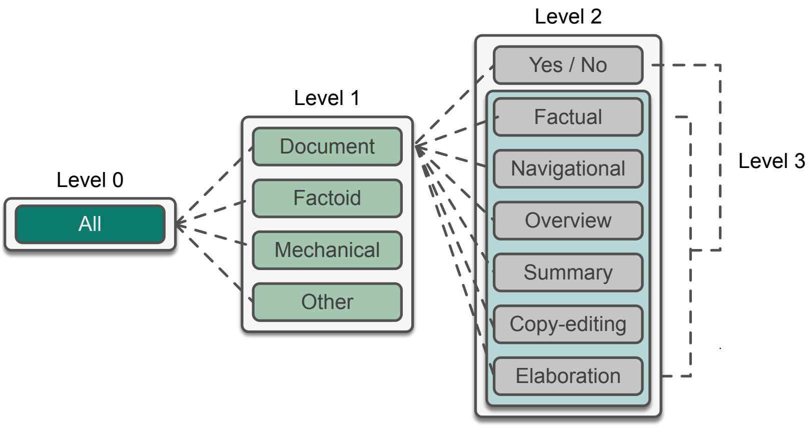


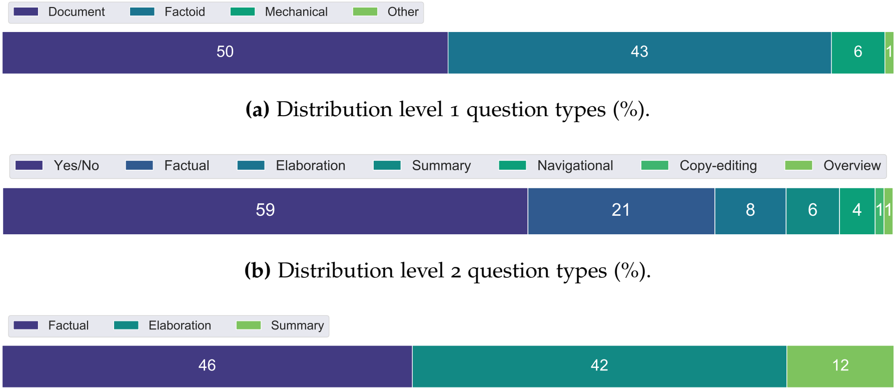
Figure 2.4: Question hierarchy.

The division of category labels for level 2 is shown in Figure 2.5b. As can be seen, the majority of questions are closed form yes / no questions. Figure 2.5c shows how these questions were categorized on level 3, yielding only 3 copy-editing questions, 2 overview questions, and 1 navigational question, rounding down to 0% in Figure 2.5c.

Table 2.3: Question type descriptions and examples.

Level	Question type	Examples
L1	**Document:** These are document-centered questions. That is, the question's phrasing explicitly or implicitly references the document. When asking such a question, a user is not looking for encyclopedic knowledge, yet rather for assistance that can help them to author the document. These types of questions are not present in existing QA datasets.	Does the document have specifications to the type of activity and sector improvement that will be offered?

	Factoid: Fact-oriented question that co-owners of a document are unlikely to ask. Answers are often only a few words long. Existing QA datasets cover these types of questions very well.	What is the date of the festival?
	Mechanical: Questions that can be answered with simple rule-based systems.	Highlight "Capability workers"
	Other: Questions that fall outside the above categories.	Read the email to me.
L2	**Yes / No:** Closed form (can be answered with 'yes' or 'no').	Does the document state who is teaching the course?
	Factual: Questions that can be answered by returning a short statement or span extracted from the document.	Where does the document state study was done?
	Navigational: Referring to position(s) in the document.	Go to policies and priorities in the doc.
	Overview: Questions that refer to the aim of the document.	What is the overall focus of the article?
	Summary: Questions that ask for a summary of the document or of a particular part of the document.	Find and summarize coaching principles in the document.
	Copy-editing: Questions when editing a document. They require a good understanding of the document to answer.	Highlight text related to application of epidemiologic principles in the document
	Elaboration: Questions that require complex reasoning and often involve a longer response.	Please detail the process to get access to grant funds prior to confirmation.



(a) Distribution level 1 question types (%).

(b) Distribution level 2 question types (%).

(c) Distribution level 3 question types (%).

Figure 2.5: Distribution of question types per hierarchical level. (Best viewed in color.)

Table 2.4: Question type classification results. Mean accuracy and variance after 5-fold cross validation.

Level 1	**Level 2**	**Level 3**
$0.92\ (\pm 8.6e^{-5})$	$0.90\ (\pm 1.3e^{-4})$	$0.67\ (\pm 1.0e^{-3})$

2.3.5 *Classifying Question Types*

We trained a simple, yet effective logistic regression classifier to classify the question types. From Table 2.4 it becomes clear that we can accurately learn to classify different question types, especially at higher levels in the hierarchy. These labels are extremely helpful for a number of tasks: they are useful to decide what type of answer the user is expecting, or the type of model that should deliver a response. An accurate classification on the first level is important for this task: do we want to use a rule-based system, a factoid QA model, or a newly trained document-centered QA model? The results on the second level can be used to decide whether or not we face a yes / no question and therefore may have to start the answer with "yes" or "no." In a question generation setting, the labels can also be used to condition the question generation process.

2.3.6 Answering Research Question 1.1 and Research Question 1.2

The results of the survey allow us to answer our first two research questions. We have identified a range of capabilities that users would like to see in a document-centered assistance scenario, and we have identified a hierarchy of questions that users would ask. Document-centered questions are different from factoid QA questions and form an interesting new category of questions to research.

2.4 STEP 2 – DATA COLLECTION

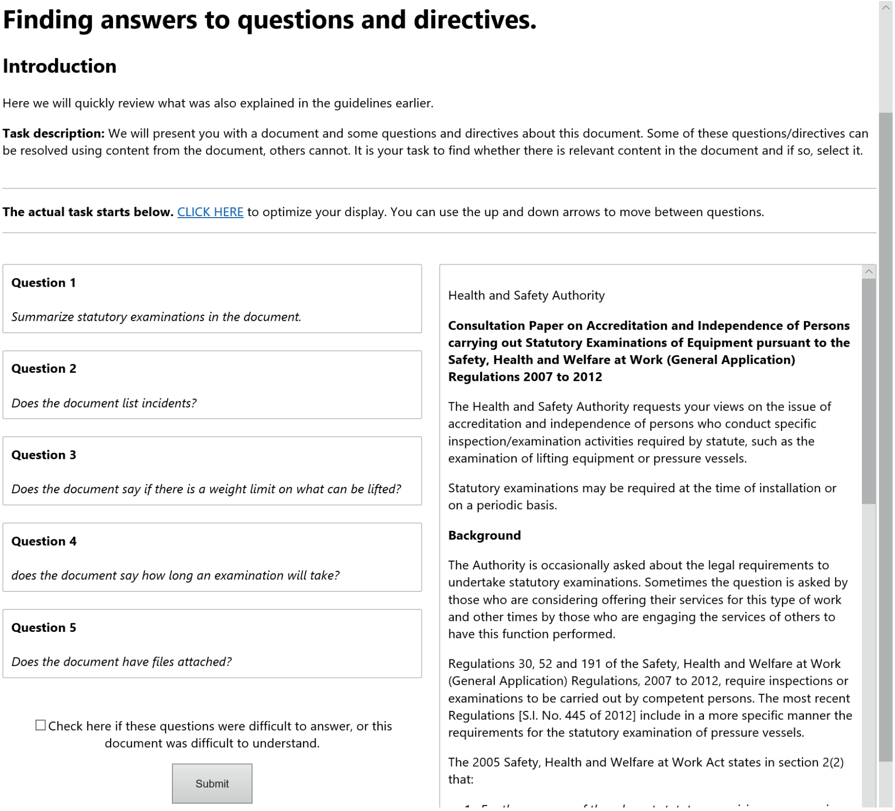
The first step of our work shows that users pose different types of questions to a digital assistant when seeking document-centered assistance than are typically present in modern QA datasets. To dive deeper, we first scale up our data collection to gather more questions and proposed answers to those questions. In this section, we describe our data collection process and the statistics of the collected data. We refer to the collected data as "DQA" dataset, short for Document Question-Answering.

2.4.1 Question Collection

For the question collection, we randomly selected another 36 documents using the same selection criteria as in Section 2.3.3. We asked the same set of participants as in Step 1, now acting as crowd workers, to generate questions for these documents. This time we omitted the survey questions about the scenario and capabilities; we asked them to pose five questions about the document. Since we only presented the workers with the document introduction, it is likely that workers will also ask questions that cannot be answered from the document, more closely resembling a real life situation.

2.4.2 Answer Collection

Once we collected the questions, we asked the same pool of crowd workers to select answers for these questions. We presented workers with the full docu-



Finding answers to questions and directives.

Introduction

Here we will quickly review what was also explained in the guidelines earlier.

Task description: We will present you with a document and some questions and directives about this document. Some of these questions/directives can be resolved using content from the document, others cannot. It is your task to find whether there is relevant content in the document and if so, select it.

The actual task starts below. CLICK HERE to optimize your display. You can use the up and down arrows to move between questions.

Question 1

Summarize statutory examinations in the document.

Question 2

Does the document list incidents?

Question 3

Does the document say if there is a weight limit on what can be lifted?

Question 4

does the document say how long an examination will take?

Question 5

Does the document have files attached?

□ Check here if these questions were difficult to answer, or this document was difficult to understand.

Health and Safety Authority

Consultation Paper on Accreditation and Independence of Persons carrying out Statutory Examinations of Equipment pursuant to the Safety, Health and Welfare at Work (General Application) Regulations 2007 to 2012

The Health and Safety Authority requests your views on the issue of accreditation and independence of persons who conduct specific inspection/examination activities required by statute, such as the examination of lifting equipment or pressure vessels.

Statutory examinations may be required at the time of installation or on a periodic basis.

Background

The Authority is occasionally asked about the legal requirements to undertake statutory examinations. Sometimes the question is asked by those who are considering offering their services for this type of work and other times by those who are engaging the services of others to have this function performed.

Regulations 30, 52 and 191 of the Safety, Health and Welfare at Work (General Application) Regulations, 2007 to 2012, require inspections or examinations to be carried out by competent persons. The most recent Regulations [S.I. No. 445 of 2012] include in a more specific manner the requirements for the statutory examination of pressure vessels.

The 2005 Safety, Health and Welfare at Work Act states in section 2(2) that:

Figure 2.6: Question-answering data collection overview.

ment and asked them to read it carefully. Then we asked them to answer five questions about the document. These questions were always a set of five questions that were asked by one of the crowd workers in the question collection round (not necessarily the same as the worker who is answering the questions). The questions were kept together and were presented in the same order as they were asked, due to the potential conversational nature of the questions. Note that this is only applicable to a few instances in the data, allowing us to train a single-turn QA model later. Each set of questions is answered by three crowd workers. An overview of the presented task is included in Figure 2.6.

For each question, we display the following options after a click on the question:

- This question or directive does not make sense;

- The document does not contain the answer to this question;
- Please indicate the question type:
  - This is a yes / no question;
  - This is not a yes / no question.

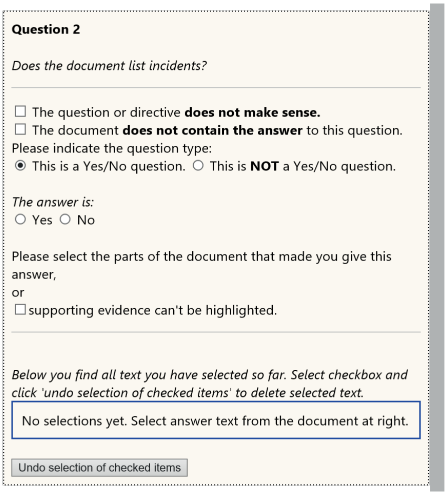
If a worker selects that the question is a yes / no question, we ask them to indicate whether the answer is "yes" or "no" and to select parts of the document with supporting evidence. If no supporting evidence could be found in the document (e.g., because the question was *"does the document contain information about topic X?"* and the answer was "no") we asked workers to tick the box that supporting evidence cannot be highlighted. An example of the task including the expansion that is shown if a worker selects that the question is a yes / no question is given in Figure 2.7. If the worker has not clicked any of the above mentioned options, it means the question is valid, open-ended, and answerable. For these questions, we asked workers to select the minimal spans of text necessary to answer the question. Workers could select up to three spans in the document; each span was at most 700 characters in length. Since some documents can be challenging to understand, we included a checkbox to indicate that the questions were difficult to answer or the document was hard to understand. Figure 2.8 shows an example of the highlighting tool. Text highlighted in the document (right-hand pane), is populated as a selected span in the left-hand pane (blue box).

We again performed 2 training rounds with the crowd workers, in which we ensured workers fully understand the task. During the data collection phase an expert spot-checked answer quality.

2.4.3 *Dataset Statistics*

Table 2.5 describes the distribution of annotations about the questions that were collected from the crowd workers. Recall that each question was judged and answered by 3 workers. Here we present the raw numbers.

During the question generation phase workers were not shown the full document, whereas the workers have access to the full text while selecting answers. This disparity is reflected in the statistic that 40% of questions were considered unanswerable from the text. This ensures that our dataset is suitable for training a system that can identify unanswerable questions.



Question 2

Does the document list incidents?

☐ The question or directive **does not make sense.**
☐ The document **does not contain the answer** to this question.
Please indicate the question type:
● This is a Yes/No question. ○ This is **NOT** a Yes/No question.

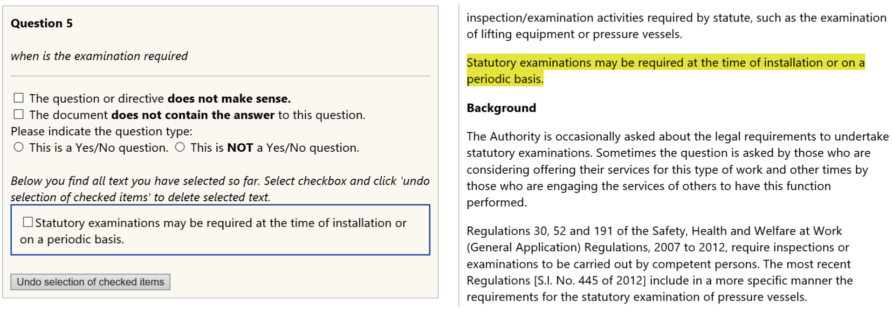
The answer is:
○ Yes ○ No

Please select the parts of the document that made you give this answer,
or
☐ supporting evidence can't be highlighted.

Below you find all text you have selected so far. Select checkbox and click 'undo selection of checked items' to delete selected text.

No selections yet. Select answer text from the document at right.

Figure 2.7: Question-answering data collection yes/no expansion.



Question 5

when is the examination required

☐ The question or directive **does not make sense.**
☐ The document **does not contain the answer** to this question.
Please indicate the question type:
○ This is a Yes/No question. ○ This is **NOT** a Yes/No question.

Below you find all text you have selected so far. Select checkbox and click 'undo selection of checked items' to delete selected text.

☐ Statutory examinations may be required at the time of installation or on a periodic basis.

inspection/examination activities required by statute, such as the examination of lifting equipment or pressure vessels.

Statutory examinations may be required at the time of installation or on a periodic basis.

Background

The Authority is occasionally asked about the legal requirements to undertake statutory examinations. Sometimes the question is asked by those who are considering offering their services for this type of work and other times by those who are engaging the services of others to have this function performed.

Regulations 30, 52 and 191 of the Safety, Health and Welfare at Work (General Application) Regulations, 2007 to 2012, require inspections or examinations to be carried out by competent persons. The most recent Regulations [S.I. No. 445 of 2012] include in a more specific manner the requirements for the statutory examination of pressure vessels.

Figure 2.8: Question-answering data collection selected text.

Table 2.5: Answer and question types.

	Number	**% (of total)**
Annotated documents	56	–
Valid questions (= annotation tasks)	16,375	100.00
Invalid questions (discarded)	425	–
Open questions	9,442	57.66
Yes/no questions	6,933	42.34
No answer	6,543	39.96
No evidence	1,748	25.21

Table 2.6: Span statistics. Span length in tokens.

	Statistic
Total number of spans	11,702
Average number of spans per question (all)	0.715
Average number of spans per question with answer	1.45
Average span length per question (all)	26.69
Average span length per question with answer	37.35

Table 2.6 gives an overview of the number of spans and the lengths of spans that were selected by crowd workers. The average span length is substantially larger than the average span length of only a few words in most existing QA datasets. This supports our claim that the current document-centered scenario requires different types of data to train on. Table 2.7 describes the distribution of annotation responses, in particular the fraction of questions where workers were in full agreement about the impossibility of answering a question

Table 2.7: Agreement statistics.

	Metric
Impossible full agreement (%)	52.09
Impossible partial agreement (%)	47.91
ROUGE-1 F-score avg (questions with span)	$52.44_{\pm 8.79}$
ROUGE-2 F-score avg (questions with span)	$44.92_{\pm 11.14}$
ROUGE-L F-score avg (questions with span)	$46.89_{\pm 9.54}$

from the text (52%) (random full agreement would be 25%), as well as ROUGE-scores describing the mean self-similarity of selected spans across judges who responded to the same question. Hence, participants agreed well with each other.

2.5 STEP 3 – BASELINE MODELING

We present baseline models for passage retrieval and answer selection on our dataset. Our aim is to answer Research Question 1.3 *(How well do initial baseline models perform in a document-centered scenario?).*

2.5.1 *Data Preprocessing*

We use exactly the same format as the popular SQuAD2.0 (Rajpurkar et al., 2018) dataset for our preprocessing output. We keep all questions and answers for a random sample of 25% of the documents as a separate hold-out set. Recall that we have collected 3 answers per question, as we had 3 workers answer each question. We discarded all invalid questions and we ensured that the remaining labels (such as "yes / no questions") were consistent as follows. First we looked at workers' answers for whether the question was a yes / no question and computed the majority vote. We kept the answers of the workers who agreed with the majority vote and discarded the rest (if any). The majority vote has been shown to be a strong indication for the true label (Li, 2019). In case of a tie, we chose to treat this question as a yes / no question as it provided us with most information about the question, which is beneficial for training. If the question is now labeled as a yes / no question we continue to the answer (i.e., "yes" or "no"). Again we computed the majority vote and only kept the answers from workers who agreed with the majority vote. In case of a tie we chose "yes" as the answer, as this results in the richest label for the question. Then we followed the same procedure for the "no-evidence" checkbox, choosing to include spans in the event of a tie. Lastly, if the question was not labeled as a yes / no question, we applied the same majority vote and tie-breaking strategy for whether the document contains the answer. Using this approach, we kept approximately half of the collected question-answer pairs, but ensured that no model is trained on contradictory answers. This improved model performance. During training, we used the collected question-answer pairs as individual training examples, i.e., if we have 2 answers for a question given by 2 workers, we added them separately to our training set. This way we increased the number of training samples. At this stage, we also chose to add all selected spans for an answer separately to our training set. We leave multiple span selection for future work. During evaluation we treated all selected answers for a question as valid answers.

2.5.2 Passage Ranking

In this section, we describe our approach for initial passage ranking experiments on our new DQA dataset. We explore three baseline methods: random selection, BM25-based ranking, and selecting the first passage in the document.

Passage Construction

During data collection, crowd workers selected answers to questions, yet they did not select the paragraphs or passages that include these answers. Therefore we constructed passages for all questions with answers as follows. We discard questions without answers in this experiment. We split each document in the dataset into sentences. We adopted a sliding window approach, moving our window one sentence at the time, constructing passages of size *window size*. We set the window size to 5. We also divided the selected answers into sentence chunks (or smaller, if only parts of sentences were selected). For each answer, we scored each passage by the number of chunks it contains. That is, a passage received a point for each chunk that is also in the answer.

Baseline Passage Ranking 1 – Random

For this baseline we retrieve a random passage. For each retrieved passage we compute the Rouge-1 F-score, Rouge-2 F-score, and Rouge-L F-score (based on retrieved passage and ground truth) (Lin, 2004) and the *Precision@1*. Recall that we scored paragraphs based on the number of overlapping chunks with the selected answer. Therefore some paragraphs contain only part of the answer, and some contain the full answer. To account for this difference we computed a so-called *hard* and *soft Precision@1*. For the hard version, we assigned binary labels to retrieved passages; 1 if the retrieved passage contains (part of) the answer, 0 if it does not. For the soft version, we scored each retrieved passage as follows: we took the number of overlapping chunks of the retrieved passage and the answer and divided this by the maximum number of overlapping chunks. Since annotators may select answers from different passages, we optimistically took the best passage score per question, i.e., we returned a valid match if the selected passage matched any annotator response.

Baseline Passage Ranking 2 – First passage

For this baseline, we select the document's first passage as an answer to each question. We compute the same metrics as in Baseline 1. The purpose of this baseline is to establish to what extent answers to questions are biased by their presence in the preamble of the document, which was shown to study participants at question generation time.

Baseline Passage Ranking 3 – BM25

For this baseline, we retrieve the best matching passage with BM25 (Robertson, Zaragoza, et al., 2009) and compute the same metrics as in Baselines 1 and 2.

2.5.3 *Results for Passage Ranking*

In Table 2.8 the results for the passage ranking experiments are shown. Analysis of variance (ANOVA), $F(2,54) > 8.9$, $p < 0.0002$ yields significant differences between the three approaches. A post-hoc Tukey test $p < 0.05$ shows that first passage selection significantly outperforms Random for all measures, and BM25 for all measures except Rouge-L. BM25 significantly outperforms Random only for Rouge-L. We hypothesize that the performance of first passage selection can have a number of causes. Firstly, because workers have been shown the introduction of the document, many questions can be tailored towards information located in the introduction. Secondly, workers have read the document from beginning to end, which may have biased them towards selecting from the first part of the document and not from the later parts once they found the answer.

Table 2.8: Results for passage ranking.

Model	P@1 soft	P@1 hard	Rouge-1 F-score	Rouge-2 F-score	Rouge-L F-score
Random	0.29	0.31	0.26	0.13	0.19
First	0.55	0.56	0.32	0.20	0.23
BM25	0.32	0.34	0.29	0.16	0.22

2.5.4 *Answer Selection*

In this section, we discuss how state-of-the-art models for answer selection perform on the DQA data and DQA enhanced with data from the SQuAD2.0 dataset (Rajpurkar et al., 2018). We select this dataset for two reasons: first, it is a standard dataset for benchmarking question-answering tasks and, second, like DQA, it contains questions marked as unanswerable, making it closely compatible with our collected data. All baselines were evaluated using the DQA hold-out set.

Passage Construction

For the answer selection experiments, we selected the passages for each answer using the same windowing method as in the passage ranking experiments. The only difference is that we now only considered passages that contain the full answer. For unanswerable questions, we selected the best matching paragraph with BM25. Even though our previous experiments showed that the answer is often in the first paragraph, we chose BM25 as a less biased and more informed selection procedure.

Baseline Answer Selection 1 – Fine-tuned BERT on SQuAD2.0

For QA, BERT is fine-tuned as follows. A question and a passage are fed to a pre-trained BERT language model. They are separated with a separator token. The final output layer is trained to select the start and end index of the answer, from the input passage. If no answer is detected in the passage, 0 is selected as index for both start and end. For the current baseline we fine-tuned HuggingFace's implementation of BERT Large (Wolf et al., 2020) on 8 Titan XP GPUs, using SQuAD2.0. First, we ensured we got similar scores as reported in the repository for the SQuAD2.0 tasks. Then, we evaluated the model on the DQA hold-out set. We included this baseline to test how a pre-trained and fine-tuned BERT model on a very popular QA dataset performed on our DQA dataset without any adaption.

Baseline Answer Selection 2 – Fine-tuning on SQuAD2.0 with query rewriting

For this baseline, we used the same fine-tuned BERT model as for Baseline 1, yet this time we performed some simple query rewriting on the hold-out set to make our questions more comparable to those the model is fine-tuned on. For query rewriting, we computed the most common n-grams in our document train set. We manually inspected those n-grams and chose to delete the following document and conversational related patterns from our questions, expressed as Python regular expressions:

- `'^does( the)? document (\S)+ (you)? '`
- `'^does it (\S)+ '`
- `'^what does( the)? document (\S)+ (you)? ')`
- `'according to( the)? document(\s,\s|,\s|\s)')`
- `'in( the)? document '`
- `'^assistant, '`

Baseline Answer Selection 3 – Fine-tuning on DQA

For this experiment, we fine-tuned BERT Large using the DQA dataset, and again used the same fine-tuning implementation as used previously. We evaluated on the DQA hold-out dataset.

Baseline Answer Selection 4 – Fine-tuning on DQA with query rewriting

This experiment resembles Baseline 3, but we used the same query rewriting as in Baseline 2 to the train and the hold-out set.

Baseline Answer Selection 5 – Fine-tuning on SQuAD2.0 & DQA

This baseline is similar to Baseline 1, but now we added our data to the existing SQuAD2.0 data set while fine-tuning the BERT Large model. We did this since our DQA dataset is not very large. We expected an improvement in performance when we enhanced our data with more data points. We shuffled the training input randomly. We evaluated on the DQA hold-out set.

Baseline Answer Selection 6 - Fine-tuning on SQuAD2.0 & DQA with query rewriting

This baseline resembles Baseline 5, but we performed the same query rewriting to DQA part of the train set and to the DQA hold-out set as in Baselines 2 and 4.

2.5.5 *Results for Answer Selection*

Table 2.9 shows the results for the answer selection experiments. Fine-tuning BERT on SQuAD2.0 and the DQA data significantly outperforms the other baselines. These results look promising, but reveal an interesting new problem to work on as the scores are significantly lower than we are used to from the typical QA task leader boards such as the SQuAD2.0 challenge. It is interesting to see that query rewriting is not beneficial. We assume that our approach may have been too simplistic. We would like to experiment with different types of query rewriting in future work (e.g., Zhang et al., 2007; Grbovic et al., 2015).

2.5.6 *Answering Research Question 1.3*

We have shown that the initial baseline models perform reasonably well on our new document-centered domain. For the answer selection task, it is beneficial to add data from the Wikipedia domain (SQuAD2.0) during training. This improves the results, but also shows that document-centered assistance is a very different novel domain. While our initial experimental results are promising, there is still plenty of opportunity to improve the models in future work, for example by increasing the dataset size. We also expect improvements if we would train BERT on data similar to the DQA data. As BERT has been trained on the Wikipedia domain — the same domain as the SQuAD2.0 data — BERT could 'memorize' certain parts of the data during training, which could give an advantage when fine-tuning on the SQuAD2.0 data. DQA does not have this advantage. In some specific scenarios, using the meta-structure of the document might help to improve results. However, we consider not relying on

Table 2.9: Results answer selection. All models fine-tuned BERT Large and were evaluated on the DQA hold-out set. AS means Answer Selection. AS 5 significantly outperforms the other baselines (Wilcoxon Signed-rank, $p < 0.001$).

Baseline	Training source	F1	EM
AS 1	SQuAD2.0	27.24	13.21
AS 2	SQuAD2.0 with Eval Query rewriting	26.79	13.09
AS 3	DQA	38.84	18.93
AS 4	DQA with Query rewriting	36.73	17.83
AS 5	SQuAD2.0 + DQA	**41.02****	**20.30****
AS 6	SQuAD2.0 + DQA with Query rewriting	37.28	18.52

this structure as the preferred option since this allows us to generalize quickly over a wide variety of documents.

2.6 CONCLUSIONS AND FUTURE WORK

In this chapter, we explored the novel domain of document-centered digital assistance. We focused on a consumption scenario, in which individuals are a (co-)owner of a document. We answered the first research question of this thesis in three steps. Through a survey, we identified a set of primary capabilities people expect from a digital assistant in a document-centered scenario, as well as a large set of questions that gave us insight into the types of queries that people might pose about a document when they have an approximate or good idea what the document is about. Our explorations shed light on the hierarchy of questions that might be posed, and demonstrate that the types of questions people ask in a document-centered scenario are different from the factoid questions in conventional QA datasets. We show that state-of-the-art QA models can be fine-tuned to perform with reasonable accuracy on the new DQA data. Yet, it has proven to be an unsolved task with many possibilities for future work, e.g., deeper explorations of query rewriting to better tailor document-centered questions to conventional QA systems, and exploring ways to scale up the data to a much larger and broader range of documents.

In the next chapter we continue our human-centered investigation, but for a different task: automatic text summarization. We will propose a survey methodology to investigate what a good and useful summary is for users. In this survey, we also explore an interactive approach to automatic summarization. For this part of the survey we are inspired by the user study questions regarding the most and least useful features for a digital assistant from Section 2.3.

3

WHAT MAKES A GOOD AND USEFUL SUMMARY?

In this chapter we turn our focus to automatic text summarization. This task has enjoyed great progress over the years and is used in numerous applications, impacting the lives of many. Despite this development, there is little research that meaningfully investigates how the current research focus in automatic summarization aligns with users' needs. In this chapter[1] we bridge this gap, and answer the second research question of this thesis:

Research Question 2: *What makes a good and useful summary for users of automatically generated summaries?*

To answer this question, we propose a survey methodology that can be used to investigate the needs of users of automatically generated summaries. Importantly, these needs are dependent on the target group. Hence, we design our survey in such a way that it can be easily adjusted to investigate different user groups. In this work we focus on university students, who make extensive use of summaries during their studies. We find that the current research directions of the automatic summarization community do not fully align with students' needs. Motivated by our findings, we present ways to mitigate this mismatch in future research on automatic summarization: we propose research directions that impact the design, the development and the evaluation of automatically generated summaries.

[1] This chapter is based on (ter Hoeve et al., 2022d).

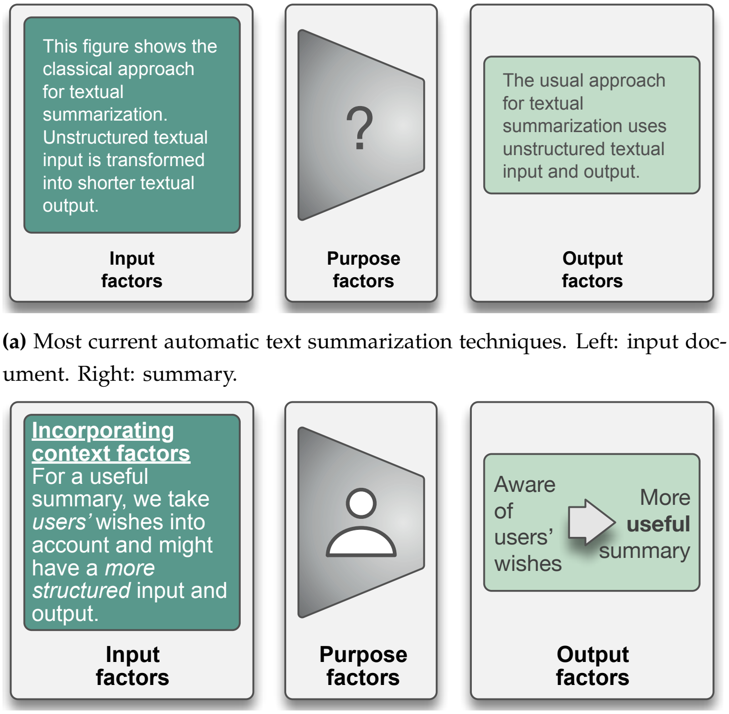
3.1 INTRODUCTION

The field of automatic text summarization has experienced great progress over the last years, especially since the rise of neural sequence to sequence models (e.g., Cheng and Lapata, 2016; See et al., 2017; Vaswani et al., 2017). The introduction of self-supervised transformer language models like BERT (Devlin et al., 2019) has given the field an additional boost (e.g., Liu et al., 2018; Liu and Lapata, 2019; Lewis et al., 2020; Xu et al., 2020b).

The — often *implicit* — goal of automatic text summarization is to generate a condensed textual version of the input document(s), whilst preserving the main message. This is reflected in today's most common evaluation metrics for the task; they focus on aspects such as informativeness, fluency, succinctness and factuality (e.g., Lin, 2004; Nenkova and Passonneau, 2004; Paulus et al., 2018; Narayan et al., 2018b; Goodrich et al., 2019; Wang et al., 2020; Xie et al., 2021). The *needs* of the users of the summaries are often not explicitly addressed, despite their importance in *explicit* definitions of the goal of automatic summarization (Spärck Jones, 1998; Mani, 2001a). Mani defines this goal as: *"to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs."*

Different user groups have different needs. Investigating these needs explicitly is critical, given the impact of adequate information transfer (Bennett et al., 2012). We propose a survey methodology to investigate these needs. In designing the survey, we take stock of past work by Spärck Jones (1998) who argues that in order to generate useful summaries, one should take the context of a summary into account — a statement that has been echoed by others (e.g., Mani, 2001a; Aries et al., 2019). To do this in a structured manner, Spärck Jones introduces three *context factor* classes: *input factors*, *purpose factors* and *output factors*, which respectively describe the input material, the purpose of the summary, and what the summary should look like. We structure our survey and its implications around these factors. In Figure 3.1 we give an example of incorporating the context factors in the design of automatic summarization methods.

Our proposed survey can be flexibly adjusted to different user groups. Here we turn our focus to university students as a first stakeholder group. University students are a particularly relevant group to focus on first, as they bene-


This figure shows the classical approach for textual summarization. Unstructured textual input is transformed into shorter textual output.
Input factors
?
Purpose factors
The usual approach for textual summarization uses unstructured textual input and output.
Output factors
Incorporating context factors
For a useful summary, we take users' wishes into account and might have a more structured input and output.
Input factors
Purpose factors
Aware of users' wishes
More useful summary
Output factors

(a) Most current automatic text summarization techniques. Left: input document. Right: summary.

(b) Example of summarizing while taking users' wishes and desires into account. Left: input document. Right: summary.

Figure 3.1: Example of most current summarization techniques vs. summarization while incorporating the users in the process.

fit from using pre-made summaries in a range of study activities (Reder and Anderson, 1980), but the desired characteristics of these pre-made summaries have not been extensively investigated. We use the word *pre-made* to differentiate such summaries from the ones that users write themselves. Automatically generated summaries fall in the pre-made category, and should thus have the characteristics that users wish for pre-made summaries.

Motivated by our findings, we propose important future research directions that directly impact the design, development, and evaluation of automatically generated summaries. We contribute the following:

- We design a survey that can be easily adapted and reused to investigate and understand the needs of the wide variety of users of automatically generated summaries;

- We develop a thorough understanding of how automatic summarization can optimally benefit users in the educational domain, which leads us to unravel important and currently underexposed research directions for automatic summarization;
- We propose a new, feasible and comprehensive evaluation methodology to explicitly evaluate the usefulness of a generated summary for its intended purpose.

3.2 RELATED WORK

In Section 3.1 we introduced the context factors as proposed by Spärck Jones (1998). Each context factor class can be divided into more fine-grained subclasses. To ensure the flow of the chapter, we list an overview in Appendix 3.A. Below, we explain and use the context factors and their fine-grained subclasses to structure the related work. As our findings have implications for the evaluation of automatic summarization, we also discuss evaluation methods. Lastly, we discuss the use-cases of automatic summaries in the educational domain.

3.2.1 Automatic Text Summarization

Input Factors

We start with the fine-grained input factor *unit*, which describes how many sources are to be summarized at once, and the factor *scale*, which describes the length of the input data. These factors are related to the difference between single and multi-document summarization (e.g., Chopra et al., 2016; Cheng and Lapata, 2016; Wang et al., 2016; Yasunaga et al., 2017; Nallapati et al., 2017; Narayan et al., 2018b; Liu and Lapata, 2019). *Scale* plays an important role when material shorter than a single document is summarized, such as sentence summarization (e.g., Rush et al., 2015). Regarding the *genre* of the input material, most current work focuses on the news domain or Wikipedia (e.g., Sandhaus, 2008; Hermann et al., 2015; Koupaee and Wang, 2018; Liu et al., 2018; Narayan et al., 2018a). A smaller body of work addresses different input genres, such as scientific articles (e.g., Cohan et al., 2018), forum data (e.g., Völske et al., 2017), opinions (e.g., Amplayo and Lapata, 2020) or dialogues (e.g., Liu et al., 2021a). These differences are also closely related to the input factor *subject type*, which describes the difficulty level of the input material. The factor *medium* refers to the input language. Most automatic summarization work is concerned with English as language input, although there are exceptions, such as Chinese (e.g., Hu et al., 2015) or multilingual input (Ladhak et al., 2020). The last input factor is *structure*. Especially in recent neural approaches, explicit structure of the input text is often ignored. Exceptions include graph-based approaches, where implicit document structure is used to summarize a document (e.g., Tan et al., 2017; Yasunaga et al., 2017), and summarization of tabular data (e.g., Zhang et al., 2020b) or screenplays (e.g., Papalampidi et al., 2020).

Purpose Factors

Although identified as the most important context factor class by Spärck Jones (1998) — and followed by, for example, Mani (2001a) — purpose factors do not receive a substantial amount of attention. There are some exceptions, e.g., query-based summarization (e.g., Nema et al., 2017; Litvak and Vanetik, 2017), question-driven summarization (e.g., Deng et al., 2020), personalized summarization (e.g., Móro and Bieliková, 2012) and interactive summarization (e.g., Hirsch et al., 2021). They take the *situation* and the *audience* into account. The *use*-cases of the generated summaries are also clearer in these approaches.

Output Factors

We start with the output factors *style* and *material*. The latter is concerned with the degree of coverage of the summary. Most generated summaries have an *informative* style and cover most of the input material. There are exceptions, e.g., the XSum dataset (Narayan et al., 2018a) which constructs summaries of a single sentence and is therefore more *indicative* in terms of style and inevitably less of the input material is covered. Not many summaries have a *critical* or *aggregative* style. Aggregative summaries put different source texts in relation to each other, to give a topic overview. Most popular summarization techniques focus on a *running format*. Work on template-based (e.g., Cao et al., 2018) and faceted (e.g., Meng et al., 2021) summarization follows a more *headed* (structured) *format*. Falke and Gurevych (2017) build concept maps and Wu et al. (2020) make knowledge graphs. The difference between abstractive and extractive summarization is likely the best known distinction in output type (e.g., Nallapati et al., 2017; See et al., 2017; Narayan et al., 2018b; Gehrmann et al., 2018; Liu and Lapata, 2019), although it is not entirely clear which output factor best describes the difference.

In Section 3.5 we use the context factors to identify future research directions, based on the difference between our findings and the related work.

3.2.2 Evaluation

Evaluation methods for automatic summarization can be grouped in *intrinsic* vs. *extrinsic* methods (Mani, 2001b). Intrinsic methods evaluate the model itself, e.g., on informativeness or fluency (Paulus et al., 2018; Liu and Lapata, 2019). Extrinsic methods target how a summary performs when used for a task (Dorr et al., 2005; Wang et al., 2020). Extrinsic methods are resource intensive, explaining the popularity of intrinsic methods.

Evaluation methods can also be grouped in *automatic* vs. *human* evaluation methods. Different automatic metrics have been proposed, such as Rouge (Lin, 2004) and BERTScore (Zhang et al., 2020c) which respectively evaluate lexical and semantic similarity. Other methods use an automatic question-answering evaluation methodology (Wang et al., 2020; Durmus et al., 2020). Most human evaluation approaches evaluate intrinsic factors such as informativeness, readability and conciseness (DUC, 2003; Nallapati et al., 2017; Paulus et al., 2018; Liu and Lapata, 2019) — factors that are difficult to evaluate automatically. There are also some extrinsic human evaluation methods, where judges are asked to perform a certain task based on the summary (e.g., Narayan et al., 2018b). So far, *usefulness*[2] has not been evaluated in a feasible and comprehensive manner, whereas it is an important metric to evaluate whether summaries fulfill users' needs. Therefore, we bridge the gap by introducing a feasible and comprehensive evaluation methodology to evaluate usefulness.

[2] We follow the definition of the English Oxford Learner's Dictionary (`www.oxfordlearnersdictionaries.com/definition/english/`) for usefulness: *"the fact of being useful or possible to use"*, where *useful* is defined as *"that can help you to do or achieve what you want"*.

3.2.3 Automatic Summarization for Education

Summaries play a prominent role in education. Reder and Anderson (1980) find that students who use a pre-made summary score better on a range of study activities than students who do not use such a summary. As the quality of automatically generated summaries increases (e.g., Lewis et al., 2020; Xu et al., 2020b), so does the potential to use them in the educational domain, especially given the increasing importance of digital tools and devices for education (Luckin et al., 2012; Hashim, 2018). With these developments in mind, it is critical that educators are aware of the pedagogical implications; they need to understand how to best make use of all new possibilities (Hashim, 2018; Amhag et al., 2019). The outcomes of our survey result in concrete suggestions for developing methods for automatic summarization in the educational domain, whilst taking students' needs into account.

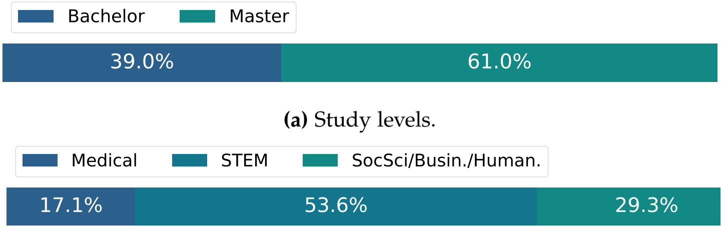
3.3 SURVEY PROCEDURE AND PARTICIPANTS

Here we detail our survey procedure. For concreteness, we present the details with our intended target group in mind. The context factors form the backbone of our survey and the setup can be easily adjusted to investigate the needs of different target groups. For example, we ask participants about a pre-made summary for a recent study activity, but it is straightforward to adapt this to a different use-case that is more suitable for other user groups.

3.3.1 Participants

We recruited participants among students at universities across the Netherlands by contacting ongoing courses and student associations, and by advertisements on internal student websites. As an incentive, we offered a ten euro shopping voucher to ten randomly selected participants.

A total of 118 participants started the survey and 82 completed the full survey, resulting in a 69.5% completion rate. We only include participants who completed the study in our analysis. Participants spent 10 minutes on average on the survey. In the final part of our survey we ask participants to indicate



(a) Study levels.

(b) Study backgrounds.

Figure 3.2: Participant details.

their current level of education and main field of study. The details are given in Figure 3.2.

3.3.2 Survey Procedure

Figure 3.3 shows a brief overview of our survey procedure. A detailed account is given in Appendix 3.B. We arrived at the final survey version after a number of pilot runs where we ensured participants understood their task and all questions. We ran the survey with SurveyMonkey (`surveymonkey.com`). A verbatim copy is released under CC BY license.[3]

Introduction

The survey starts with an introduction where we explain what to expect, how we process the data and that participation is voluntary. After participants agree with this, an explanation of the term *pre-made summary* follows. As we do not want to bias participants by stating that the summary was automatically generated, we explain that the summary can be made by anyone, e.g., a teacher, a good performing fellow student, the authors of the original material, or a computer. Recall that an automatically generated summary is a pre-made summary. Hence, our survey identifies the characteristics an automatically generated summary should have. We also give examples of types of pre-made summaries; based on the pilot experiments we noticed that people missed this

[3] `https://github.com/maartjeth/survey_useful_summarization`

information. We explicitly state that these are just examples and that participants can come up with any example of a helpful pre-made summary.

Context Factors

In the main part of our survey we focus on the context factors. First, we ask participants whether they have made use of a pre-made summary in one of their recent study activities. If so, we ask them to choose the study activity where a summary was most useful. We call this group the *Remembered group*, as they describe an existing summary from memory. If people indicate that they have not used a pre-made summary in one of their recent study activities, we ask them whether they can imagine a situation where a pre-made summary would have been helpful. If not, we ask them why not and lead them to the final background questions and closing page. If yes, we ask them to keep this imaginary situation in mind for the rest of the survey. We call this group the *Imagined group*.

Now we ask the Remembered and Imagined groups about the input, purpose and output factors of the summary they have in mind. We ask questions for each of the context factor subclasses that we discussed in Section 3.2. At this point, the two groups are in different branches of the survey. The difference is mainly linguistically motivated: in the Imagined group we use verbs of probability instead of asking to describe an existing situation. Some questions can only be asked in the Remembered group, e.g., how helpful the summary was.

In the first context factor question we ask what the study material consisted of. We give a number of options, as well as an 'other' checkbox. To avoid position bias, all answer options for multiple choice and multiple response questions in the survey are randomized, with the 'other' checkbox always as the last option. If participants do not choose the 'mainly text' option, we tell them that we focus on textual input in the current study[4] and ask whether they can think of a situation where the input did consist of text. If not, we lead them to the background questions and closing page. If yes, they proceed to the questions that give us a full overview of the input, purpose and output factors of the situation participants have in mind. Finally, we ask the Remembered group to suggest how their described summary could be turned into their ideal

[4] Different modalities are also important to investigate, but we leave this for future work to ensure clarity of our results.

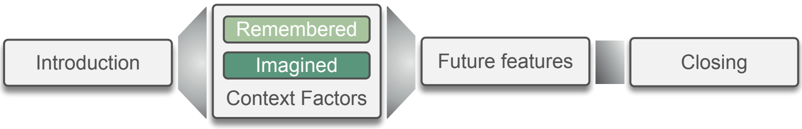


Figure 3.3: Overview of the survey procedure.

summary. We then ask both groups for any final remarks about the summary or input material.

Trustworthiness and Future Features Questions

So far we have included the option that the summary was machine-generated, but also explicitly included other possibilities to not bias participants. At this point we acknowledge that machine-generated summaries could give rise to additional challenges and opportunities. Hence, we include some exploratory questions to get an understanding of the trust users would have in machine-generated summaries and to get ideas for the interpretation of the context factors in exploratory settings.

For the first questions we tell participants to imagine that the summary was made by a computer, but contained all needs identified in the first part of the survey. We then ask them about trust in computer- and human-generated summaries. Next, we ask them to imagine that they could interact with the computer program that made the summary in the form of a digital assistant. We tell them not to feel restricted by the capabilities of today's digital assistants. We ask participants to select the three most and the three least useful features for the digital assistant, similar to ter Hoeve et al. (2020).

3.4 RESULTS

For each question we examine the outcomes of all respondents together and of different subgroups (Table 3.1). For space and clarity reasons, we present the results of all respondents together, unless interesting differences between groups are found. We use the question formulations as used for the Remembered group and abbreviate answer options. Answers to multiple choice and multiple response questions are presented in an aggregated manner and we

Table 3.1: Levels of investigation. We did not find significant differences for each, but add all for completeness.

1	All respondents together
2	Remembered branch vs Imagined branch
3	Different study fields
4	Different study levels
5	Different levels of how helpful the summary was according to participants, rated on a 5-point Likert scale (note that only the *remembered* group answered this question)

ensure that none of the open answers can be used to identify individual participants.

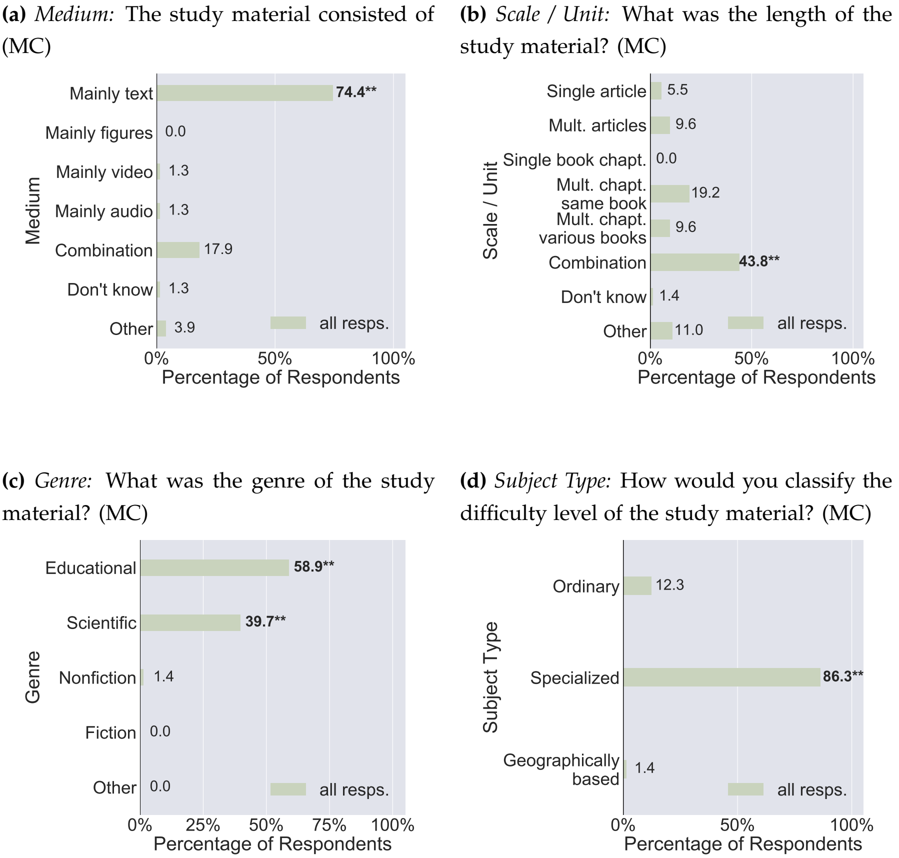
3.4.1 *Identifying Branches*

Of our participants, 78.0% were led to the Remembered branch and of the remaining 22.0%, 78.2% were led to the Imagined branch. We asked the few remaining participants why they could not think of a case where a pre-made summary could be useful for them. People answered that they would not trust such a summary and that making a summary themselves helped with their study activities.

3.4.2 *Input Factors*

Figure 3.4 shows the input factor results. We highlight some here. Textual input is significantly more popular than other input types (Figure 3.4a),[5] stressing the relevance of automatic text summarization. People described a diverse input for *scale* and *unit* (Figure 3.4b), much more than the classical focus of automatic summarization suggests. Most input had a considerable amount of structure (Figure 3.4e). Structure is often discarded in automatic summarization, although it can be very informative.

[5] This is based on people's initial responses and not on the follow-up question if they selected another option than 'text'.



(a) *Medium:* The study material consisted of (MC)

(b) *Scale / Unit:* What was the length of the study material? (MC)

(c) *Genre:* What was the genre of the study material? (MC)

(d) *Subject Type:* How would you classify the difficulty level of the study material? (MC)

Figure 3.4: *Figure continues on next page.* Results for the ***input factor*** questions. Specific input factor in italics. Answer type in brackets: MC = Multiple Choice, MR = Multiple Response. ** indicates significance (χ^2), after Bonferroni correction, with $p \ll 0.001$. If two options are flagged with **, these options are not significantly different from each other, yet both have been chosen significantly more often than the other options.

(e) *Structure:* How was the study material structured? (MR)

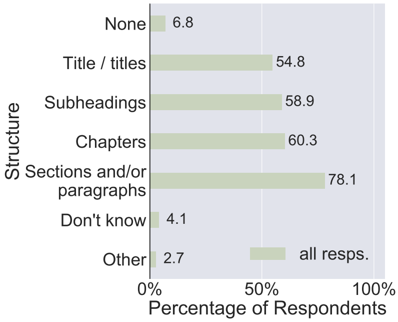


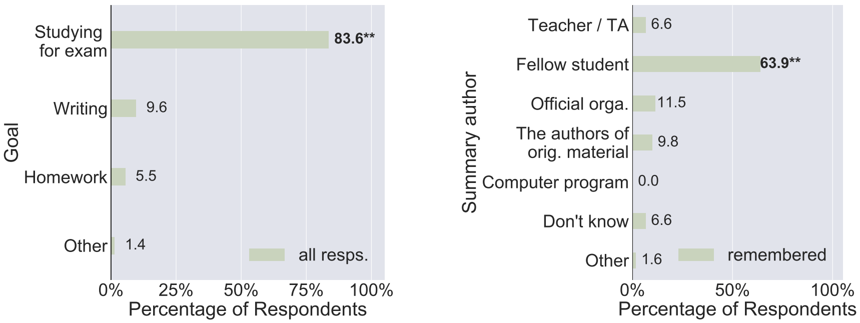
Figure 3.4: *Figure continued from previous page.* Results for the ***input factor*** questions. Specific input factor in italics. Answer type in brackets: MC = Multiple Choice, MR = Multiple Response. ** indicates significance (χ^2), after Bonferroni correction, with $p \ll 0.001$. If two options are flagged with **, these options are not significantly different from each other, yet both have been chosen significantly more often than the other options.

3.4.3 Purpose Factors

Figure 3.5 shows the purpose factor results. Participants indicated that the summary was *helpful* or *very helpful* (Figure 3.5f), which allows us to draw valid conclusions from the survey.[6] We now highlight some results from the other questions in this category. For the intended *audience* of the summaries, students selected level (4) and (5) (*"a lot (4) or full (5) domain knowledge is expected from the users of the summary"*) significantly more often than the other options (Figure 3.5d). Although perhaps unsurprising given our target group, it is an important outcome as this requires a different level of detail than, for example, a brief overview of a news article. People used the summaries for many different use-cases (Figure 3.5e), whereas current research on automatic summarization mainly focuses on giving an overview of the input. We show the

[6] Because we do not find significant differences in the overall results when we exclude the few participants who did not find their summary helpful and we do not find many correlations w.r.t. how helpful a summary was and a particular context factor, we include all participants in the analysis, regardless of how helpful they found their summary, for completeness.

results for the Remembered vs. Imagined splits, as the Imagined group chose *refresh memory* and *overview* more often than the Remembered group (Fisher's exact test, $p < 0.05$). Although not significant after a Bonferroni correction, this can still be insightful for future research directions. Lastly, participants in the Imagined group ticked more boxes than participants in the Remembered group: 3.33 vs. 2.57 per participant on average, stressing the importance of considering many different use-cases for automatically generated summaries.



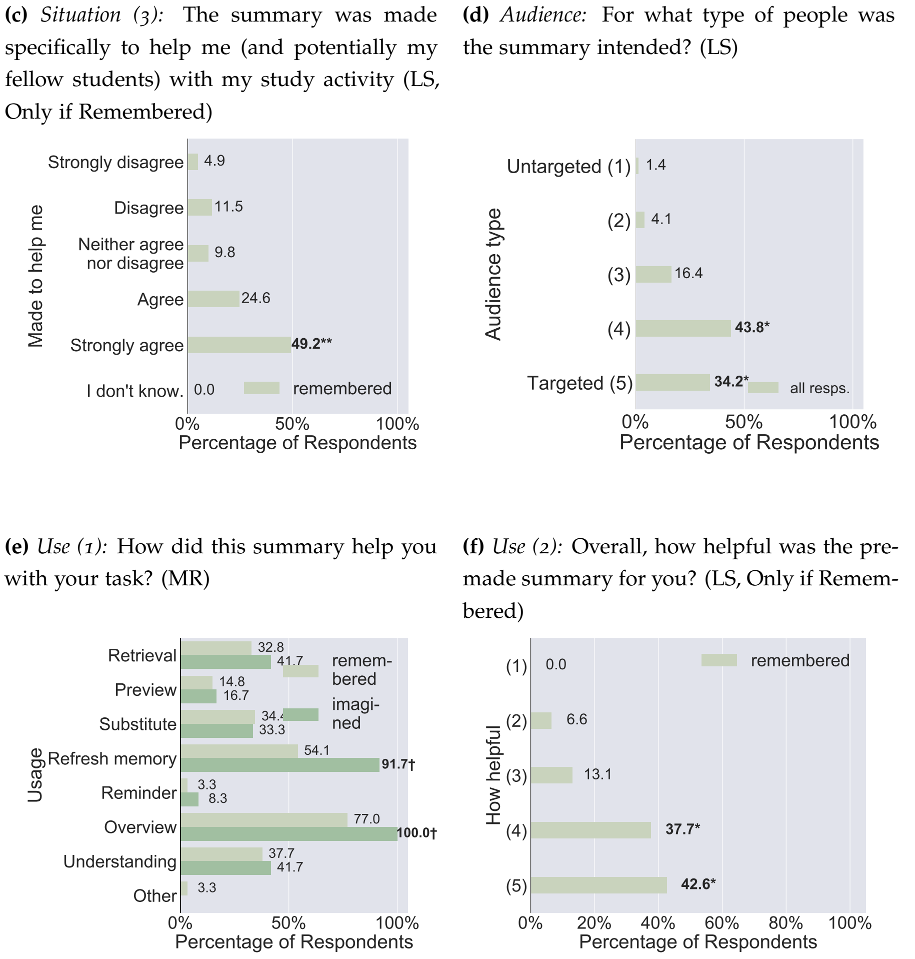
(a) *Situation (1):* What was the goal of this study activity? (MC)

(b) *Situation (2):* Who made this pre-made summary? (MC, Only if Remembered)

Figure 3.5: *Figure continues on next page.* Results for the ***purpose factor*** questions. Specific purpose factor in italics. Answer type in brackets: MC = Multiple Choice, MR = Multiple Response, LS = Likert Scale. ** indicates significance (χ^2), after Bonferroni correction, with $p \ll 0.001$, * with $p < 0.05$. † indicates noteworthy results where significance was lost after correction for the number of tests. If two options are flagged, these options are not significantly different from each other, yet both were chosen significantly more often than the other options.

3.4.4 *Output Factors*

Figure 3.6 shows the results for the output factor questions. Textual summaries were significantly more popular than other summary types (Figure 3.6a), which again stresses the importance of automatic text summarization. Most participants indicated that the summary covered (or should cover) most of the input *material* (Figure 3.6c). For the output factor *style* we find an interesting difference between the Remembered and Imagined group (Figure 3.6d). Whereas the



(c) *Situation (3):* The summary was made specifically to help me (and potentially my fellow students) with my study activity (LS, Only if Remembered)

(d) *Audience:* For what type of people was the summary intended? (LS)

(e) *Use (1):* How did this summary help you with your task? (MR)

(f) *Use (2):* Overall, how helpful was the pre-made summary for you? (LS, Only if Remembered)

Figure 3.5: *Figure continued from previous page.* Results for the ***purpose factor*** questions. Specific purpose factor in italics. Answer type in brackets: MC = Multiple Choice, MR = Multiple Response, LS = Likert Scale. ** indicates significance (χ^2), after Bonferroni correction, with $p \ll 0.001$, * with $p < 0.05$. † indicates noteworthy results where significance was lost after correction for the number of tests. If two options are flagged, these options are not significantly different from each other, yet both were chosen significantly more often than the other options.

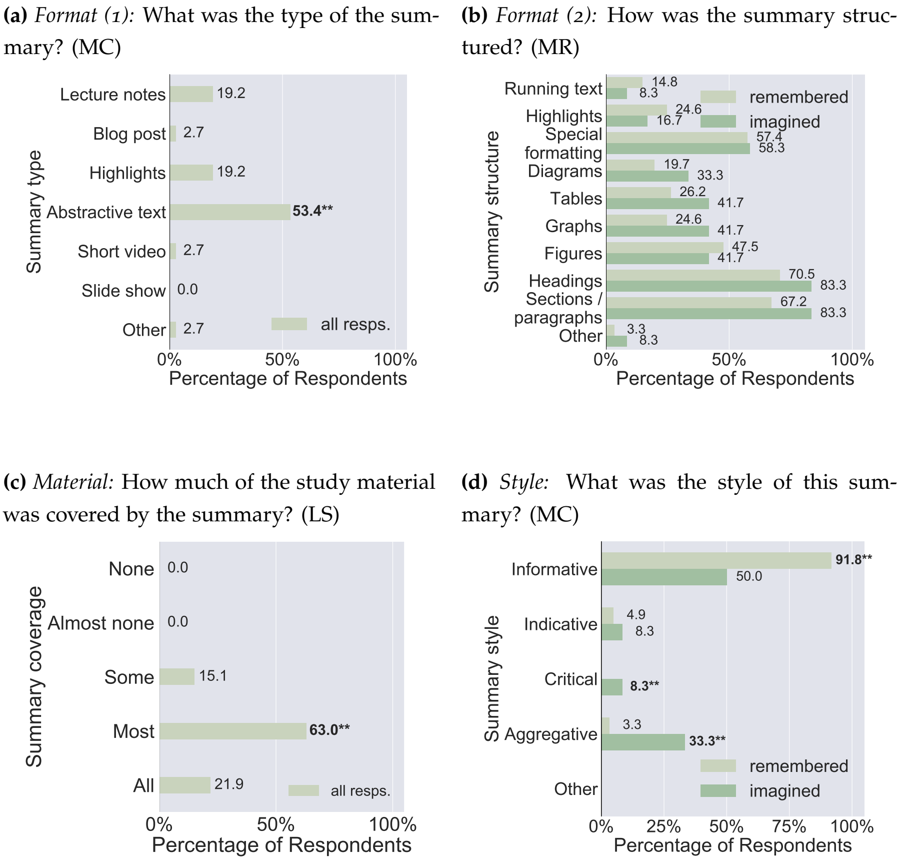
Remembered group described significantly more often an *informative* summary, the Imagined group opted significantly more often for a *critical* or *aggregative* summary. Most research on automatic summarization focusses on informative summaries only. For the output factor *structure* (Figure 3.6b), people described a substantially richer format of the pre-made summaries than adopted in most research on automatic summarization. Instead of simply a running text, the vast majority of people indicated that the summary contained (or should contain) structural elements such as special formatting, diagrams, headings, etc. Moreover, the Imagined group ticked more answer boxes on average than the Remembered group: 4.17 vs. 3.56 per participant, indicating a desire for structure in the generated summaries, which is supported by the open answer questions.

Open Answer Questions

We asked participants in the Remembered group how the summary could be transformed into their ideal summary and 86.9% of these participants made suggestions. Many of those include adding additional structural elements to the summary, like figures, tables or structure in the summary text itself. For example, one of the participants wrote: *"An ideal summary is good enough to fully replace the original (often longer) texts contained in articles that need to be read for exams. The main purpose behind this is speed of learning from my experience. More tables, graphs and visual representations of the study material and key concepts / links would improve the summary, as I would faster comprehend the study material."* Another participant wrote: *"– colors and a key for color-coding – different sections, such as definitions on the left maybe and then the rest of the page reflects the structure of the course material with notes on the readings that have many headings and subheadings."*

Another theme is the desire to have more examples in the summary. One participant wrote: *"More examples i think. For me personally i need examples to understand the material. Now i needed to imagine them myself"*.

Some participants wrote that they would like a more personalized summary, for example: *"I'd highlight some things I find difficult. So I'd personalise the summary more."* Another participant wrote: *"Make it more personalized may be. These notes were by another student. I might have focussed more on some parts and less on others."*



(a) *Format (1):* What was the type of the summary? (MC)

(b) *Format (2):* How was the summary structured? (MR)

(c) *Material:* How much of the study material was covered by the summary? (LS)

(d) *Style:* What was the style of this summary? (MC)

Figure 3.6: Results for the ***output factor*** questions. Specific output factor in italics. Answer type in brackets: MC = Multiple Choice, MR = Multiple Response, LS = Likert Scale. ** indicates significance (χ^2 or Fisher's exact test), after Bonferroni correction, with $p \ll 0.001$, * with $p < 0.05$.

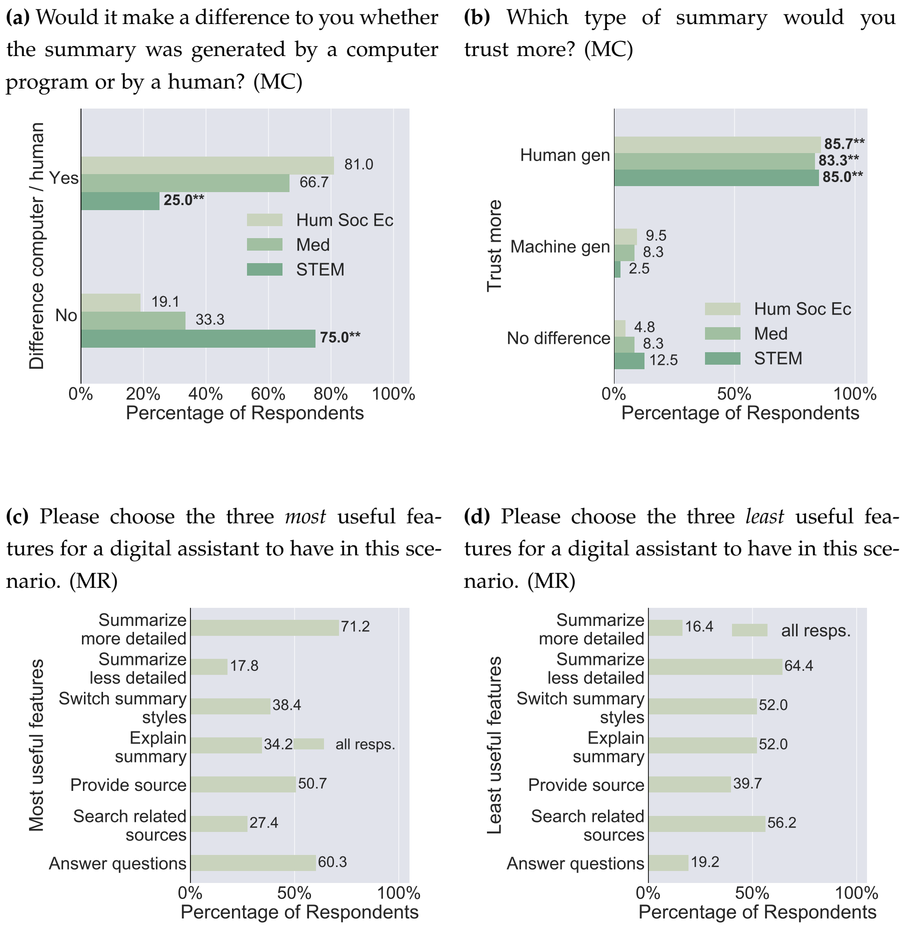
Trustworthiness and Future Features

In this section we report the results for the exploratory questions that we asked about the trustworthiness of a summary generated by a machine versus a human, as well as the results for the questions about features for summarization with a digital voice assistant.

We find that participants are divided on the question whether it would make a difference to them whether the summary was generated by a machine or a computer. If we look at all participants together, we find that 48.0.% of the participants answered that it would make a difference, whereas 52.0% answered that it would not. However, if we split the participants based on study background, an interesting difference emerges (Figure 3.7a). Participants with a background in STEM indicated significantly more often that it would not make a difference to them, whereas the other groups of students indicated the opposite. Almost all participants who answered that it would make a difference said that they would not trust a computer on being able to find the relevant information, i.e., all seemed to favor the human generated summary. Only one participant advocated for the computer-generated summary as a *"computer is more objective."* Almost all participants who said it would not matter to them did add the condition that the quality of the generated summary should be as good as if a human had generated it. One person wrote: *"If the summary captures all previously discussed elements it is effectively good for the same purpose. So then it does not matter who generated it."* This comment exactly captures the motivation of the setup of our survey.

This caution regarding automatically generated summaries is confirmed by the question in which we asked which type of summary participants would trust more — a human-generated one or a machine-generated one. People chose the human-generated summary significantly more often (Figure 3.7b). This also holds for the participants with a STEM background, which aligns with the responses to the open questions we reported earlier — apparently participants do not fully trust that the condition they raised earlier would be satisfied, namely that only if the machine was just as good as the human, it would not matter for them whether the summary was generated by a machine or a human.

The results for the most and least useful features for a digital assistant in a summarization scenario are given in Figure 3.7c and 3.7d. Adding more details to the summary and answering questions based on the content of the



(a) Would it make a difference to you whether the summary was generated by a computer program or by a human? (MC)

(b) Which type of summary would you trust more? (MC)

(c) Please choose the three *most* useful features for a digital assistant to have in this scenario. (MR)

(d) Please choose the three *least* useful features for a digital assistant to have in this scenario. (MR)

Figure 3.7: Results for the future feature questions. Answer type in brackets. MC = Multiple Choice, MR = Multiple Response. ** indicates significance (χ^2 or Fisher's exact test), after Bonferroni correction, with $p \ll 0.001$.

summary are very popular features, whereas summarizing parts of the input material with less detail is not.

Lastly, we asked participants whether they could think of any other features that they would like their digital assistant to have in the outlined scenario. A number of participants answered that they would like the digital assistant to generate questions based on the summary, so that they could test their own understanding. For example, one participant said: *"Make questions for me (to test me)"* and another participant had a related comment: *"Maybe the the digital assistant could find old exam questions to link to parts of the summary where the question is related to, so that there is a function to test if you've understood the summary."* Another line of answers pointed towards giving explicit relations between the input material and summary, for example: *"Show links between subject materials and what their relation is"* and another person wrote: *"Dynamic linking from summary to original source is a great added value of generating a summary"*.

3.5 IMPLICATIONS AND PERSPECTIVES

3.5.1 *Future Research Directions*

Our findings have important implications for the design and development of future automatic summarization methods. We present these in Table 3.2, per context factor. Summarizing, the research developments as summarized in Section 3.2 are encouraging, yet given that automatic summarization methods increasingly mediate people's lives, we argue that more attention should be devoted to its stakeholders, i.e., to the purpose factors. Here we have shown that students, an important stakeholder group, have different expectations of pre-made summaries than what most automatic summarization methods offer. These differences include the type of input material that is to be summarized, but also how these summaries are presented. Presumably, this also holds for other stakeholder groups and thus we hope to see our survey used for different target groups in the future.

Table 3.2: Implications for future research directions.

Input Factors
Stronger focus on developing methods that can: • handle a wide variety and a mixture of **different types** of input documents at once; • understand the **relationships** between different input documents; • use the **structure** of the input document(s).
Purpose Factors
• Explicitly define a **standpoint** on the purpose factors in each research project; • Include a comprehensive **evaluation** methodology to evaluate usefulness. We propose this in Section 3.5.2.
Output Factors
Stronger focus on developing methods that can: • output different summary **styles**, e.g., informative, aggregative or critical. Especially the last two require a **deeper understanding** of the input material than current models have; • explicitly model and understand **relationships** between different elements in the summary and potentially relate this back to the input document(s).

Datasets

To support these future directions we need to expand efforts on using and collecting a wide variety of datasets. Most recent data collection efforts are facilitating different input factors — the purpose and output factors need more emphasis.

Our findings also impact the evaluation of summarization methods. We discuss this next.

3.5.2 Usefulness as Evaluation Methodology

Following Spärck Jones (1998) and Mani (2001a), we argue that a good choice of context factors is crucial in producing useful summaries for users. It is important to explicitly evaluate this. The few existing methods to evaluate usefulness are very resource demanding (e.g., Riccardi et al., 2015) or not comprehensive enough (e.g., DUC, 2003; Dorr et al., 2005). Thus, we propose a feasible and comprehensive method to evaluate usefulness.

For the evaluation methodology, we again use the context factors. Before the design and development of the summarization method the intended purpose factors need to be defined. Especially the fine-grained factor *use* is important here. Next, the output factors need to be evaluated on the use factors. For this, we take inspiration from research on simulated work tasks (Borlund, 2003). Evaluators should be given a specific task to imagine, e.g., *writing a news article*, or *studying for an exam*. This task should be relatable to the evaluators, so that reliable answers can be obtained (Borlund, 2016). With this task in mind, evaluators should be asked to judge two summaries in a pairwise manner on their usefulness, in the following format: *The [output factor] of which of these two summaries is most useful to you to [use factor]?* For example: *The style of which of these two summaries is most useful to you to substitute a chapter that you need to learn for your exam preparation?* It is critical to ensure that judges understand the meaning of each of the evaluation criteria — *style* and *substitute* in the example. We provide example questions for each of the use and output factors in Appendix 3.C.

3.6 ETHICAL IMPACT

With this work we hope to take a step in the right direction to make research into automatic summarization more inclusive, by explicitly taking the needs of users of these summaries into account. As stressed throughout the chapter, these needs are different per user group and therefore it is critical that a wide variety of user groups will be investigated. There might also be within group differences. For example, in this work we have focussed on students from universities in one country, but students attending universities in other geographical locations and with different cultures might express different needs. It is important to take these considerations into account, to limit the risk of overfitting on a particular user group and potentially harming other user groups.

3.7 CONCLUSION

In this chapter we focused on users of automatically generated summaries and answered the second research question of this thesis. We argued for a stronger emphasis on their needs in the design, development and evaluation of automatic summarization methods. We led by example and proposed a survey methodology to identify these needs. Our survey is deeply grounded in past work by Spärck Jones (1998) on context factors for automatic summarization and can be re-used to investigate a wide variety of users. In this work we use our survey to investigate the needs of university students, an important target group of automatically generated summaries. We found that the needs identified by our participants are not fully supported by current automatic summarization methods and we proposed future research directions to accommodate these needs. Finally, we proposed an evaluation methodology to evaluate the usefulness of automatically generated summaries.

In the next chapter we continue with user-centered automatic summarization, and focus on one of the aspects that participants of our survey missed: summaries with more graphical elements, such as arrows that connect different parts of the summaries.

CHAPTER APPENDIX

3.A OVERVIEW CONTEXT FACTORS

Table 3.A.1: Overview of different context factor classes defined by Spärck Jones (1998), with descriptions of the factors within these classes.

Input Factors	Purpose Factors	Output Factors
Form	***Situation***	***Material***
Structure: How is the input text structured? E.g., subheadings, rhetorical patterns, etc.	*Tied:* It is known who will use the summary, for what purpose and when.	*Covering:* The summary covers all the important information in the source text.
Scale: How large is the input data that we are summarizing? E.g., a book, a chapter, a single article, etc.	*Floating:* It is not (exactly) known who will use the summary, for what purpose or when.	*Partial:* The summary (intentionally) covers only parts of the important information in the source text.
Medium: What is the input language type? E.g., full text, telegraphese style, etc. This also refers to which natural language is used.	***Audience***	***Format***

Genre: What type of literacy does the input text have? E.g., description, narrative, etc.	*Targetted:* A lot of domain knowledge is expected from the readers of the summary.	*Running:* The summary is formatted as an abstract like text.
Subject Type	*Untargetted:* No domain knowledge is expected from the readers of the summary.	*Headed:* The summary is structured following a certain standardized format with headings and other explicit structure.
Ordinary: Everyone could understand this input type.	***Use***	***Style***
Specialized: You need to speak the jargon to understand this input type.	*Retrieving:* Use the summary to retrieve source text.	*Informative:* The summary conveys the raw information that is in the source text.
Restricted: The input type text is only understandable for people familiar with a certain area, for example because it contains local names.	*Previewing:* Use the summary to preview a text.	*Indicative:* The summary just states the topic of the source text, nothing more.
Unit	*Substitutes:* Use the summary to substitute the source text.	*Critical:* The summary gives a critical review of the merits of the source text.

Single: Only one input source is given.	*Refreshing:* Use the summary to refresh one's memory of the source text.	*Aggregative:* Different source texts are put in relation to one another to give an overview of a certain topic.
Multi: Multiple input sources are given.	*Prompts:* Use the summary as an action prompt to read the source text.	

3.B SURVEY OVERVIEW

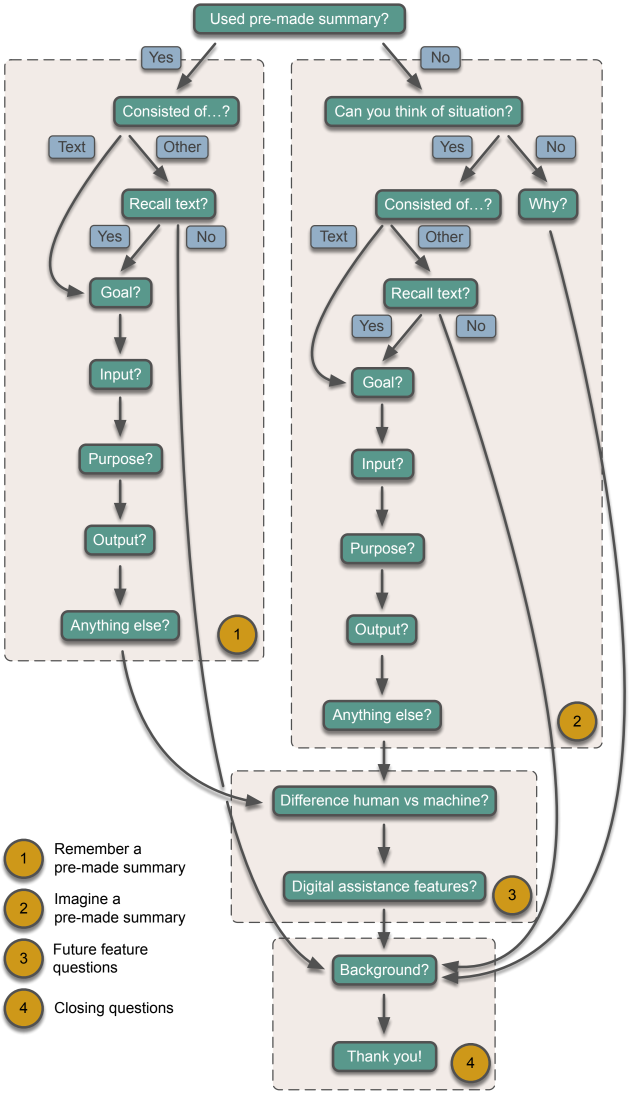


Figure 3.B.1: Overview survey design.

3.C EXAMPLES EVALUATION QUESTIONS

Here we give additional examples for the evaluation questions that can be used for our proposed evaluation methodology. The phrase *"a document that is important for your task"* should be substituted to match the task at hand. For example, in the case of exam preparations, this could be replaced with *"a chapter that you need to learn for your exam preparation"*. Only the questions with the intended purpose factors should be used in the evaluation.

Purpose factor *Use* & Output factor *Style*:

- The *style* of which of these two summaries is most useful to you to *retrieve* a document that is important for your task?
- The *style* of which of these two summaries is most useful to you to *preview* a document that is important for your task?
- The *style* of which of these two summaries is most useful to you to *substitute* a document that is important for your task?
- The *style* of which of these two summaries is most useful to you to *refresh your memory* about a document that is important for your task?
- The *style* of which of these two summaries is most useful to you to *prompt* you to read a source text that is important for your task?

Purpose factor *Use* & Output factor *Format*:

- The *format* of which of these two summaries is most useful to you to *retrieve* a document that is important for your task?
- The *format* of which of these two summaries is most useful to you to *preview* a document that is important for your task?
- The *format* of which of these two summaries is most useful to you to *substitute* a document that is important for your task?
- The *format* of which of these two summaries is most useful to you to *refresh your memory* about a document that is important for your task?
- The *format* of which of these two summaries is most useful to you to *prompt* you to read a source text that is important for your task?

Purpose factor *Use* & Output factor *Material*:

- The *coverage* of which of these two summaries is most useful to you to *retrieve* a document that is important for your task?
- The *coverage* of which of these two summaries is most useful to you to *preview* a document that is important for your task?

- The *coverage* of which of these two summaries is most useful to you to *substitute* a document that is important for your task?
- The *coverage* of which of these two summaries is most useful to you to *refresh your memory* about a document that is important for your task?
- The *coverage* of which of these two summaries is most useful to you to *prompt* you to read a source text that is important for your task?

4

AUTOMATIC SUMMARIZATION WITH GRAPHICAL ELEMENTS

In Chapter 3 and in (ter Hoeve et al., 2022d) we found that summaries generated by automatic summarization methods are not always in line with users' needs. Amongst others, we found a need for summaries with more graphical elements. This is in line with the psycholinguistics literature about how humans process text. In this chapter[1] we are motivated from these two angles, and we answer the third research question of this thesis:

Research Question 3: *How can we fulfill users' request for summaries that include graphical elements?*

To answer this research question, we propose a new task: *summarization with graphical elements*. We verify that these types of summaries are helpful for a critical mass of people. Next, we collect a high-quality human labeled dataset to support research into the task. We then present a number of baseline methods that show that the task is interesting, feasible, and challenging. That is, in this chapter we open a new line of research for the automatic summarization community.

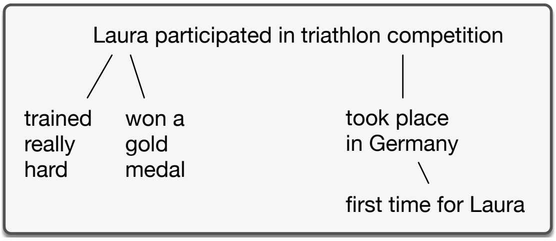
[1] This chapter is based on (ter Hoeve et al., 2022c).

4.1 INTRODUCTION

Automatic text summarization has experienced impressive progress in recent years, with the introduction of neural sequence to sequence models (e.g., Cheng and Lapata, 2016; See et al., 2017; Vaswani et al., 2017). Large, pre-trained language models such as BERT (Devlin et al., 2019) have given the performance another boost (e.g., Liu and Lapata, 2019; Lewis et al., 2020). Typically, progress is measured using automatic evaluation metrics such as Rouge (Lin, 2004) and human evaluation metrics such as informativeness, fluency, succinctness and factuality (e.g., Lin, 2004; Nenkova and Passonneau, 2004; Paulus et al., 2018; Narayan et al., 2018b; Goodrich et al., 2019; Wang et al., 2020; Xie et al., 2021). Importantly, the *purpose* of a generated summary is often not explicitly addressed. A similar observation has been made by ter Hoeve et al. (2022d), who, in line with Spärck Jones (1998), argue that the users of automatically generated summaries are often ignored when designing automatic summarization methods. Moreover, by means of a survey amongst heavy users of automatically generated summaries, they show that users' needs do not fully align with current approaches to automatic text summarization. For example, participants indicated being interested in summaries that contain more graphical elements, such as arrows and colored text to quickly comprehend the summarized material and the relations between different parts of the material. In this work, we build upon the conclusions and recommendations of ter Hoeve et al. (2022d) and take the next step to include user preferences in automatic summarization research. We do this by introducing a new task, in which we specifically target users' wishes to include more graphical elements in the summary. We call our task *summarization with graphical elements*.

In designing the specifics of our task, we also take inspiration from the literature in cognitive science and psycholinguistics on human text understanding. A popular model to capture human text understanding is the *given-new strategy* (Clark and Haviland, 1974; Haviland and Clark, 1974; Clark and Haviland, 1977), which states that when humans process text, they attach new information to already known, i.e., *given*, information in their memory, in order to build up a mental model of the information as a whole. To make this more concrete, we show an example in Table 4.1. In the first row one can read a short story about Laura who took part in a triathlon competition. The second row depicts how a mental model is formed when reading the story according

Table 4.1: Example of the given-new strategy of human text comprehension. While reading the story in the first row, a mental model like in the second row is built. Adapted based on an example in (Carroll, 2008).

Laura participated in a triathlon competition. She had trained really hard and won a gold medal. The competition took place in Germany. For Laura it was her first time in Germany.	

to the given-new strategy; new information is attached to given information as one continues to read the text. While building up this mental model, humans (unconsciously) select which information to keep, and which information can be forgotten (Kintsch and van Dijk, 1978). That is, this process can intuitively be linked to summarization, as also noted by Cardenas et al. (2021).

We arrive at the final task description of *summarization with graphical elements* by combining this psycholinguistic perspective with the user-centered recommendations from ter Hoeve et al. (2022d). A detailed description of the task design is given in Section 4.3. Here we already give an example of a summary that includes graphical elements in Figure 4.1. The summary is different from a standard raw text summary, as it includes the graphical elements that were identified by ter Hoeve et al. (2022d), and it is built up in the style of the given-new strategy: new information is attached to given information. For example, *in Dordrecht* is attached to *a 1£ million full-scale replica of Noah's ark*. In Section 4.3.3 we present the results of a first human evaluation that shows that a critical mass of people is interested in these kinds of summaries.

We divide the remainder of this work into three steps, structured in line with our three main contributions:

Step 1 We introduce a new task: *summarization with graphical elements*. We discuss the task design in detail and confirm that a critical number of people is interested in our proposed summaries.

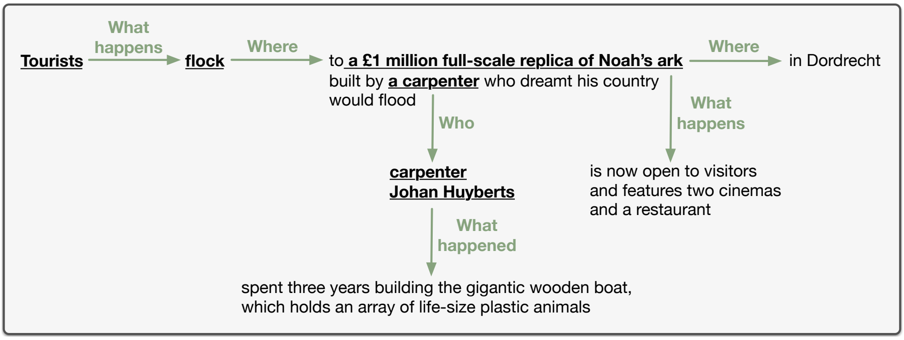

Tourists
What happens
flock
Where
to a £1 million full-scale replica of Noah's ark built by a carpenter who dreamt his country would flood
Where
in Dordrecht
Who
carpenter Johan Huyberts
What happens
is now open to visitors and features two cinemas and a restaurant
What happened
spent three years building the gigantic wooden boat, which holds an array of life-size plastic animals

Figure 4.1: Example of a summary with graphical elements. In this example, nodes with outgoing edges are boldfaced and underlined. The relations are marked with arrows, with the type of relation written on the arrow.

Step 2 Encouraged by these positive findings, we collect a dataset, which we call GraphelSums, to support research into the task; and

Step 3 We present the results of the first baseline experiments. By means of both an automatic evaluation and a human evaluation, we show that our task is feasible and challenging, and can inspire a lot of future work.

We make the code to run the experiments and to obtain the dataset freely available.[2]

4.2 BACKGROUND AND RELATED WORK

In this section we first discuss related work on automatic text summarization (Section 4.2.1). As we also use techniques from information extraction (IE) for our baselines, we discuss this in Section 4.2.2. Lastly, we discuss Snorkel (Ratner et al., 2020) (Section 4.2.3) as a means to automatically generate a larger set of labeled data.

[2] Experiments: https://github.com/maartjeth/summarization_with_graphical_elements
Data: https://github.com/maartjeth/GraphelSums

4.2.1 Automatic Text Summarization

Spärck Jones (1998) formulates three *context factors* for automatic summarization: (i) *input*, (ii) *purpose*, and (iii) *output factors*. These factors describe (i) the input material, (ii) the goal of the summary, and (iii) what the final summary looks like. Most of the work on automatic summarization is concerned with the input and the output factors. Within this space, we mostly find work that focuses on generating a condensed textual version (*output factors*) of a single or multiple document(s), often in the news or Wikipedia domain (*input factors*) (e.g., See et al., 2017; Koupaee and Wang, 2018; Liu and Lapata, 2019; Lewis et al., 2020). As noted by Spärck Jones (1998), and later echoed by Mani (2001a) and ter Hoeve et al. (2022d), not a lot of work in the automatic summarization community focuses on the purpose factors, even though they were identified as the most important context factors by Spärck Jones (1998). This limited focus on the purpose factors is remarkable, as good summaries have the potential to help people with a wide variety of tasks, such as study activities (Reder and Anderson, 1980; ter Hoeve et al., 2022d), and Balasuriya et al. (2021) show that summaries are useful to help people with intellectual disabilities to access information. It also sets automatic summarization aside from other information retrieval and natural language processing tasks where user-centered research plays a much more prominent role, such as search (e.g., Hendriksen et al., 2020; Vakkari, 2020; Ariannezhad et al., 2022; Ariannezhad, 2022), recommendation (e.g., ter Hoeve et al., 2017; Huang et al., 2022), and question-answering and dialogues (e.g., ter Hoeve et al., 2020; Cambazoglu et al., 2021; Siro et al., 2022).

On the side of the *input factors*, there has recently been increased interest in a variety of different interpretations, such as timeline summarization (e.g., Li et al., 2021; Yu et al., 2021), opinion summarization (e.g., Angelidis et al., 2021; Bražinskas et al., 2021) and dialogue summarization (e.g., Feigenblat et al., 2021; Liu et al., 2021a). A smaller body of work has focused on different interpretations of the *output factors*, for example in the form of concept maps (Falke and Gurevych, 2017) or knowledge graphs (Wu et al., 2020). Another example is faceted summarization (e.g., Meng et al., 2021), which aims at bringing structure into summaries by explicitly including different facets of the input text in the summary.

An alternative approach to categorizing different summarization methods is to differentiate between *extractive* and *abstractive* summarization. In *extractive* approaches summaries are constructed by extracting literal parts of the input text. A summary is formed by concatenating the extracted pieces of text (e.g., Narayan et al., 2018b; Ju et al., 2021). *Abstractive* approaches construct summaries by generating new pieces of text, often in an auto-regressive manner (e.g., Gehrmann et al., 2018; Lewis et al., 2020). With the recent introduction of pre-trained, transformer-based models like BART (Lewis et al., 2020) and T5 (Xue et al., 2021), the grammaticality and fluency of abstractive summaries has substantially increased, although they still struggle with factual consistency (e.g., Cao et al., 2020; Maynez et al., 2020).

In our work, we contribute a new interpretation of the output factors, as we create summaries with graphical elements. In designing our task, we take the purpose factors into account; we are inspired by the recommendations from ter Hoeve et al. (2022d) and we explicitly focus on the usefulness of the summaries for users. Importantly, our summaries are *abstractive* in nature and contain longer phrases of text, contrasting our work with previous *extractive* work on knowledge graphs and concept maps, and making our task more challenging.

4.2.2 Information Extraction

The summaries in our task can be represented as *abstractive relation triples* (see Section 4.3.1). For example, *(a £1 million full-scale replica of Noah's ark,* WHERE, *in Dordrecht)* would be one of the abstractive triples of the summary in Figure 4.1. In this work we present a number of baselines to generate these abstractive summary triples. For these baselines, we also use techniques from information extraction (IE). Jurafsky and Martin (2014) define information extraction as the process that "*turns the unstructured information embedded in texts into structured data.*" That is, many well-known IR and NLP tasks are IE tasks, for example, named entity recognition (NER), (co)reference solution, relation extraction and classification, event extraction and classification, and temporal analysis. (Jurafsky and Martin, 2014). Many of these tasks, such as NER-tagging, coreference resolution and relation extraction, play a role in our baselines. Recently, a couple of general frameworks for IE have been proposed (e.g., Qian et al., 2019;

Luan et al., 2019; Wadden et al., 2019). That is, whereas previous methods for IE operated in a cascading manner by, for example, first extracting named entities and then performing relation extraction (e.g., Nadeau and Sekine, 2007; Chan and Roth, 2011), these general approaches combine several IE tasks in a multi-task framework. In this work we make use of DyGIE++ (Wadden et al., 2019), a 100M parameter, BERT-based, well-documented architecture for IE. Briefly, DyGIE++ uses BERT to encode sentences in a token-wise manner, using a sliding window approach. Next, spans of text are constructed, after which a graph structure is generated based on the spans in the document. This graph structure represents the entities, events and relations, and is trained in an end-to-end manner.

Importantly, we cannot only use information extraction to arrive at our desired *abstractive* summary triples, as IE is an *extractive* approach. Hence, we combine IE with abstractive summarization models like BART and T5.

4.2.3 Snorkel

In this work we collect human labels for our task of *summarization with graphical elements* (see Section 4.4), but we also make use of Snorkel (Ratner et al., 2020) to generate a larger amount of labeled data for training purposes. Snorkel is a well-documented method[3] that generates (weak) labels for unlabeled data using (noisy) labeling functions and trained generative models on top of these labeling functions. The labeling functions are written based on heuristics. As an example, imagine that we want to write a labeling function that determines whether two phrases are connected by a when relation. An example of a labeling function that we could write would be:

```
@labeling_function()
def when(x):
    if contain_year(x.phrase_a, x.phrase_b) and are_close(x.phrase_a, x.phrase_b):
        return WHEN
    return ABSTAIN
```

Listing 1: Example of a Snorkel labeling function for the when relation.

[3] https://www.snorkel.org/

This labeling function checks whether one of the two phrases contains a year and whether the two phrases occur close together in the text. Of course, this function does not capture all occurrences of the WHEN relation. First of all, we need additional labeling functions that capture other temporal occurrences, like months and dates. Moreover, the current function still contains ambiguities, making it a *fuzzy* labeling function. For example, a reasonable approach to implement the function to check whether the phrases contain a year, would be to use a regular expression. However, this regular expression will not have perfect precision and recall. Snorkel works by combining a number of these fuzzy, potentially ambiguous labeling functions. The outputs of each of the labeling functions are combined into a probabilistic model that outputs a probabilistic label for each of the data points that were fed to the fuzzy labeling functions. These labels can be used directly, or they can be used as weak labels to train another model for the task at hand. In our work we use them in both ways.

4.3 STEP 1 – DESIGNING THE TASK

In Section 4.1 and 4.2 we already intuitively explained the *summarization with graphical elements* task by means of the example in Figure 4.1. In this section we specify the task in detail, which is the first important step of our work. We discuss the task description (Section 4.3.1), the task domain (Section 4.3.2), and we show that a critical number of people finds such summaries with graphical elements helpful (Section 4.3.3).

4.3.1 *Task Description*

Here we describe our task, *summarization with graphical elements*, more formally. Given an input document $D = [x_0, \ldots, x_n]$, where x_i refers to the i^{th} token in document D, our task is to generate summarizing triples of the form $(y_{a_0} \ldots\ y_{a_k}, relation, y_{b_0} \ldots\ y_{b_k})$, where y_{a_i} and y_{b_i} are tokens, generated in an abstractive fashion. The triples can be thought of in a more graphical way as $(y_{a_0} \ldots\ y_{a_k})$ being a node with an outgoing edge and $(y_{b_0} \ldots\ y_{b_k})$ as a node with an incoming edge. The connecting edge is labeled with *relation*.

In line with the given-new strategy, we improve the conciseness of a summary by merging nodes with outgoing edges that refer to the same entity. As an example, recall Table 4.1, where the phrases *trained really hard* and *won a gold medal* are linked to *Laura*, instead of making a new node for the word *She*. This makes coreference resolution an important part of the task.

Relations could be of any form, ranging from an open to an empty set. We choose to use a closed set of relations L. Explicitly labeling relations, instead of leaving them empty, makes for a more interesting task. Given the nature of our data, we define $L = \{$*who*, *what*, *what happens*, *what happened*, *what will happen*, *where*, *when*, *why*$\}$. We motivate this choice in more detail in the next section.

4.3.2 *Task Domain*

As we are the first to propose the task of *summarization with graphical elements*, there is no standard dataset available for the task, and thus we set up a human labeling effort to collect one ourselves. We call our dataset GraphelSums, short for *summaries with graphical elements*. In Section 4.4 we discuss the specifics of GraphelSums in more detail. In this section we share our considerations for choosing the domain of the dataset.

Requirements

In choosing the specifics of GraphelSums, we need to satisfy a number of requirements:

1. The domain of the data needs to be fully understandable by human annotators in order to ensure high-quality annotations. That is, we cannot choose a domain that can only be fully understood by domain experts, such as scientific documents. Annotators also need to be fluent in the language of the data.

2. The data should naturally fit the task description, i.e., it should be clear how summary triples can be constructed in a meaningful way.

3. The data needs to be easily accessible for others to reproduce and build upon our work.

Decisions

Keeping the requirements in mind, we opt to use the CNN/DM dataset (Hermann et al., 2015),[4] a dataset for English news summarization (e.g., See et al., 2017; Baan et al., 2019a; Baan et al., 2019b), as the basis of our data collection, as it fits all conditions: (i) news documents do not require specific domain knowledge from annotators and our annotators are fluent in English, the language of the dataset; (ii) the abstractive ground truth summaries from the CNN/DM dataset give us a way to construct abstractive summary triples and we can use the 5W's — that are in the nature of news articles and have been used for automatic summarization before (Parton et al., 2009) — as our relations; and (iii) we are able to release the code to obtain the collected labels, making our work reproducible. We acknowledge that the CNN/DM dataset and the (English) news domain in general have been well studied for classical approaches to automatic summarization, i.e., without graphical elements, yet a thorough exploration of alternatives convinced us that the CNN/DM dataset fits our requirements best to start with. We hope that more datasets will be collected for summarization with graphical elements in different domains and languages in the future.

As mentioned, we choose to use the 5W's (*who, what, where, when, why*) as our relations. We add three additional labels to provide more temporal nuance and make the task more challenging: *what happens*, *what happened*, and *what will happen*.

4.3.3 Human Evaluation of the Task Design

Now that we have formally defined the task description and decided on the domain of the task, we run a first proof-of-concept human evaluation, where we investigate whether people find our proposed summaries useful. We stress that we do not necessarily aim for a majority of people; as noted in previous work (Spärck Jones, 1998; ter Hoeve et al., 2022d), different people have different preferences in different contexts. We follow the recommendations from ter Hoeve et al. (2022d) and run a human evaluation to evaluate the usefulness of three types of summaries in a pairwise manner on two purpose factors: *previewing* and *substituting* (Spärck Jones, 1998). We use the following scenario

[4] `https://cs.nyu.edu/~kcho/DMQA/`

outline: *"Imagine that you would like to quickly gather information about a certain news event. To help you quickly find your information, you have access to a summary that describes the news article."* The summaries that we compare are:

1. *A text only summary*: the CNN/DM abstract;
2. *A summary with graphical elements*, for which we use the human labeled summary triples and convert them into a graphical representation;[5]
3. *Typeset control summary*: a purely textual, but formatted summary. We boldface the first sentence and color all phrases that contain outgoing edges in the human-labeled equivalent.

We run this evaluation on Amazon Mechanical Turk.[6] We compare summaries for two different news articles, to control for potential bias towards an article. We also randomize the position of the summaries on the page, to control for position bias. We request 20 judgments per comparison per news article, i.e., 40 judgments in total. Crowd workers are U.S. based, have a HIT approval rate of at least 90% and at least 1000 accepted HITs. For quality control, we add two questions about the content of the summary, which are unique for each pairwise comparison (Appendix 4.A.2). Significant failure to answer these questions results in the rejection of the HIT. We request new labels for rejected HITS and arrive at a total of 120 judgments. An example of the task is given in Appendix 4.A.1, Figure 4.A.1 and Figure 4.B.2.

Findings

Here we report the aggregated results on both news articles (Table 4.2), as we did not find substantial differences between documents (Appendix 4.A.3, Table 4.A.1 and 4.A.2). For both purpose factors, a substantial fraction of the workers favored a summary with graphical elements over the other two summaries and a substantial fraction used the graphical summary to answer the questions about the summary content. Although not a majority, the group is large enough to be convinced that summaries with graphical elements are important to focus on, in line with the findings by ter Hoeve et al. (2022d). We

[5] We acknowledge that the precise lay-out may affect people's judgments. The results of this human evaluation should therefore be taken as an indication of people's preference and taking this work in production should include an additional design step.

[6] http://www.requester.mturk.com/

Table 4.2: Results first human evaluation. Pairwise comparisons (%).

	Pair (A/B)	Prefer A	Prefer B
Used	Graphical/Text	35.0	65.0
	Graphical/Typeset	50.0	50.0
	Typeset/Text	72.5	27.5
Prev	Graphical/Text	35.0	65.0
	Graphical/Typeset	47.5	52.5
	Typeset/Text	75.0	25.0
Sub	Graphical/Text	30.0	70.0
	Graphical/Typeset	32.5	67.5
	Typeset/Text	65.0	35.0

also find that typeset summaries are much more popular than purely textual summaries, making another case for expanding research efforts beyond purely textual summaries. In this work we choose to focus on the more challenging task of including graphical elements in the summaries, but our collected dataset can also be used in follow-up work that focuses on these typeset summaries.

Motivated by these results, we now proceed to collect a larger labeled dataset, to support research into our task of *summarization with graphical elements*.

4.4 STEP 2 — COLLECTING THE DATASET

As stated in Section 4.3.2, we collect a human-labeled dataset called GraphelSums to support the task of *summarization with graphical elements*. In the previous section we discussed our considerations for choosing the CNN/DM dataset as the basis for our data collection. In this section we first discuss the CNN/DM dataset in more detail, including some specific adaptations we make for our task (Section 4.4.1). Next, we discuss our human labeling procedure to collect GraphelSums (Section 4.4.2) and the statistics of GraphelSums (Section 4.4.3).

4.4.1 The CNN/DailyMail Dataset

The CNN/DailyMail dataset (Hermann et al., 2015) is a dataset for English news summarization. Summaries are constructed based on a single news article. The ground truth is formed by so-called "story highlights," which were part of the original news article and are written by the human editors. These story highlights are *abstractive* in nature. Importantly, these story highlights do not include the article title. In preparing our labeling task, we noticed that many of the story highlights are hard to understand without title, as the highlights often refer back to information that was introduced in the title. Therefore, we add the titles to the summary abstracts. In our labeling procedure we confirm that the title is essential in 80% of the abstracts (Section 4.4.3).

4.4.2 Human Labeling Procedure

Here we describe our human labeling procedure to collect GraphelSums. Given the intensity of the labeling procedure (see below), we opt to collect a human-labeled *test set*. Each document in the set is labeled by *three* annotators. This allows us to account for the ambiguity in the summarization process, as there are multiple ways of correctly constructing a summary. The CNN/DM dataset is a popular dataset for automatic summarization, increasing the risk for methods to gradually optimize for the test set, as work that reports a significant improvement on the test set typically has the highest chance of being accepted. Therefore, we also use this opportunity to shuffle the standard train, validation and test sets of the CNN/DM dataset.

Annotators

In order to construct a high quality test set, we recruit three annotators with NLP expertise. The annotators need to construct summary triples, based on input abstracts from the CNN/DM dataset. Annotators are instructed via a detailed instruction manual, a video call in which the manual was discussed, and there was room for asynchronous communication via a chat channel. This allowed annotators to ask questions while doing the annotations and to report mistakes, which improved the quality control of the annotations (Section 4.4.2). Annotators were paid at an hourly rate, removing the incentive to rush responses. Annotators were first asked to annotate a set of six articles, after which they received feedback. During the remaining annotation task we checked the quality of the incoming annotations, by sampling annotations at random and inspecting them manually, to make sure annotators were still on the right track. All annotators annotated the same set of documents, resulting in three annotations per document (apart from a handful of documents for which we had to discard the annotations of an annotator, see Section 4.4.2).

Annotation Task Description

Each document is annotated in a human intelligence task (HIT). Examples are given in Appendix 4.B. Each HIT consists of:

1. *An introduction*, where we iterate the most important parts of the instruction manual, and

2. *The actual task.* Annotators are presented with an abstract taken from our test set, including title. The abstract is divided into sentences. We present annotators with the constituents of the abstract, per sentence, that we obtain with the Berkeley constituency parser (Kitaev and Klein, 2018). Annotators need to select the relation triples and have the option to indicate whether something was wrong with the presented abstract or the constituents, and whether the first sentence of the abstract (i.e., the article title) was fully redundant. If they click that option, we ask them to not select any constituents from the first sentence. Lastly, annotators can check a box if they were particularly uncertain about their annotations.

Quality Control

In addition to the checks during the annotation process, we also inspect the answers to the quality control questions: (i) whether something was wrong with the abstract or presented constituents, and (ii) whether annotators were uncertain about their annotations. We also analyze reported annotation mistakes and the overlap in annotations per HIT.

1. *Issues with the presented abstract or constituents.* Here, we manually inspect the annotations of the HITs where a majority of the annotators indicated that there was something wrong, such as a mistake with the automatically generated constituents. This does not necessarily mean that the

Table 4.3: Overlap statistics of annotators on human-labeled test set. *Const A* refers to the first and *Const B* to the second constituent in a summary triple.

Metric	**Avg ± Std**
Hard F1	$0.21_{\pm 0.16}$
Soft F1	$0.47_{\pm 0.18}$
Jaccard w.r.t. Triple	$0.13_{\pm 0.11}$
Jaccard w.r.t. Const A	$0.28_{\pm 0.16}$
Jaccard w.r.t. Const B	$0.32_{\pm 0.16}$
Jaccard w.r.t. Relations	$0.61_{\pm 0.17}$

annotation is also wrong. We discard eight documents based on these answers. Moreover, we discard one more article because of preprocessing issues with the abstract later in the pipeline.

2. *Annotator uncertainty.* We manually inspect all annotations where annotators indicated to be uncertain about a HIT and we discard twelve HITs.

3. *Reported issues by annotators.* Annotators had the option to report mistakes they made via chat. For example, sometimes annotators realised after submitting a HIT that they chose an incorrect label. Based on these reports we manually made changes to the annotations of six documents.

4. *Overlap in annotations.* We also compute the overlap in annotations for all annotator pairs for each HIT. We compute three types of scores, to evaluate different aspects.

   - *Hard pairwise macro F_1-scores.* In this setting, we only count a selected triple if both annotators have exactly that triple in their annotations.
   - *Soft greedy pairwise macro F_1-scores.* The hard F_1-scores are extremely conservative, especially given the nature and the ambiguity of the annotation task. Intuitively, we also want to assign points if the triple partially overlaps, but not entirely, for example because one annotator decided to include an article, whereas another annotator did not. Therefore, we also compute a soft score, where we greedily align the best matching triples based on the $F1$-scores of lexical overlap. We add a zero score for all triples that could not be matched due to a

different number of triples in the two annotations. We report the average lexical overlap.

- *Pairwise Jaccard scores.* Finally, we compute the average Jaccard scores for annotator pairs. Specifically, we compute these scores for the entire triple, and for each individual component of the summary triple.

All scores are given in Table 4.3. From these scores, it becomes clear that the annotators are aligned in their annotations, yet it also shows that there are different ways to construct the summaries. This underlines our choice of collecting multiple annotations per data point. During evaluation of the task, the best matching annotation can be chosen as ground truth.

4.4.3 Dataset Statistics

We present the statistics of GraphelSums in Table 4.4. Annotators spent around 7 minutes on average per HIT. Within our budget we were able to obtain annotations for 295 documents, i.e., 885 annotations in total. After quality control, our dataset consists of 286 documents, which is comparable to earlier work that constructed human-annotated test sets for summarization (e.g., Wu et al., 2020). The vast majority of documents have three annotations. We also confirm our intuition about including the titles: in almost 80% of the cases the title was needed to understand the summary abstract. Lastly, we find that all relations are represented, with some popular relations such as *what happened* and *what*.

4.5 STEP 3 – BASELINES

As the third and final step of our work we provide a variety of baselines for our task. Each of these baselines consists of an *abstractive* text summarization component, followed by a component in which the final summary triples are constructed. We leave entire end-to-end solutions as an interesting and important direction for future work. In this section we discuss the baselines in detail (Section 4.5.1 and Section 4.5.2), followed by our findings on an automatic evaluation and a human evaluation (Section 4.5.3 and Section 4.5.3).

Table 4.4: Dataset statistics of GraphelSums.

	Counts	**%**
# Documents	286	100
# Three annotations/doc	268	93.7
# Two annotations/doc	18	6.29
# Triples	5942	–
Avg # triples/doc	7.07$_{\pm 3.28}$	–
Title redundant (majority vote)	58	20.3
# Who	439	7.39
# What	1,407	23.7
# What happens	967	16.3
# What happened	2,075	34.9
# What will happen	149	2.51
# Where	339	5.71
# When	333	5.60
# Why	233	3.92

4.5.1 Abstractive Summarization Component

As stressed throughout this work, our task is to generate *abstractive* summary triples, meaning that only using information *extraction* approaches is not enough for our task. We introduce the abstractive component by fine-tuning two well-known and well performing models for abstractive summarization: BART-Large (Lewis et al., 2020) and T5 (Raffel et al., 2020). BART and T5 are both transformer based, pre-trained models for sequence-to-sequence tasks. BART-Large has 406M parameters, T5 has 220M parameters. BART is pre-trained as a denoising autoencoder, where the task is to correctly reconstruct corrupted text. T5 is trained in a multi-task scenario, on self-supervised and fully supervised tasks. As summarization of the original CNN/DM dataset was part of T5's pre-training objective, we only add T5 here for reference and completeness. We perform our final human evaluation only on the summary triples that are constructed with BART-Large as its abstractive component. We fine-tune and validate BART-Large and T5 on our train and validation splits for the CNN/DM dataset (296,444 and 15,100 documents respectively). We

Table 4.5: Rouge scores of the summarization components on our test set; scores on the validation set are included in Appendix 4.C.1, Table 4.C.1.

Model	R1	R2	RL	RLsum
BART-Large	48.16	22.81	32.84	44.49
T5	46.63	22.41	32.97	43.43

use the standard setup from the Hugging Face library[7] (Wolf et al., 2020) and fine-tune both models for 3 epochs on 4 GPUs with 12GB of RAM each.

Intermediate results for the abstractive summarization components

We report inference scores on the abstracts in our GraphelSums test set in Table 4.5 and confirm that our summarization models score on par with the scores that are reported on the original CNN/DM test set. Interestingly, T5 performs worse than BART-Large, despite having summarization as part of its pre-training objective.

4.5.2 Summary Triples Component

The summaries that we have generated so far satisfy the *abstractive* objective, but are still purely textual, i.e., they do not contain the graphical elements yet. This is what we focus on next. We propose two approaches to generate the final summary triples, that both take the abstractive summaries from the previous step as input: (i) generating summary triples with Snorkel, and (ii) trained information extraction. We discuss each of these approaches below.

Generating summary triples with Snorkel

In Section 4.2.3 we discussed how Snorkel (Ratner et al., 2020) can be used as a means to obtain probabilistic labels for unsupervised data, by creating fuzzy labeling functions. We use Snorkel to generate summary triples, and use the obtained triples for two purposes: (i) to extract relations directly on the output of the summarization models, and (ii) to use as weak labels for training a relation extraction model (Section 4.5.2). Our Snorkel pipeline consists of

[7] `https://github.com/huggingface/transformers`

three stages: (i) candidate pairs selection, (ii) relation labeling, and (iii) final filtering. We discuss each of these steps in more detail below.

(1) *Candidate pairs selection.* In this step our objective is to find the phrase pairs in a text that can potentially be part of a summary triple, if linked by a relation. We use the Berkeley constituency parser (Kitaev and Klein, 2018) as a fast and high quality parser to obtain all constituents in a summary abstract. Next, we combine the constituents to make potential candidate pairs. Naively, one could simply combine all constituents that are found for an abstract. However, due to its quadratic complexity, this approach is very resource demanding. Instead, we make use of a few heuristics to make the candidate pairs. First, we discard all single token constituents that are of type IN, DT or CC, as these would not lead to valid candidate pairs. For the same reason, we also discard all constituents that only consist of a special token, such as a punctuation mark. Moreover, we make sure to not pair overlapping constituents. Finally, we set a threshold that constituent pairs can be at most two sentences apart. With this heuristic, we miss a small fraction of correct candidate pairs, yet it considerably speeds up the computation.

(2) *Relation labeling.* In the second step, we aim to find the relations that link the candidate pairs that we found in the previous step. Note that not all candidate pairs can and will be linked. For each possible relation, we construct Snorkel labeling functions. We use AllenNLP's[8] NER-tagger[9] and implementation of SpanBERT[10] (Joshi et al., 2020) to obtain NER-tags and coreference relations for our corpus, which we use in our labeling functions. For example, in one of our labeling functions for the WHO relation, we check whether the two phrases in a candidate pair are both referring to the same referent and whether one of the phrases is tagged with the PERSON NER-tag. A full overview of all labeling functions can be found in our code.

(3) *Final filtering.* As a last step, we apply filtering to obtain the final set of weakly labeled data points. First, we determine the edge directions. Let

[8] https://allenai.org/allennlp

[9] https://storage.googleapis.com/allennlp-public-models/ner-model-2020.02.10.tar.gz

[10] https://github.com/allenai/allennlp-models/blob/main/allennlp_models/modelcards/coref-spanbert.json

us take constituents connected by the WHEN relation as an example. In these cases, the constituent that indicates the time should be the second constituent in the summary triple. Next, we filter out overlapping constituents. For example, imagine three possible triples: *(event,* WHEN*, June 7)*, *(event,* WHEN*, 2012)*, *(event,* WHEN*, June 7, 2012)*. In this case, we make sure that only one of these potential data points is included in the weakly labeled training set. As a heuristic, we choose to include the data point with the longest string, arguing that this provides us with most information. Finally, we merge all coreferences and use the first occurrence of the referent in our summary triples, corresponding to the instructions for our human annotators.

Trained Relation Extraction

We choose DyGIE++ (Wadden et al., 2019) as our trained relation extraction method. DyGIE++ has shown to be a well-performing method across several tasks and datasets, it is well documented and can be easily adapted to our task. We use the code as provided,[11] but make a number of adaptations. Firstly, we adapt the DyGIE++ implementation to allow cross-sentence relations, instead of only within-sentence relations. Concretely, that means that we treat the entire document as if it were a single sentence. We leave all punctuation marks in. We train DyGIE++ for 20 epochs on 4 GPUs with 12GB RAM each. A cross-sentence approach is substantially more memory demanding than a within sentence approach. Hence, we resort to training DyGIE++ on $10,000$ summary abstracts from the training set, the maximum number of abstracts we could process with our machines.[12] Secondly, we evaluate the extracted relations on the same metrics that we used to assess the human annotations.

4.5.3 Evaluation Baseline Methods

In this section we evaluate our baseline models with an automatic evaluation (Section 4.5.3), a human evaluation (Section 4.5.3), and a qualitative analysis (Section 4.5.3).

[11] `https://github.com/dwadden/dygiepp`

[12] We also found that results did not substantially improve by adding more training data.

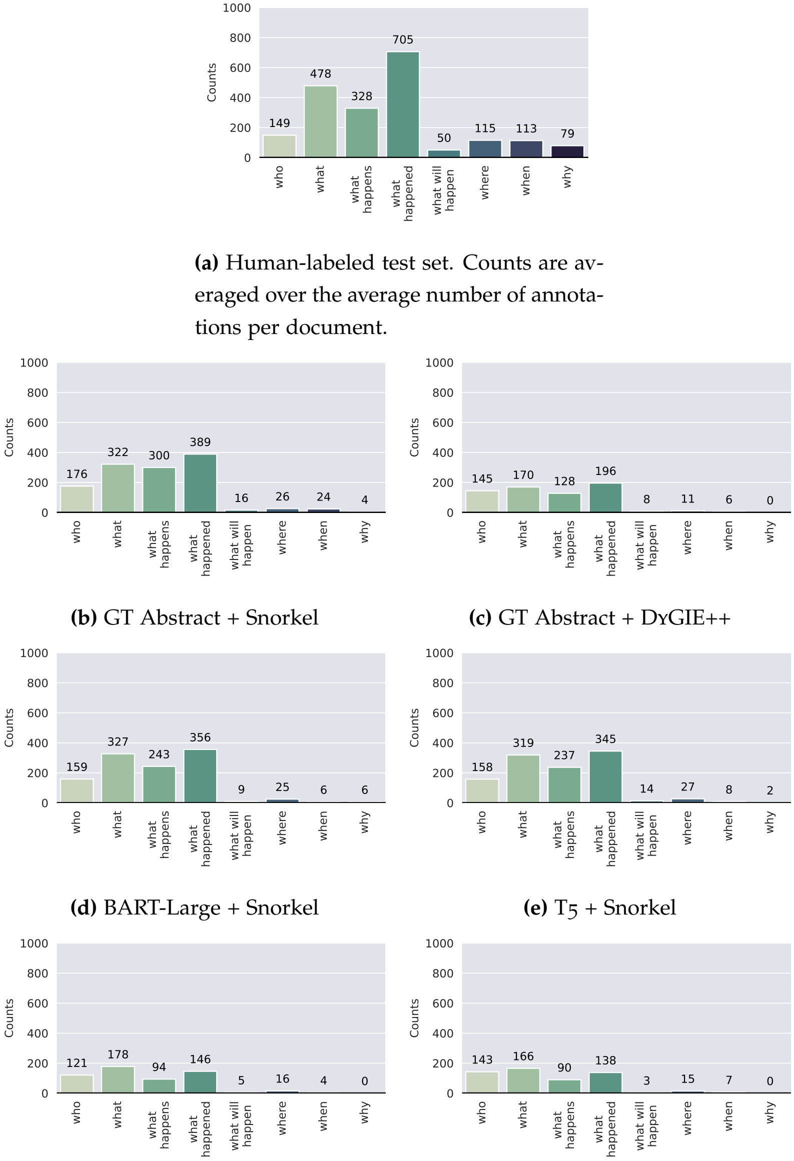
Automatic Evaluation

Here we evaluate which baseline model performs best on our task. First, we evaluate Snorkel and DYGIE++ with the ground truth abstracts as input. Next, we combine the abstractive summarization and the summary triples component. We generate summaries with BART-Large and T5 and feed them to Snorkel and DYGIE++. For evaluation, we use the best scoring ground truth to compute the final scores.[13] We evaluate on the same metrics as used for the quality control of the human labels, as defined in Section 4.4.2: (i) hard pairwise macro F_1-scores (Table 4.6), (ii) soft greedy pairwise macro F_1-scores (Table 4.6), and (iii) pairwise Jaccard scores (Table 4.7). In this section, we also report the *precision* and *recall* scores for the hard evaluation in Table 4.6, which we compute in a similar manner as the hard pairwise macro F_1-scores. Finally, there are different ways in how one can account for the intuition that the hard matching scores are too conservative. So far, we have used a soft metric where we greedily matched summary triples. For completeness, we add one additional soft scoring metric, to complement the soft greedy pairwise macro F_1-scores. For this metric, we score whether or not predicted summary triples have any lexical overlap with the ground truth triples. Based on these scores we compute the precision, recall and F_1-scores (Table 4.8).

Generally, we find that the scores are still far from the human agreement scores, indicating that our task is challenging and that there are many opportunities for future work. More specifically, the settings where we use Snorkel directly perform better than relation extraction with DYGIE++. Moreover, even though the summaries produced by BART-Large and T5 are of high quality, the additional summarization step decreases the performance substantially. We postulate that the BART-Large/T5-generated summaries are still quite different from the human-written summaries, therefore decreasing the performance of models trained on labels for the human-written summaries.

We also inspect the predicted relations in more detail. In Figure 4.2 we share how often each relation occurs in each baseline. Figure 4.C.1 in Appendix 4.C shows the same, yet measured in percentages. Our findings are in line with the scores in Table 4.6; settings with lower recall scores predict fewer relations. The ratios of the predicted relations are comparable across settings, but differ more for settings with lower scores.

[13] Some documents could not be parsed in the Snorkel pipeline and we leave these out.



(a) Human-labeled test set. Counts are averaged over the average number of annotations per document.

(b) GT Abstract + Snorkel

(c) GT Abstract + DyGIE++

(d) BART-Large + Snorkel

(e) T5 + Snorkel

(f) BART-Large + DyGIE++

(g) T5 + DyGIE++

Figure 4.2: Histograms of relation counts for the baselines.

Table 4.6: Precision, recall and F_1-scores for different methods on our task of generating summaries with graphical elements. In this table we report hard scores and the greedy soft F_1-score.

	Hard P	**Hard R**	**Hard F1**	**Soft Greedy F1**
GT Abstract + Snorkel	$0.130_{\pm 0.165}$	$0.102_{\pm 0.136}$	$0.111_{\pm 0.141}$	$0.446_{\pm 0.121}$
GT Abstract + DyGIE++	$0.071_{\pm 0.157}$	$0.037_{\pm 0.084}$	$0.046_{\pm 0.098}$	$0.256_{\pm 0.145}$
BART-Large + Snorkel	$0.003_{\pm 0.024}$	$0.002_{\pm 0.013}$	$0.002_{\pm 0.017}$	$0.323_{\pm 0.087}$
T5 + Snorkel	$0.008_{\pm 0.056}$	$0.004_{\pm 0.027}$	$0.005_{\pm 0.035}$	$0.329_{\pm 0.086}$
BART-Large + DyGIE++	$0.001_{\pm 0.015}$	$0.00_{\pm 0.007}$	$0.001_{\pm 0.009}$	$0.188_{\pm 0.120}$
T5 + DyGIE++	$0.003_{\pm 0.033}$	$0.001_{\pm 0.009}$	$0.001_{\pm 0.014}$	$0.184_{\pm 0.115}$

Table 4.7: Jaccard scores for different methods on our task of generating summaries with graphical elements.

	Triple	**Const A**	**Const B**	**Relation**
GT Abstract + Snorkel	$0.065_{\pm 0.089}$	$0.238_{\pm 0.175}$	$0.19_{\pm 0.148}$	$0.582_{\pm 0.256}$
GT Abstract + DyGIE++	$0.026_{\pm 0.057}$	$0.161_{\pm 0.204}$	$0.095_{\pm 0.103}$	$0.409_{\pm 0.262}$
BART-Large + Snorkel	$0.001_{\pm 0.009}$	$0.086_{\pm 0.132}$	$0.016_{\pm 0.041}$	$0.535_{\pm 0.221}$
T5 + Snorkel	$0.003_{\pm 0.021}$	$0.091_{\pm 0.134}$	$0.017_{\pm 0.047}$	$0.543_{\pm 0.224}$
BART-Large + DyGIE++	$0.000_{\pm 0.005}$	$0.056_{\pm 0.126}$	$0.017_{\pm 0.047}$	$0.352_{\pm 0.270}$
T5 + DyGIE++	$0.001_{\pm 0.008}$	$0.065_{\pm 0.132}$	$0.014_{\pm 0.045}$	$0.345_{\pm 0.259}$

Human Evaluation

We also evaluate our baseline methods using a human evaluation, with a similar setup as in Section 4.3.3. We compare three settings: (i) the human-labeled ground truth, (ii) BART-Large output followed by Snorkel labels, and (iii) BART-Large output followed by DieGIE++ labels. We compare summaries for 15 different articles and we request 3 annotations per HIT to be able to compute the majority vote afterwards. We select our summaries randomly, but filter out summaries with potentially sensitive topics for crowd workers. We randomly select one of the annotations per summary. Participants do not know which summary is generated by which setting.

In addition to the questions asked in the first human evaluation, we ask which summaries workers find more *informative* and which ones they find

Table 4.8: Soft binary precision, recall and F_1-scores for different methods on our task of generating summaries with graphical elements.

	Soft Binary		
	P	**R**	**F1**
GT Abstract + Snorkel	$0.342_{\pm 0.241}$	$0.262_{\pm 0.200}$	$0.286_{\pm 0.202}$
GT Abstract + DyGIE++	$0.237_{\pm 0.303}$	$0.102_{\pm 0.139}$	$0.132_{\pm 0.162}$
BART-Large + Snorkel	$0.239_{\pm 0.219}$	$0.173_{\pm 0.164}$	$0.191_{\pm 0.170}$
T5 + Snorkel	$0.250_{\pm 0.247}$	$0.179_{\pm 0.195}$	$0.198_{\pm 0.194}$
BART-Large + DyGIE++	$0.145_{\pm 0.268}$	$0.056_{\pm 0.101}$	$0.075_{\pm 0.129}$
T5 + DyGIE++	$0.137_{\pm 0.261}$	$0.055_{\pm 0.102}$	$0.072_{\pm 0.128}$

more *concise*, in line with previous work (e.g., Paulus et al., 2018; Narayan et al., 2018a). We do not evaluate on metrics like fluency, as the nature of our summaries with graphical elements does not align with this metric. As an additional quality control question, we ask workers to indicate their favorite summary and to provide a short justification for their choice. As in Section 4.3.3, we obtain new annotations to replace rejected HITs and we apply some filtering to ensure good quality judgments. This leaves us with 134 annotations. An example of the task is given in Appendix 4.C.3, Figure 4.B.1.

The results, based on majority votes, are given in Table 4.9 and 4.10. These results show that crowd workers prefer the human-annotated summaries on all metrics, followed by the version where we used BART-Large and Snorkel. Summaries with graphical elements produced by BART-Large and DyGIE++ are preferred least. Table 4.10 shows that these summaries often do not contain the answer to questions. These results are in line with the results on the automatic metrics.

The question in which we asked people to indicate their favorite summary was mostly intended as a control question. However, by reading participants' answers we find that their motivations closely match the quantitative results. Many comments are related to the amount of detail in the summary, and how that affects the informativeness of the summary. Most comments in this space argue that the favorite summary contains more information. For example, one participant preferred the human-labeled summary over the DyGIE++ summary because *"it contains much information that can be required to answer questions with-*

Table 4.9: Results human evaluation of baselines. Pairwise comparisons. Results (%) based on majority votes.

	Pair (A/B)	Prefer A	Prefer B
Inform	Human/Snorkel	80.0	20.0
	Human/DyGIE++	93.3	6.70
	Snorkel/DyGIE++	86.7	13.3
Cons	Human/Snorkel	60.0	40.0
	Human/DyGIE++	73.3	26.7
	Snorkel/DyGIE++	73.3	26.7
Prev	Human/Snorkel	73.3	26.7
	Human/DyGIE++	86.7	13.3
	Snorkel/DyGIE++	80.0	20.0
Sub	Human/Snorkel	86.7	13.3
	Human/DyGIE++	100.0	0.00
	Snorkel/DyGIE++	86.7	13.3
Fav	Human/Snorkel	86.7	13.3
	Human/DyGIE++	93.3	6.7
	Snorkel/DyGIE++	86.7	13.3

Table 4.10: Results human evaluation of baselines. Which summary was used to answer the questions. *No ans* if none of the summaries included the answer. Pairwise comparisons. Aggregated scores. Results (%) based on majority votes.

Pair (A/B)	Use A	Use B	No ans
Human/Snorkel	73.3	20.0	6.7
Human/DyGIE++	73.3	13.3	13.3
Snorkel/DyGIE++	26.7	20.0	53.3

out referring to the main article." Another participant chose the human-labeled summary as their favorite, over the Snorkel generated summary, because it "*describes the article well and comprehensive. And also it was understandable, easy to read.. It gives the elaborate details."* Another participant preferred the Snorkel generated summary over the DyGIE++ summary, because the latter "*has almost no information and barely makes sense."* This comment also fits in a category of comments that mention that some of the summaries that are generated by the baselines are still of low quality. For example, one participant compared a DyGIE++ summary with a human-labeled summary and found the DyGIE++ summary "*disjointed. "Expanded his wine to water narrative —> What –>nonprofit" makes no sense in almost any context*". Another line of comments discussed how easy it was to understand the different summaries. For example, one participant found the human-labeled summary "*a bit easier to understand and read to get an image of what is being asked*" than the Snorkel summary, and another participant preferred the human-labeled summary over the Snorkel summary, because "*it provides a coherent and understandable summary."* Finally, one participant preferred the human-labeled summary over the DyGIE++ summary, because it "*gives a better and clearer picture of the event that ensued. You will understand and get the full story [...] without reading the full story."*

Qualitative Observations

We also manually inspect the outputs of the different methods that we compared in the human evaluation. First, we note that DyGIE++ still misses many relations. This is in line with the automatic scores for recall and the human evaluation. An example is given in Appendix 4.C.3, Figure 4.C.3c. Second, both settings still miss coreferences, resulting in summaries that are less well-connected than their human-labeled counterparts. An example is given in Appendix 4.C.3, Figure 4.C.4b.

4.6 DISCUSSION AND CONCLUSION

In this chapter we addressed the third research question of this thesis as we proposed a new task: *summarization with graphical elements*. We collected a high-quality human-labeled dataset, GraphelSums, for the task and presented the experimental results for a number of baselines. By means of automatic and human evaluations, we showed that our task is a much wanted, feasible, and challenging addition to the existing types of automatic summarization tasks. As such, our work can inspire a lot of follow-up work in this direction.

There are still some limitations of our work that we would like to address in future work. For example, as discussed in Section 4.4, we have used data of a single domain (news), in a single language (English). Although we believe this to be a good start and detailed how our choice was motivated by our requirements, different languages and domains need to be investigated in future research. This can be achieved by recruiting annotators with different skill sets, e.g., annotators who are experts in different domains, or who are fluent in different languages than English. Additionally, biases that are present in the original CNN/DM dataset will likely be present in our GraphelSums dataset as well. It is also to be noted that the CNN/DM dataset was collected in 2015, hence very recent news articles are not yet included. Moreover, given the intensity of the labeling task (recall from Section 4.4 that annotators spent approximately 7 minutes on a single annotation), we have been able to collect labels for a test set only. This limits the types of methods that can be used for our task.

For our human evaluations, we have manually constructed graphical representations of summary triples. As mentioned in Section 4.5.3, the precise lay-out of the summaries may affect people's judgments. Therefore, the results of the human evaluations should be taken as an indication of people's preferences. A detailed investigation of the effect of different designs is not within the scope of this chapter, but we encourage more design driven research into this question. Finally, we have focused on a relatively general user group, but future work with specific use-cases of summaries with graphical elements may target very specific groups of users. In that case, it is important to explicitly include these users in the evaluation.

For future work we also plan to explore graph-based summarization methods to directly learn the summarization triples. Moreover, we see many opportunities to investigate model understanding with our task. As mentioned in Section 4.2.1, current automatic summarization models have difficulties with factual consistency and being able to generate correct summary triples, including correct relations, may require an additional level of factual consistency.

So far we have focused on English as a language for our modeling efforts. From a user-centered perspective this is limited. In the next chapter we shift our focus to a wide variety of languages, and we specifically investigate how the focus on high-resource languages, like English, can bias the results for languages with fewer resources.

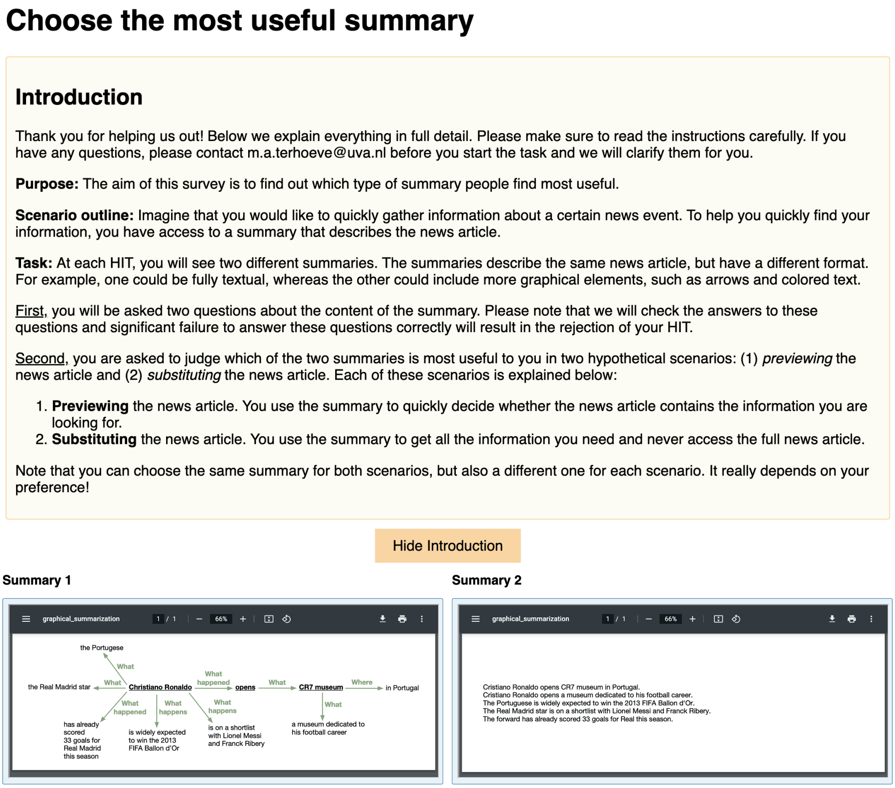
CHAPTER APPENDIX

4.A STEP 1 – DESIGNING THE TASK

In this section we give additional details of the initial human evaluation we performed in the first step of our investigation. We give examples of the task setup in the form of screenshots in Section 4.A.1. In Section 4.A.2 we list all open control questions that we asked during the evaluation. We also give additional results in Section 4.A.3.

4.A.1 *Examples of Task Setup First Human Evaluation*

Figure 4.A.1 shows screenshots of the human evaluation task setup where we investigate whether a critical mass of people is interested in summaries with graphical elements.



Choose the most useful summary

Introduction

Thank you for helping us out! Below we explain everything in full detail. Please make sure to read the instructions carefully. If you have any questions, please contact m.a.terhoeve@uva.nl before you start the task and we will clarify them for you.

Purpose: The aim of this survey is to find out which type of summary people find most useful.

Scenario outline: Imagine that you would like to quickly gather information about a certain news event. To help you quickly find your information, you have access to a summary that describes the news article.

Task: At each HIT, you will see two different summaries. The summaries describe the same news article, but have a different format. For example, one could be fully textual, whereas the other could include more graphical elements, such as arrows and colored text.

First, you will be asked two questions about the content of the summary. Please note that we will check the answers to these questions and significant failure to answer these questions correctly will result in the rejection of your HIT.

Second, you are asked to judge which of the two summaries is most useful to you in two hypothetical scenarios: (1) *previewing* the news article and (2) *substituting* the news article. Each of these scenarios is explained below:

1. **Previewing** the news article. You use the summary to quickly decide whether the news article contains the information you are looking for.
2. **Substituting** the news article. You use the summary to get all the information you need and never access the full news article.

Note that you can choose the same summary for both scenarios, but also a different one for each scenario. It really depends on your preference!

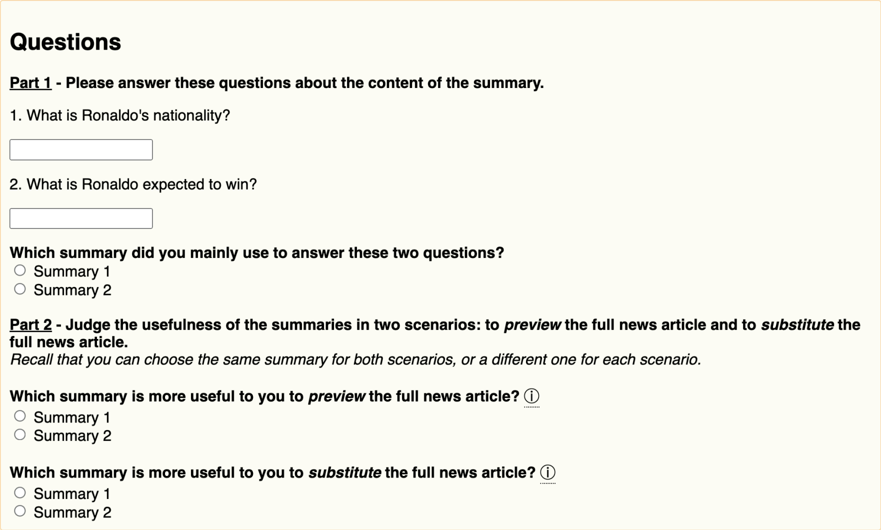
Summary 1

graphical_summarization
1 / 1
66%
the Portugese
What
the Real Madrid star
What
Christiano Ronaldo
What happened
opens
What
CR7 museum
Where
in Portugal
What happened
What happens
What happens
What
has already scored 33 goals for Real Madrid this season
is widely expected to win the 2013 FIFA Ballon d'Or
is on a shortlist with Lionel Messi and Franck Ribery
a museum dedicated to his football career

Summary 2

Cristiano Ronaldo opens CR7 museum in Portugal.
Cristiano Ronaldo opens a museum dedicated to his football career.
The Portuguese is widely expected to win the 2013 FIFA Ballon d'Or.
The Real Madrid star is on a shortlist with Lionel Messi and Franck Ribery.
The forward has already scored 33 goals for Real this season.

Figure 4.A.1: Example of human evaluation task in Step 1 – Example continues on the next page.



Questions

Part 1 - Please answer these questions about the content of the summary.

1. What is Ronaldo's nationality?

2. What is Ronaldo expected to win?

Which summary did you mainly use to answer these two questions?
○ Summary 1
○ Summary 2

Part 2 - Judge the usefulness of the summaries in two scenarios: to *preview* the full news article and to *substitute* the full news article.
Recall that you can choose the same summary for both scenarios, or a different one for each scenario.

Which summary is more useful to you to *preview* the full news article? ⓘ
○ Summary 1
○ Summary 2

Which summary is more useful to you to *substitute* the full news article? ⓘ
○ Summary 1
○ Summary 2

Figure 4.A.1: Example of human evaluation task in Step 1 – Continued from previous page. These are three screenshots. Summary 1 and 2 have a larger font in the actual task interface, so they are more easily readable for workers.

4.A.2 *List of Open Questions First Human Evaluation*

Document 1

1. How expensive was the replica of the ark?
2. What is the name of the carpenter?
3. Where is the replica of the ark?
4. How long did Johan Huyberts spend on building the ark?
5. What did the carpenter dream?
6. What can visitors do in the ark?

Document 2

1. What is Ronaldo's nationality?
2. What is Ronaldo expected to win?
3. Where did Ronaldo open a museum?
4. How many goals has Ronaldo scored for Real Madrid this season?
5. What is the CR7 museum?
6. Who are on the short list together with Ronaldo?

4.A.3 Additional Results First Human Evaluation

In Table 4.A.1 and 4.A.2 we give the results of the human evaluation per document.

Table 4.A.1: Results Human Evaluation. Pairwise comparisons. Results for Document 1.

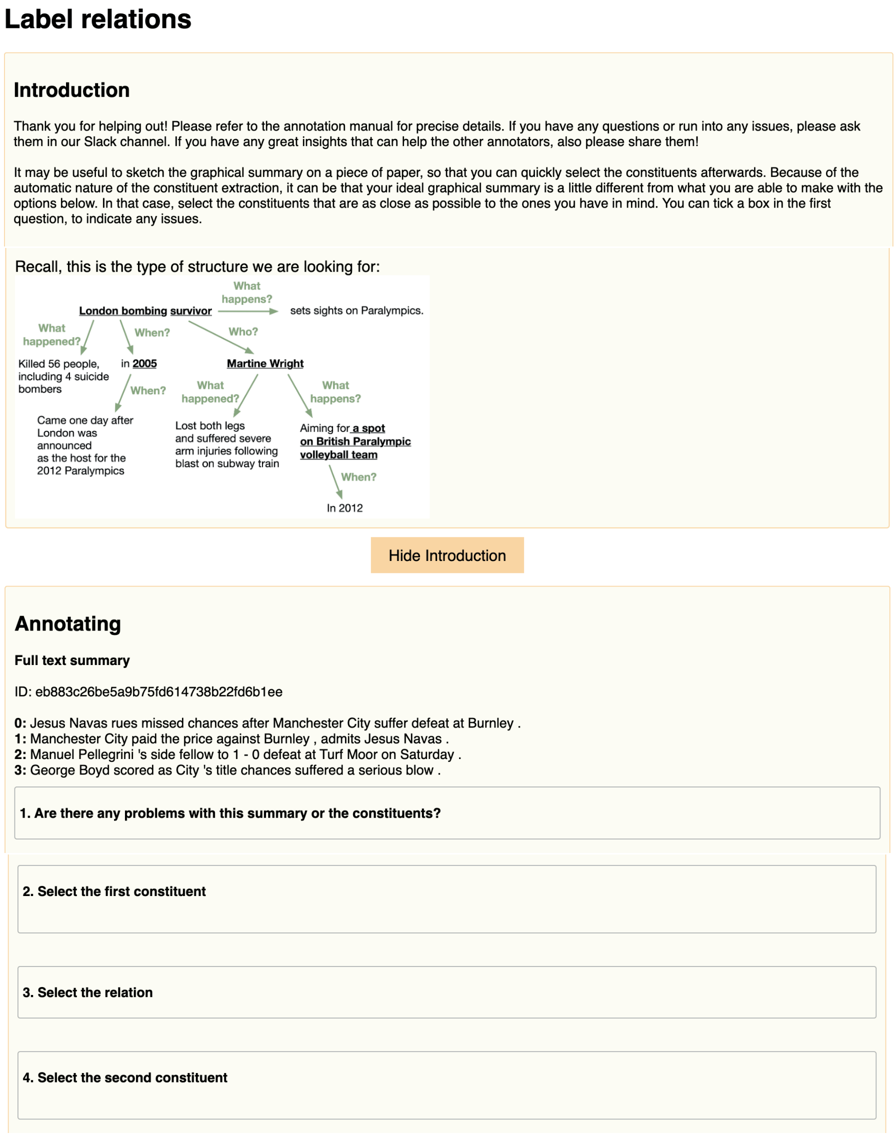
	Pair (A/B)	Prefer A (%)	Prefer B (%)
Used	Graphical / Text	40.0	60.0
	Graphical / Typeset	50.0	50.0
	Typeset / Text	85.0	15.0
Prev	Graphical / Text	35.0	65.0
	Graphical / Typeset	55.0	45.0
	Typeset / Text	90.0	10.0
Sub	Graphical / Text	20.0	80.0
	Graphical / Typeset	30.0	70.0
	Typeset / Text	75.0	25.0

Table 4.A.2: Results Human Evaluation. Pairwise comparisons. Results for Document 2.

	Pair (A/B)	Prefer A (%)	Prefer B (%)
Used	Graphical / Text	30.0	70.0
	Graphical / Typeset	50.0	50.0
	Typeset / Text	60.0	40.0
Prev	Graphical / Text	35.0	65.0
	Graphical / Typeset	40.0	60.0
	Typeset / Text	60.0	40.0
Sub	Graphical / Text	40.0	60.0
	Graphical / Typeset	35.0	65.0
	Typeset / Text	55.0	45.0

4.B STEP 2 – COLLECTING THE DATA

Figure 4.B.1 and 4.B.2 show screenshots of the human labeling task.



Label relations

Introduction

Thank you for helping out! Please refer to the annotation manual for precise details. If you have any questions or run into any issues, please ask them in our Slack channel. If you have any great insights that can help the other annotators, also please share them!

It may be useful to sketch the graphical summary on a piece of paper, so that you can quickly select the constituents afterwards. Because of the automatic nature of the constituent extraction, it can be that your ideal graphical summary is a little different from what you are able to make with the options below. In that case, select the constituents that are as close as possible to the ones you have in mind. You can tick a box in the first question, to indicate any issues.

Recall, this is the type of structure we are looking for:

London bombing survivor
What happens?
sets sights on Paralympics.
What happened?
When?
Who?
Killed 56 people, including 4 suicide bombers
in 2005
Martine Wright
When?
What happened?
What happens?
Came one day after London was announced as the host for the 2012 Paralympics
Lost both legs and suffered severe arm injuries following blast on subway train
Aiming for a spot on British Paralympic volleyball team
When?
In 2012
Hide Introduction

Annotating

Full text summary

ID: eb883c26be5a9b75fd614738b22fd6b1ee

0: Jesus Navas rues missed chances after Manchester City suffer defeat at Burnley .
1: Manchester City paid the price against Burnley , admits Jesus Navas .
2: Manuel Pellegrini 's side fellow to 1 - 0 defeat at Turf Moor on Saturday .
3: George Boyd scored as City 's title chances suffered a serious blow .

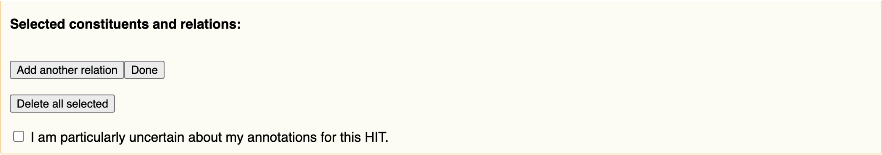
1. Are there any problems with this summary or the constituents?

2. Select the first constituent

3. Select the relation

4. Select the second constituent

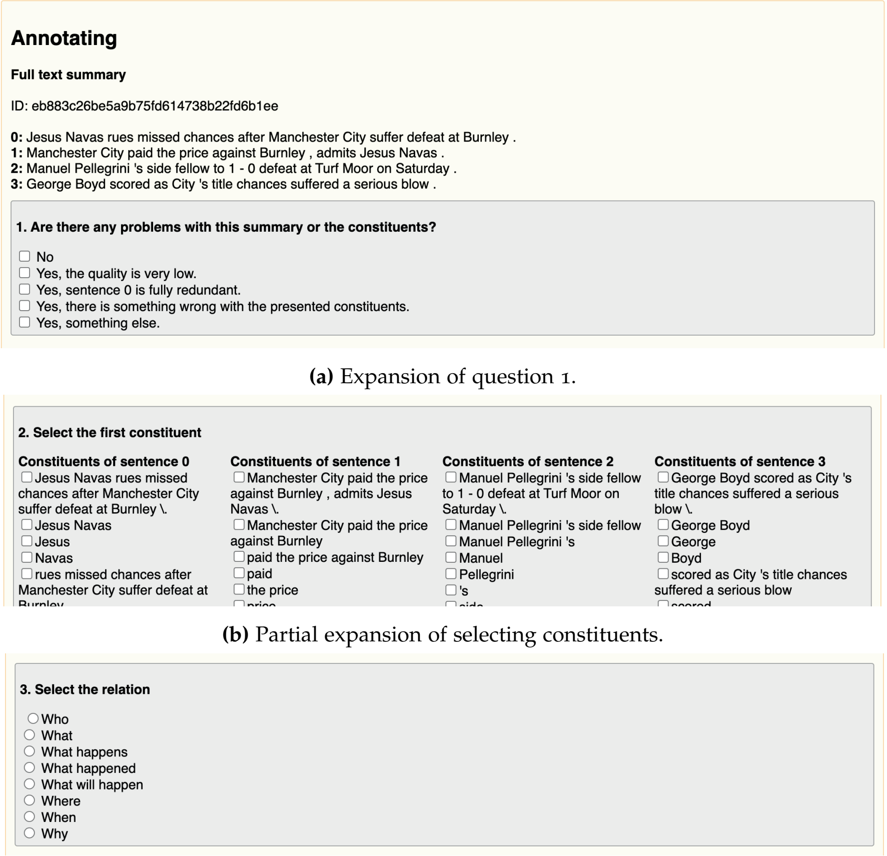
Figure 4.B.1: Overview of human labeling task – Example continues on the next page.



Selected constituents and relations:

□ I am particularly uncertain about my annotations for this HIT.

Figure 4.B.1: Overview of human labeling task – Example continued from previous page. Figure consists of several screenshots of the task.



Annotating

Full text summary

ID: eb883c26be5a9b75fd614738b22fd6b1ee

0: Jesus Navas rues missed chances after Manchester City suffer defeat at Burnley .
1: Manchester City paid the price against Burnley , admits Jesus Navas .
2: Manuel Pellegrini 's side fellow to 1 - 0 defeat at Turf Moor on Saturday .
3: George Boyd scored as City 's title chances suffered a serious blow .

1. Are there any problems with this summary or the constituents?

□ No
□ Yes, the quality is very low.
□ Yes, sentence 0 is fully redundant.
□ Yes, there is something wrong with the presented constituents.
□ Yes, something else.

(a) Expansion of question 1.

2. Select the first constituent

Constituents of sentence 0
□ Jesus Navas rues missed chances after Manchester City suffer defeat at Burnley \.
□ Jesus Navas
□ Jesus
□ Navas
□ rues missed chances after Manchester City suffer defeat at Burnley

Constituents of sentence 1
□ Manchester City paid the price against Burnley , admits Jesus Navas \.
□ Manchester City paid the price against Burnley
□ paid the price against Burnley
□ paid
□ the price
□ price

Constituents of sentence 2
□ Manuel Pellegrini 's side fellow to 1 - 0 defeat at Turf Moor on Saturday \.
□ Manuel Pellegrini 's side fellow
□ Manuel Pellegrini 's
□ Manuel
□ Pellegrini
□ 's
□ side

Constituents of sentence 3
□ George Boyd scored as City 's title chances suffered a serious blow \.
□ George Boyd
□ George
□ Boyd
□ scored as City 's title chances suffered a serious blow
□ scored

(b) Partial expansion of selecting constituents.

3. Select the relation

○ Who
○ What
○ What happens
○ What happened
○ What will happen
○ Where
○ When
○ Why

(c) Expansion of selecting relations.

Figure 4.B.2: Human labeling task in more detail.

4.C STEP 3 – BASELINES

In this section we provide additional results for the baseline experiments from the third part of our investigation.

4.C.1 Rouge Scores Validation Set

In Table 4.C.1 we give the ROUGE scores on the validation set for Bart-Large and T5.

Table 4.C.1: Rouge scores summarization components on the validation set.

Model	**R1**	**R2**	**RL**	**RLsum**
BART-Large	47.12	21.97	31.95	43.59
T5	46.70	22.01	32.63	43.41

4.C.2 Automatic Evaluation – Predicted Relations

In Figure 4.C.1 we show the percentages of predicted relations for different baselines.

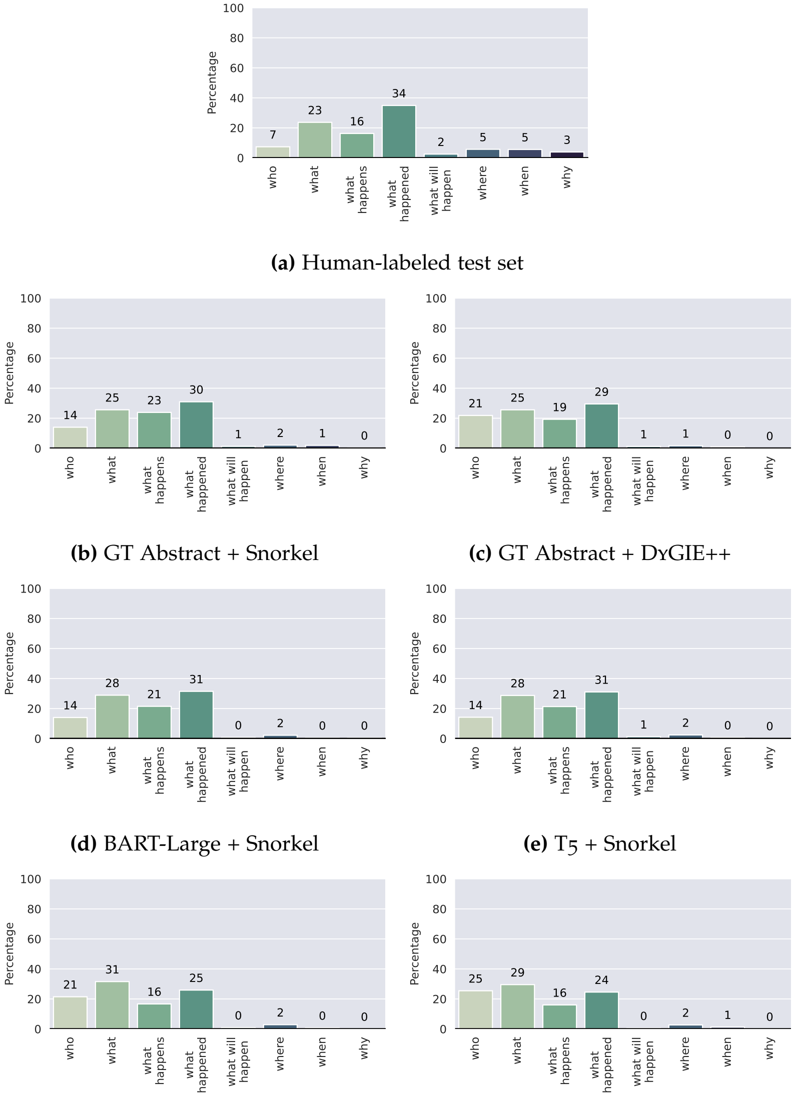
4.C.3 Human Evaluation of Baselines

Examples of Task Setup

In Figure 4.C.2 we show screenshots of the task setup for the human evaluation where we compare the results of different methods on our task.

Examples of Generated Summaries

Figure 4.C.3 and Figure 4.C.4 give examples of outputs of summaries with graphical elements, generated by different methods.



(a) Human-labeled test set

(b) GT Abstract + Snorkel

(c) GT Abstract + DyGIE++

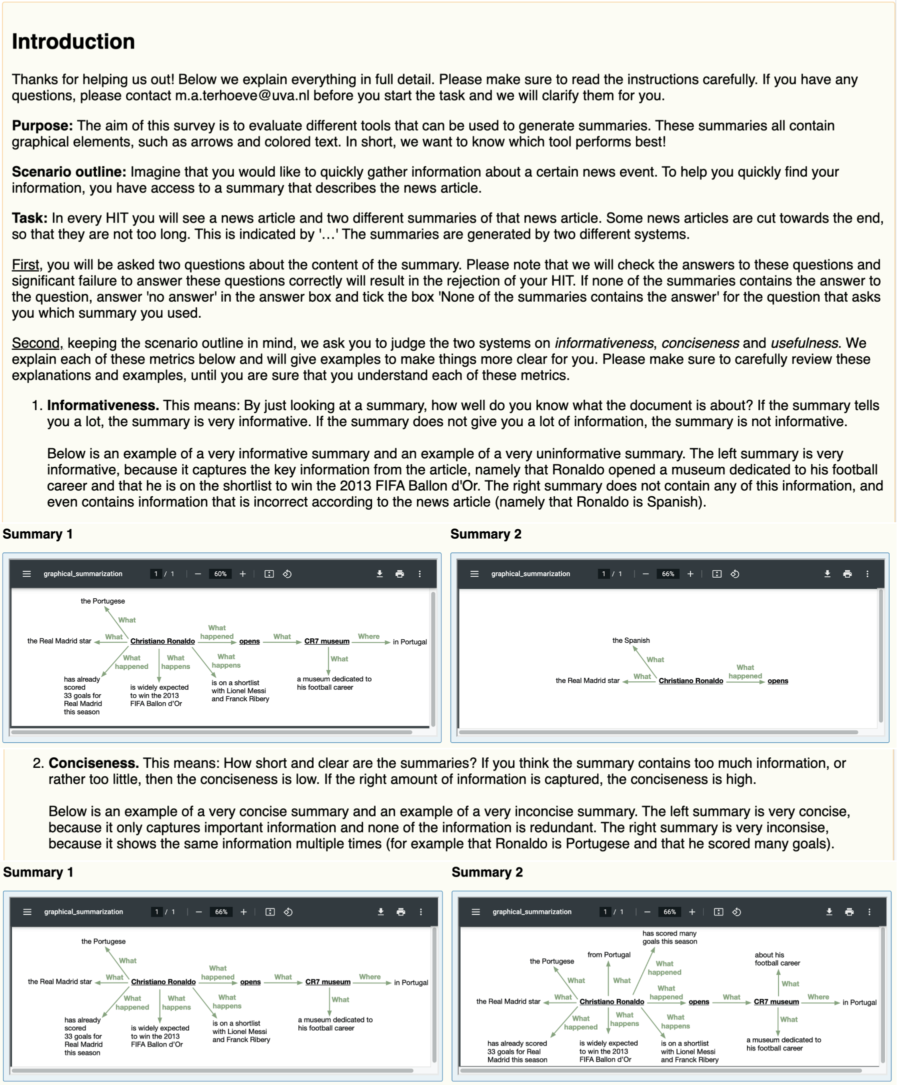
(d) BART-Large + Snorkel

(e) T5 + Snorkel

(f) BART-Large + DyGIE++

(g) T5 + DyGIE++

Figure 4.C.1: Histograms of relation percentages for the baselines.



Introduction

Thanks for helping us out! Below we explain everything in full detail. Please make sure to read the instructions carefully. If you have any questions, please contact m.a.terhoeve@uva.nl before you start the task and we will clarify them for you.

Purpose: The aim of this survey is to evaluate different tools that can be used to generate summaries. These summaries all contain graphical elements, such as arrows and colored text. In short, we want to know which tool performs best!

Scenario outline: Imagine that you would like to quickly gather information about a certain news event. To help you quickly find your information, you have access to a summary that describes the news article.

Task: In every HIT you will see a news article and two different summaries of that news article. Some news articles are cut towards the end, so that they are not too long. This is indicated by '…' The summaries are generated by two different systems.

First, you will be asked two questions about the content of the summary. Please note that we will check the answers to these questions and significant failure to answer these questions correctly will result in the rejection of your HIT. If none of the summaries contains the answer to the question, answer 'no answer' in the answer box and tick the box 'None of the summaries contains the answer' for the question that asks you which summary you used.

Second, keeping the scenario outline in mind, we ask you to judge the two systems on *informativeness*, *conciseness* and *usefulness*. We explain each of these metrics below and will give examples to make things more clear for you. Please make sure to carefully review these explanations and examples, until you are sure that you understand each of these metrics.

1. **Informativeness.** This means: By just looking at a summary, how well do you know what the document is about? If the summary tells you a lot, the summary is very informative. If the summary does not give you a lot of information, the summary is not informative.

   Below is an example of a very informative summary and an example of a very uninformative summary. The left summary is very informative, because it captures the key information from the article, namely that Ronaldo opened a museum dedicated to his football career and that he is on the shortlist to win the 2013 FIFA Ballon d'Or. The right summary does not contain any of this information, and even contains information that is incorrect according to the news article (namely that Ronaldo is Spanish).

Summary 1 **Summary 2**

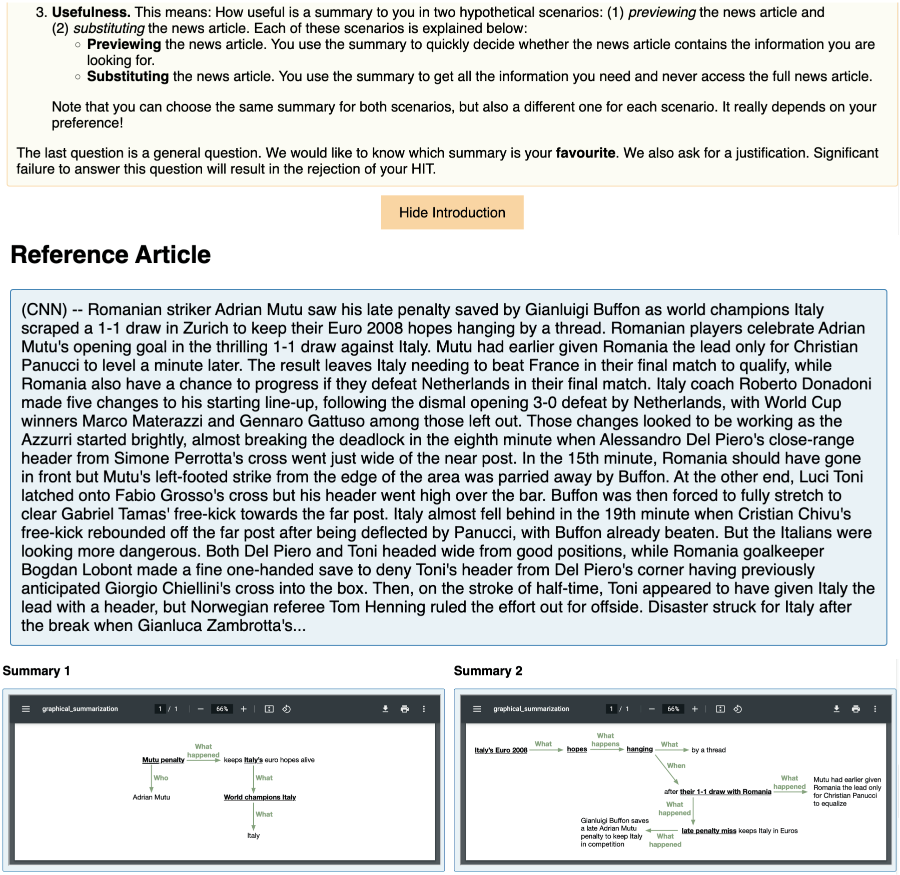
2. **Conciseness.** This means: How short and clear are the summaries? If you think the summary contains too much information, or rather too little, then the conciseness is low. If the right amount of information is captured, the conciseness is high.

   Below is an example of a very concise summary and an example of a very inconcise summary. The left summary is very concise, because it only captures important information and none of the information is redundant. The right summary is very inconsise, because it shows the same information multiple times (for example that Ronaldo is Portugese and that he scored many goals).

Summary 1 **Summary 2**

graphical_summarization
1 / 1
66%
the Portugese
What
the Real Madrid star
What
Christiano Ronaldo
What happened
opens
What
CR7 museum
Where
in Portugal
What happened
What happens
What happens
What
has already scored 33 goals for Real Madrid this season
is widely expected to win the 2013 FIFA Ballon d'Or
is on a shortlist with Lionel Messi and Franck Ribery
a museum dedicated to his football career
graphical_summarization
1 / 1
66%
has scored many goals this season
from Portugal
the Portugese
What
What
What happened
about his football career
What
the Real Madrid star
What
Christiano Ronaldo
What happened
opens
What
CR7 museum
Where
in Portugal
What happened
What happens
What happens
What
has already scored 33 goals for Real Madrid this season
is widely expected to win the 2013 FIFA Ballon d'Or
is on a shortlist with Lionel Messi and Franck Ribery
a museum dedicated to his football career

Figure 4.C.2: Example of human evaluation task in part 3 – Example continues on the next page.



3. **Usefulness.** This means: How useful is a summary to you in two hypothetical scenarios: (1) *previewing* the news article and (2) *substituting* the news article. Each of these scenarios is explained below:
   - ○ **Previewing** the news article. You use the summary to quickly decide whether the news article contains the information you are looking for.
   - ○ **Substituting** the news article. You use the summary to get all the information you need and never access the full news article.

   Note that you can choose the same summary for both scenarios, but also a different one for each scenario. It really depends on your preference!

The last question is a general question. We would like to know which summary is your **favourite**. We also ask for a justification. Significant failure to answer this question will result in the rejection of your HIT.

Hide Introduction

Reference Article

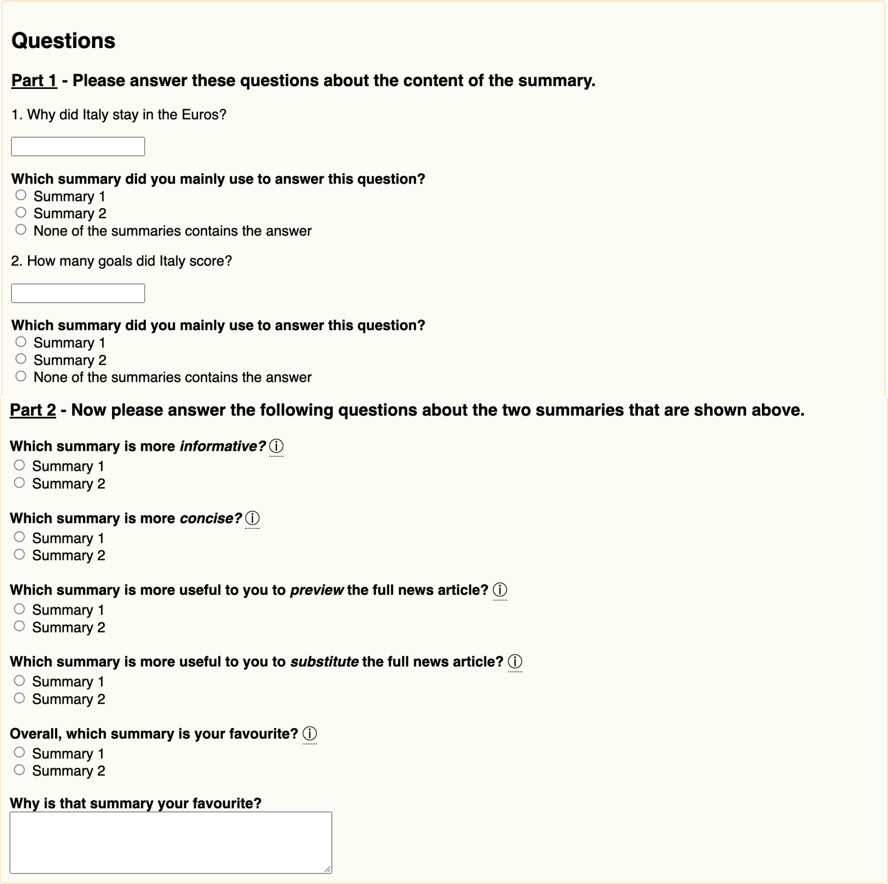
(CNN) -- Romanian striker Adrian Mutu saw his late penalty saved by Gianluigi Buffon as world champions Italy scraped a 1-1 draw in Zurich to keep their Euro 2008 hopes hanging by a thread. Romanian players celebrate Adrian Mutu's opening goal in the thrilling 1-1 draw against Italy. Mutu had earlier given Romania the lead only for Christian Panucci to level a minute later. The result leaves Italy needing to beat France in their final match to qualify, while Romania also have a chance to progress if they defeat Netherlands in their final match. Italy coach Roberto Donadoni made five changes to his starting line-up, following the dismal opening 3-0 defeat by Netherlands, with World Cup winners Marco Materazzi and Gennaro Gattuso among those left out. Those changes looked to be working as the Azzurri started brightly, almost breaking the deadlock in the eighth minute when Alessandro Del Piero's close-range header from Simone Perrotta's cross went just wide of the near post. In the 15th minute, Romania should have gone in front but Mutu's left-footed strike from the edge of the area was parried away by Buffon. At the other end, Luci Toni latched onto Fabio Grosso's cross but his header went high over the bar. Buffon was then forced to fully stretch to clear Gabriel Tamas' free-kick towards the far post. Italy almost fell behind in the 19th minute when Cristian Chivu's free-kick rebounded off the far post after being deflected by Panucci, with Buffon already beaten. But the Italians were looking more dangerous. Both Del Piero and Toni headed wide from good positions, while Romania goalkeeper Bogdan Lobont made a fine one-handed save to deny Toni's header from Del Piero's corner having previously anticipated Giorgio Chiellini's cross into the box. Then, on the stroke of half-time, Toni appeared to have given Italy the lead with a header, but Norwegian referee Tom Henning ruled the effort out for offside. Disaster struck for Italy after the break when Gianluca Zambrotta's...

Summary 1

Summary 2

graphical_summarization
1 / 1
66%
Italy's Euro 2008
What
hopes
What happens
hanging
What
by a thread
When
after their 1-1 draw with Romania
What happened
Mutu had earlier given Romania the lead only for Christian Panucci to equalize
What happened
late penalty miss keeps Italy in Euros
What happened
Gianluigi Buffon saves a late Adrian Mutu penalty to keep Italy in competition

Figure 4.C.2: Example of human evaluation task in part 3 – Example continued from previous page and continues on the next page.



Questions

Part 1 - Please answer these questions about the content of the summary.

1. Why did Italy stay in the Euros?

Which summary did you mainly use to answer this question?
○ Summary 1
○ Summary 2
○ None of the summaries contains the answer

2. How many goals did Italy score?

Which summary did you mainly use to answer this question?
○ Summary 1
○ Summary 2
○ None of the summaries contains the answer

Part 2 - Now please answer the following questions about the two summaries that are shown above.

Which summary is more *informative?* ⓘ
○ Summary 1
○ Summary 2

Which summary is more *concise?* ⓘ
○ Summary 1
○ Summary 2

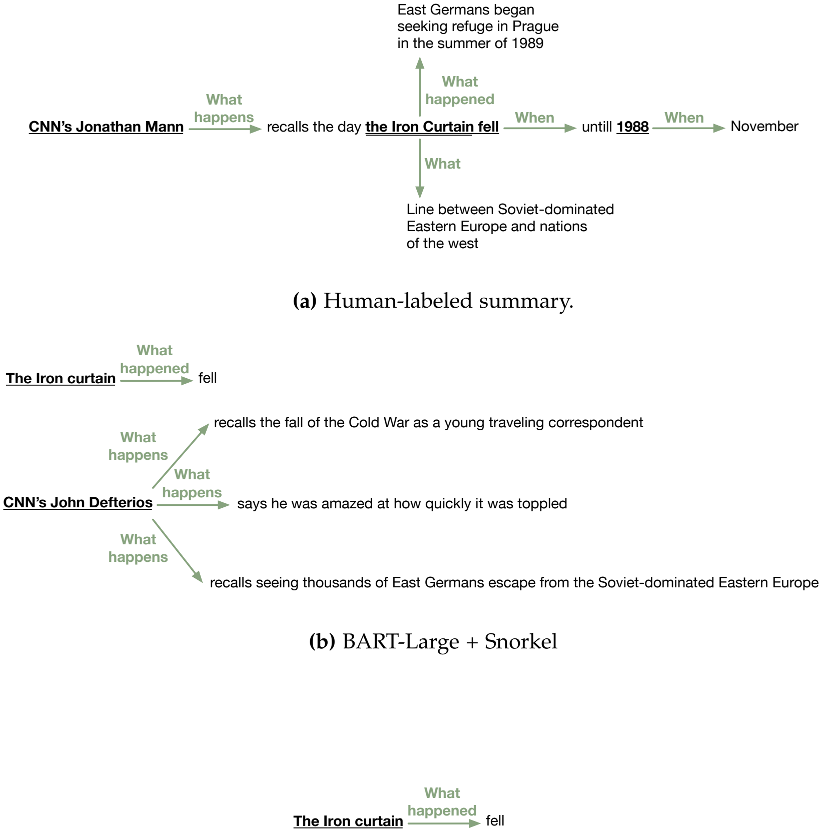
Which summary is more useful to you to *preview* the full news article? ⓘ
○ Summary 1
○ Summary 2

Which summary is more useful to you to *substitute* the full news article? ⓘ
○ Summary 1
○ Summary 2

Overall, which summary is your favourite? ⓘ
○ Summary 1
○ Summary 2

Why is that summary your favourite?

Figure 4.C.2: Example of human evaluation task in part 3. These are screenshots – Example continued from previous page. Summaries have a larger font in the actual task, so they are more readable for workers.

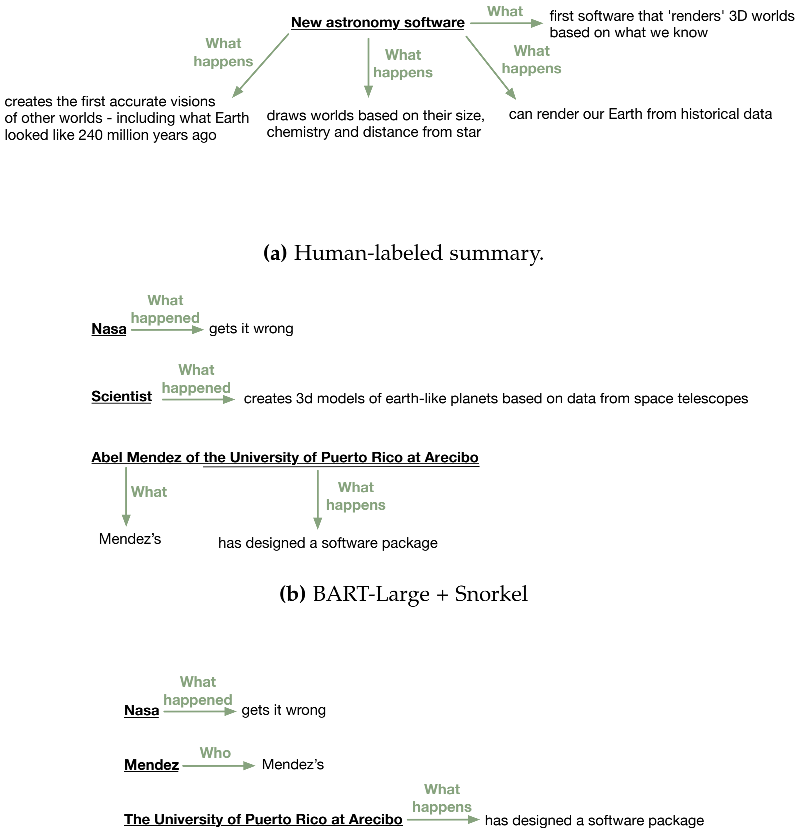

East Germans began seeking refuge in Prague in the summer of 1989
What happened
CNN's Jonathan Mann
What happens
recalls the day the Iron Curtain fell
When
untill 1988
When
November
What
Line between Soviet-dominated Eastern Europe and nations of the west
The Iron curtain
What happened
fell
recalls the fall of the Cold War as a young traveling correspondent
What happens
What happens
CNN's John Defterios
says he was amazed at how quickly it was toppled
What happens
recalls seeing thousands of East Germans escape from the Soviet-dominated Eastern Europe
The Iron curtain
What happened
fell

(a) Human-labeled summary.

(b) BART-Large + Snorkel

(c) BART-Large + DieGIE++

Figure 4.C.3: Examples of summaries with graphical elements generated by different methods.


New astronomy software
What
first software that 'renders' 3D worlds based on what we know
What happens
creates the first accurate visions of other worlds - including what Earth looked like 240 million years ago
What happens
draws worlds based on their size, chemistry and distance from star
What happens
can render our Earth from historical data
Nasa
What happened
gets it wrong
Scientist
What happened
creates 3d models of earth-like planets based on data from space telescopes
Abel Mendez of the University of Puerto Rico at Arecibo
What
Mendez's
What happens
has designed a software package
Nasa
What happened
gets it wrong
Mendez
Who
Mendez's
The University of Puerto Rico at Arecibo
What happens
has designed a software package

(a) Human-labeled summary.

(b) BART-Large + Snorkel

(c) BART-Large + DieGIE++

Figure 4.C.4: Examples of summaries with graphical elements generated by different methods.

5

HIGH-RESOURCE METHODOLOGICAL BIAS IN LOW-RESOURCE INVESTIGATIONS

So far we have focused on English as the language for our modeling efforts in this thesis. This limits the human-centered approach that we have been advocating throughout this thesis. In this chapter, we shift our focus to languages that have fewer available resources. The central bottleneck for low-resource NLP is typically regarded to be the quantity of accessible data, overlooking the contribution of data quality. This is particularly seen in the development and evaluation of low-resource systems via downsampling of high-resource language data. In this chapter[1] we investigate the validity of this approach, as we answer the fourth research question of this thesis:

Research Question 4: *How are low-resource investigations in NLP biased by high-resource approaches?*

In answering this research question, we specifically focus on two well-known NLP tasks for our empirical investigations: POS-tagging and machine translation. We show that downsampling from a high-resource language results in datasets with different properties than the low-resource datasets, impacting the model performance for both POS-tagging and machine translation. Based

[1] This chapter is based on (ter Hoeve et al., 2022a).

on these results we conclude that naive downsampling of datasets can result in a biased view of how well models trained on these downsampled datasets work in a low-resource scenario.

5.1 INTRODUCTION

The field of natural language processing (NLP) has experienced substantial progress over the last few years, with the introduction of neural sequence-to-sequence models (e.g., Kalchbrenner and Blunsom, 2013; Vaswani et al., 2017) and large, pre-trained transformer based language models (e.g., Devlin et al., 2019; Brown et al., 2020). Despite their impressive performance, these models require a lot of training resources, which are not always available. Approaches specifically targeted towards low-resource scenarios try to address this issue (e.g., Agić et al., 2016; Plank and Agić, 2018; Zhu et al., 2019; Bai et al., 2021). Resource scarcity manifests itself in various ways, such as a lack of compute power (e.g., Hedderich et al., 2020) or a lack of (labeled) training data (e.g., Adelani et al., 2021). In this work we focus on the latter.

Whether or when a scenario or language should be considered as 'low-resourced' has been a topic of debate (e.g., Bird, 2022). In this work, we add to this discussion by highlighting that many low-resource approaches have been grounded in high-resource scenarios, as has also been noted previously (e.g., Kann et al., 2020). This is problematic from a cultural or sociolinguistic perspective (e.g., Hämäläinen, 2021; Bird, 2022), as well as from a methodological perspective (e.g., Kann et al., 2020). Although both perspectives are arguably intertwined, we mostly focus on the latter in this work.

For example, a popular approach to develop and evaluate low-resource systems is to downsample uniformly from a high-resource language to simulate a low-resource scenario (e.g., Fadaee et al., 2017; Araabi and Monz, 2020; Chronopoulou et al., 2020; Ding et al., 2020; Kumar et al., 2021). The motivations for this setup are often justifiable, for example if used to investigate the effect of the dataset size, or because low-resource data is hard to obtain. However, we do believe that there are two potential issues with this downsampling approach that should be carefully considered.

First, a large dataset is potentially richer in content than a small dataset, for example in terms of the number of domains or styles. That is, the total vocabulary size is expected to be larger for a large dataset. When downsampling, this would cause a mismatch between the downsampled dataset and the real low-resource scenario, potentially affecting the scores on the task at hand.

Second, when collecting datasets, we are often faced with a quality-quantity trade-off. On one end of the spectrum we find examples of high quality, low-resource datasets, that are carefully constructed for a specific task (e.g., ter Hoeve et al., 2020; Adelani et al., 2021; ter Hoeve et al., 2022c). Obtaining high quality data points is costly, and thus, once the dataset size increases, a different trade-off has to be made (e.g., Caswell et al., 2020; Luccioni and Viviano, 2021).[2] The quality and usefulness of these large datasets stem from their size, but not necessarily from the quality of individual data points (Kreutzer et al., 2022). Downsampling from such a dataset can cause the obtained sample to be of lower quality than expected in a truly low-resource scenario. This also links our work to active learning (Cohn et al., 1996) and curriculum learning (Bengio et al., 2009), which focus on finding the most helpful data points during training.

More theoretically, we can summarize these two issues by taking a look at the estimation error that is optimized during training, typically in the form of a cross-entropy loss:

$$\mathcal{L}(\theta; D) = -\frac{1}{|D|} \sum_{y \in D} P_D(y) \log P_M(y|\theta), \tag{5.1}$$

in which D refers to the data, M to the model, y to the prediction and θ to the model parameters. Uniformly downsampling to the same size as the simulated low-resource dataset deals with the $1/|D|$ term, but it does not account for the fact that D itself is different in the low- and high-resource setting. This mismatch is also referred to as the *proxy fallacy* (Agić and Vulić, 2019).

In this work we investigate the effect of simulating a low-resource scenario by taking a uniform downsample from a high-resource setting in the context of two well-known NLP tasks: part-of-speech (POS)-tagging and machine translation (MT). We empirically find evidence for both issues raised above: (i) downsampling from a high-resource scenario increases the richness of the vocabulary of the sample, and (ii) the quality of the high-resource dataset is sometimes lower than the low-resource variant. Thus, our work serves as a reminder to be

[2] This trade-off can also affect the quality of the collected low-resource data in large multilingual datasets (Kreutzer et al., 2022).

careful when simulating low-resource scenarios by uniformly downsampling from a high-resource dataset.

5.2 RELATED WORK

Here, we discuss the definition of 'low-resource', low-resource approaches in NLP, and different training strategies.

5.2.1 *On the Definition of 'Low-Resource'*

Despite the amount of work on low-resource languages, or low-resource scenarios, it is hard to find a definition of when a scenario, or even a language, counts as low- or high-resource. It seems questionable to call a language low-resourced if it is spoken by millions of people who communicate in oral and/or written form in that language (e.g., Hämäläinen, 2021; Bird, 2022). In this work we follow the implicit definition as used in previous work (e.g., Zhu et al., 2019; Hedderich et al., 2021): we compare languages with different amounts of written data available, which is mainly indicated by the availability of the datasets that we use.

5.2.2 *Low-Resource Approaches in NLP*

With the recent surge of work on NLP systems that require a lot of resources (e.g., Devlin et al., 2019; Brown et al., 2020; Chowdhery et al., 2022), the question of designing systems that also work in a low-resource scenario has received a lot of attention. We refer to Hedderich et al. (2021) for a recent survey. Although there are many examples of approaches that ground themselves in a 'truly' low-resource scenario (e.g., Plank et al., 2016; Kann et al., 2020; Adelani et al., 2021), there are also many examples of approaches where assumptions are made that are more plausible in higher resource scenarios (e.g., Li et al., 2012; Gu et al., 2018; Ding et al., 2020; Liu et al., 2021b). For example, Kann et al. (2020) investigate the POS-tagging performance when no additional resources, like manually created dictionaries, are available, and they find that performance drops substantially. Our work focuses on the validity of the common approach to simulate a low-resource scenario by randomly downsampling from a higher resource dataset (e.g., Gu et al., 2018; Chronopoulou et al., 2020; Dehouck and Gómez-Rodríguez, 2020; Kumar et al., 2021; Park et al., 2021; Zhang et al., 2021a). We investigate the biases that occur in the dataset statistics of the downsample, and how training on such downsampled datasets affects model performance.

5.2.3 *Different Learning Strategies*

Different learning strategies have been proposed to optimally make use of available data. Curriculum learning (CL) (Bengio et al., 2009) is motivated by the idea that humans learn best when following certain curricula. For example, one effective curriculum is to learn new things in increasing order of difficulty. CL aims at finding similar curricula for artificial model training, by finding meaningful orders in which to present data to a model, such that the model learns more effectively. Some studies report improved results when using CL (Xu et al., 2020a; Chang et al., 2021; Zhang et al., 2021b), whereas for other studies CL does not seem to help yet (e.g., Liu et al., 2019; Rao Vijjini et al., 2021).

Active learning (AL) (Cohn et al., 1996) is a related learning strategy, in which a model actively selects the data that it can most effectively be trained on at different points during training, for example based on its uncertainty for certain data points. As such, AL has often been used as an effective way to decide which data points to label in an unlabeled dataset (e.g., Reichart et al., 2008; Xu et al., 2018; Ein-Dor et al., 2020; Chaudhary et al., 2021).

5.3 EMPIRICAL INVESTIGATION

We empirically investigate downsampling from a high- to a low-resource scenario on two well-known NLP tasks: POS-tagging and machine translation. Both tasks are also popular low-resource tasks (e.g., Hedderich et al., 2021; Haddow et al., 2022) for which downsampling strategies have been used (e.g., Irvine and Callison-Burch, 2014; Ding et al., 2020; Kann et al., 2020; Araabi and Monz, 2020), making them suitable for our investigation. Moreover, POS-tagging is especially suitable, as the task is relatively quick and straightforward, giving us a good starting point. We found downsampling approaches to be especially prominent in the MT literature (e.g., Irvine and Callison-Burch, 2014; Fadaee et al., 2017; Ma et al., 2019; Araabi and Monz, 2020; Kumar et al., 2021; Xu et al., 2021), making it a natural task for our investigation. Our work serves as a good start to investigate other tasks in the future. For each task we investigate the effect of downsampling on the dataset statistics, and on the modeling performance.

We emphasize that our goal is to get a general understanding of the effect of simulating a low-resource scenario by randomly downsampling from a high-resource scenario. Therefore, we also keep our investigation general. That is, we use default versions of state-of-the-art models for both tasks, instead of versions that are fully optimized to get the highest possible scores. We also explicitly do not dissect individual papers in which downsampling is used. This is not the goal of this work, and we believe that there can be good reasons to use downsampling, as discussed in Section 5.1. Instead, we aim to provide useful insights that can be taken into consideration in future work.

5.3.1 POS-tagging

Briefly, POS-tagging is the task of assigning grammatical parts of speech, such as nouns, verbs, etc., to tokens in the input text. We use the Universal Dependencies (UD) dataset (see Marneffe et al. (2021)) for our experiments.

Data Description

The Universal Dependencies project[3] consists of treebanks for over a hundred languages (Marneffe et al., 2021), with varying amounts of resources. Languages are labeled with morphosyntactic labels, such as dependency tags and POS-tags. We only make use of the POS-tags.

Effect of Downsampling on Dataset Statistics

First, we downsample datasets from several high-resource languages, until they have the same size as the lower resource language datasets in the UD. We determine size based on the number of tokens or sentences. To investigate whether tokens in different languages can be equally compared from a typo-

[3] Website: `https://universaldependencies.org/`, Github: `https://github.com/UniversalDependencies`.

logical point of view, we start with a typological inspection of the languages in the UD collection.

Typological considerations. Languages differ from each other in their morphological complexity, for example in their morpheme per word ratios (Baker et al., 2012). Although subject to some debate, this can be described as the difference between analytic and synthetic languages.[4] Analytic languages have a low morpheme per word ratio, as opposed to synthetic languages. Within the synthetic category, one can differentiate between agglutinative and fusional languages, depending on how well single morphemes can be distinguished.

To the best of our knowledge, there is no easily accessible, exhaustive list that categorizes the languages in the UD as either analytic or synthetic, and thus we use two proxies. First, we use the *inflectional synthesis of the verb* as reported by the WALS (Bickel and Nichols, 2013),[5] which measures the number of inflectional categories per verb in different languages. To do so, it uses the 'most synthetic' form of the verb. WALS defines 7 categories, ranging from 0-1 till 12-13 categories per word. We label all UD languages included in the WALS. Second, if a Wikipedia page with information about the language type exists, we use this as a proxy to label the corresponding UD language.

Motivated by the idea that the language type might affect the tokenization quality, we compute the average ratio between the unique number of tokens and the total number of tokens for the labeled languages (Table 5.1). We only find a significant difference between the agglutinative and analytic languages ($t = -2.20, p = 0.04$). Agglutinative languages have more unique tokens per total of tokens, so they could be harder to tokenize. However, as we will see next, even if we downsample from an analytic language like English, we end up with a larger vocabulary size in the majority of samples.

Investigation of data statistics. With these typological considerations in mind, we now proceed to investigate the effect of downsampling on the dataset statistics. The UD provides an excellent testbed for our inspection, as the datasets of the included languages are of different sizes. First, we filter them on a number of criteria:

1. We only include non-extinct languages;

[4] There are also still other categories, like isolating languages. As we simply base ourselves on the morpheme per word ratios for our analysis, we leave these out for simplicity.

[5] `https://wals.info/chapter/22`

Table 5.1: Average ratio of vocabulary size per total number of tokens for different language types.

	Category	Count	Avg ratio
WALS	$0-1$	1	$0.12_{\pm 0.00}$
	$2-3$	6	$0.12_{\pm 0.07}$
	$4-5$	10	$0.09_{\pm 0.05}$
	$6-7$	5	$0.18_{\pm 0.10}$
Wiki	Analytic	9	$0.11_{\pm 0.03}$
	Agglutinative	22	$0.22_{\pm 0.15}$
	Fusional	4	$0.10_{\pm 0.05}$

2. We only include languages that have a POS-tagged dataset available on the UD Github page;

3. For some corpora, tokens are not released but marked by an underscore. We filter these out;

4. Some languages have multiple corpora that are very similar, but somewhat differently tagged. We filter these corpora to avoid duplication.

Based on these selection criteria, we arrive at a total of 100 languages.[6] We select the five highest resource languages in the UD: Czech, French, German, Icelandic, and Russian. We also include English, as it is often used to downsample from and still one of the higher resourced languages in the UD.

Next, we randomly downsample each of these high-resource languages to the size of the remaining lower resource languages. We compute size based on the number of tokens and number of sentences. We report the results based on the number of tokens in the main body of this chapter.

We investigate how downsampling affects the vocabulary size by computing the difference in vocabulary size between the downsampled dataset and its respective low-resource dataset. We normalize by the number of tokens in the low-resource dataset, to make a fair comparison. We plot the results of this analysis in Figure 5.1. In this plot, a positive number indicates that the vocabulary size of the downsample is larger than the original low-resource dataset, whereas a negative number indicates the opposite. In line with our intuition

[6] We refer to the appendix in (ter Hoeve et al., 2022a) for an exhaustive overview.

from Section 5.1, we find that downsampling indeed results in a larger vocabulary in the vast majority of cases. We find the same effect for downsampling based on number of sentences (Appendix 5.A, Figure 5.A.1). For this setting we also find that the downsampled corpora mostly contain more tokens than their originals (Appendix 5.A, Figure 5.A.2).

Effect of Downsampling on Model Training

Having shown that downsampling from a higher resource dataset often results in a larger vocabulary than that of the original lower resource language, we now investigate the effect of vocabulary size on the modeling performance for POS-tagging. In line with most related work, we fully focus on English as our high-resource language. We sample a number of smaller datasets from the English UD. Each of these samples has the same number of sentences, but they differ in vocabulary size. To achieve this, we use a greedy approach for the downsampling: we shuffle all sentences and greedily add sentences until we have the desired vocabulary size and the desired number of sentences.[7] We construct training datasets of 1,000 sentences each, for three vocabulary sizes: 1,000, 2,000 and 3,000 tokens. We limit the validation sets to the same vocabulary as the training set, and use the original test set in order to be able to compare different settings equally. We sample each of these settings five times, for five different random seeds.

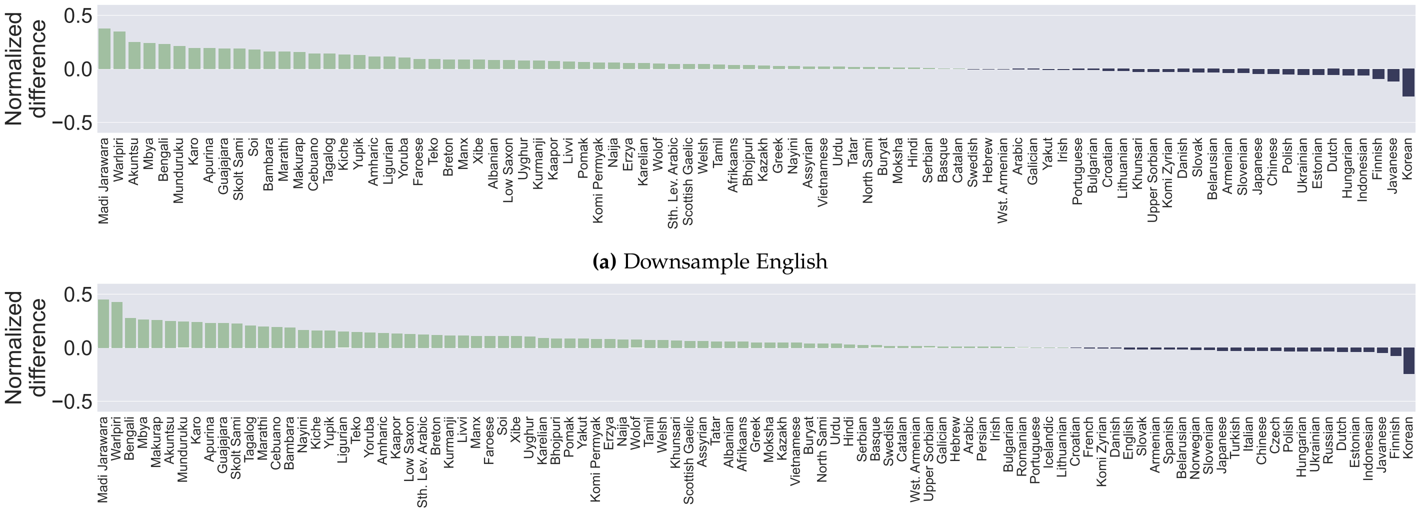
Next, we use these sampled datasets to model the POS-tagging task, for which we use the standard POS-tagging setup from the FlairNLP library.[8] We use FlairNLP's implementation of a sequence-to-sequence tagger, which defaults to a bidirectional RNN-CRF.[9] We compare three word embedding types: (i) *word2vec* embeddings (Mikolov et al., 2013) that we train from scratch on our training sets, (ii) pre-trained Glove embeddings, and (iii) pre-trained BERT embeddings. For the latter two we use the implementation from FlairNLP, for the *word2vec* embeddings we use Gensim.[10] This setting is most realistic, as it

[7] We also experimented with token-based downsampling, but did not find a good trade-off where the vocabulary size increased, whereas the number of tokens stayed the same. We also experimented with different sampling strategies, which did not change our findings.

[8] `https://github.com/flairNLP/flair/blob/master/resources/docs/TUTORIAL_7_TRAINING_A_MODEL.md`

[9] `https://github.com/flairNLP/flair/blob/master/flair/models/sequence_tagger_model.py`, 25M-125M parameters, depending on the embedding type.

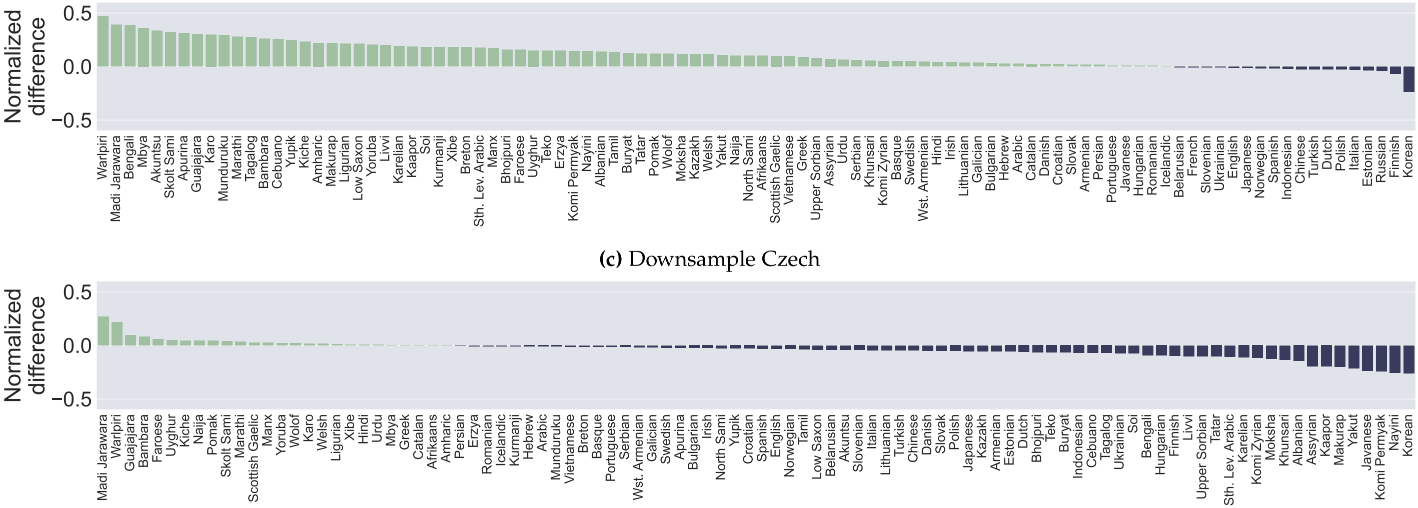
[10] `https://github.com/RaRe-Technologies/gensim`



(a) Downsample English

(b) Downsample German

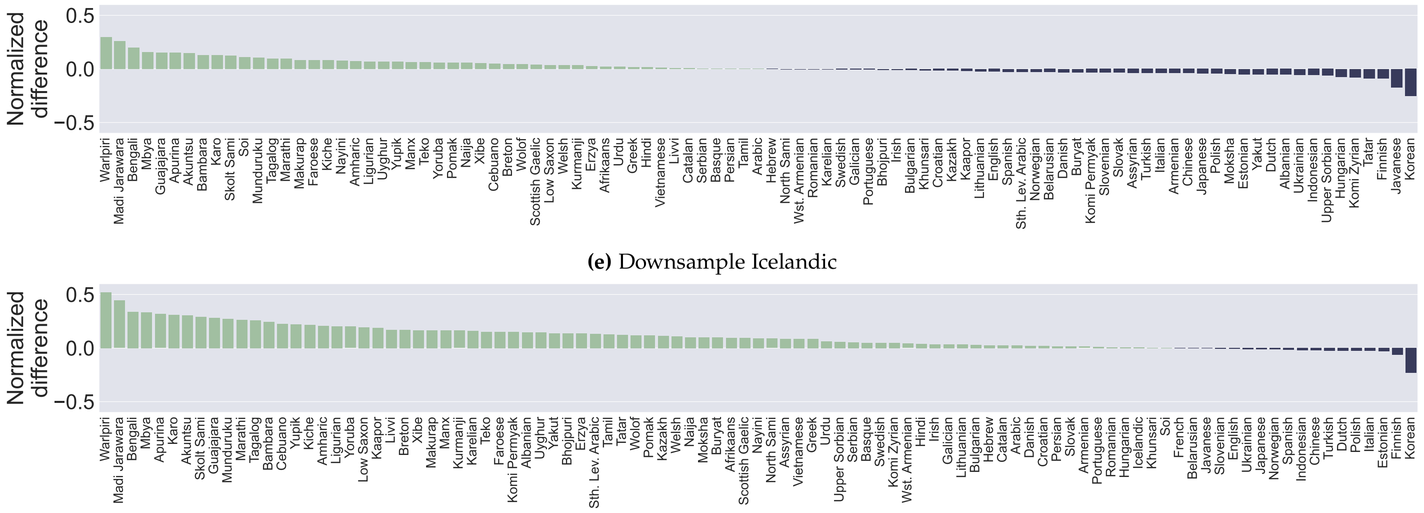
Figure 5.1: *Figure continues on next page.* Effect of downsampling on vocabulary size. Downsampling based on number of tokens. Each plot describes a different language that we are downsampling. The x-axis shows the language that we use as reference. The normalized difference in vocabulary size between the downsample and the original low-resource reference language is shown on the y-axis. A positive number indicates that the vocabulary size of the downsample is larger than the original low-resource dataset, whereas a negative number indicates the opposite. We find that downsampling indeed results in a larger vocabulary in the vast majority of cases.



(c) Downsample Czech

(d) Downsample French

Figure 5.1: *Figure continued from previous page and continues on next page.* Effect of downsampling on vocabulary size. Downsampling based on number of tokens. Each plot describes a different language that we are downsampling. The x-axis shows the language that we use as reference. The normalized difference in vocabulary size between the downsample and the original low-resource reference language is shown on the y-axis. A positive number indicates that the vocabulary size of the downsample is larger than the original low-resource dataset, whereas a negative number indicates the opposite. We find that downsampling indeed results in a larger vocabulary in the vast majority of cases.



(e) Downsample Icelandic

(f) Downsample Russian

Figure 5.1: *Figure continued from previous page.* Effect of downsampling on vocabulary size. Downsampling based on number of tokens. Each plot describes a different language that we are downsampling. The x-axis shows the language that we use as reference. The normalized difference in vocabulary size between the downsample and the original low-resource reference language is shown on the y-axis. A positive number indicates that the vocabulary size of the downsample is larger than the original low-resource dataset, whereas a negative number indicates the opposite. We find that downsampling indeed results in a larger vocabulary in the vast majority of cases.

Table 5.2: POS-tagging scores for different vocabulary sizes, while keeping the number of sentences equal. We report macro F1-scores for different word embeddings. Sents = Sentences. Toks = Tokens.

			Macro F1		
Vocab size	**Nr Sents**	**Nr Toks**	**Word2Vec**	**Glove**	**BERT**
1,000	1,000	7,235.75±174.485	0.328±0.021	0.743±0.005	0.921±0.003
2,000	1,000	11,252.0±227.885	0.350±0.024	0.773±0.005	0.937±0.003
3,000	1,000	14,867.2±292.534	**0.360±0.006**	**0.778±0.010**	**0.940±0.005**

is the only embedding type that is trained without access to another dataset or model. As low-resource work sometimes still makes use of these large pre-trained models, we include them for completeness. Moreover, a model like English BERT has been shown to be relatively multilingual (Pires et al., 2019). Table 5.2 gives the results.[11] We give additional micro F1-scores in Appendix 5.A.1, Table 5.A.1. We find that model scores increase when the vocabulary size increases.[12] In line with our downsampling analysis in the previous section, we find that the total number of tokens also increases. Unsurprisingly, we find that pre-trained word embeddings substantially outperform our own *word2vec* model.

Summarizing, in our POS-tagging investigation we find that downsampling from high-resource languages often results in a larger vocabulary size, and that a larger vocabulary size positively affects the scores on the POS-tagging task, in our settings for English. This is in line with the first issue that we raised in Section 5.1. We take our results on the POS-tagging experiments as a first strong indication that one needs to be careful with naive downsampling, as we already find differences in the current, still limited, scenario. Naturally, our findings raise many follow-up questions regarding the effects for different settings, such as for different domains, languages, or tasks. Therefore, we shift our focus to another task that is often the focus of low-resource investigations: machine translation.

[11] For the setting with a vocabulary size of 1,000 we had to remove the results of one of the seeds, as it did not find enough sentences.

[12] We also find that the scores for a vocabulary size of 2,000 and 3,000 tokens are similar, although the average for 3,000 is higher.

5.3.2 Machine Translation

Machine translation aims at translating text from a source to a target language. Machine learning systems address this task primarily by learning from bilingual documents with corresponding human translations (Koehn, 2020). These systems have shown substantial progress in recent years (e.g., Barrault et al., 2019; Barrault et al., 2020a; Akhbardeh et al., 2021) and have been applied to a growing number of language pairs (e.g., Platanios et al., 2018; Costa-jussà et al., 2022). We use the WMT datasets (see Akhbardeh et al. (2021)) for our experiments.

Data Description

The WMT is a collection of datasets for research on machine translation belonging to the WMT shared tasks, which were first organized in 2006 (Koehn and Monz, 2006). The first WMT collection consisted of three European language pairs: English-German, English-French and English-Spanish. The WMT shared tasks have been expanded each year, with additional translation pairs for the original language pairs, and with additional data for new language pairs and tasks (Callison-Burch et al., 2007; Callison-Burch et al., 2008; Callison-Burch et al., 2009; Callison-Burch et al., 2010; Callison-Burch et al., 2011; Callison-Burch et al., 2012; Bojar et al., 2013; Bojar et al., 2014; Bojar et al., 2015; Bojar et al., 2016; Bojar et al., 2017; Bojar et al., 2018; Barrault et al., 2019; Barrault et al., 2020a; Akhbardeh et al., 2021). An especially large jump in resources was made in 2017. This gives us a unique opportunity to test the effect of downsampling. In our investigation we treat the early versions of the WMT as low-resource setting, and later versions of the WMT as high-resource setting. We focus on the English-German translation pairs.

Effect of Downsampling on Dataset Statistics

To explore the effect of downsampling on the dataset statistics, we use the WMT 2014 German-English dataset (WMT14) as our low-resource dataset, and the 2018 version (WMT18) as our high-resource dataset. We focus on the English-German translation task. We apply two types of downsampling: sentence- and token-based. For the sentence-based setting we shuffle the WMT18, and sample the same number of sentences as in the WMT14. For the token-based setting, we shuffle the WMT18, and greedily add sentences until we reach the same number of tokens as in the WMT14.

We plot the downsampling effect in Figure 5.2. These plots reflect the WMT train sets over different years. We focus our investigations on WMT14 and WMT18, but plot all years from 2013 till 2019 for reference. The last two light green bars show the two downsampled datasets. If we downsample based on sentences (first light green bar right to the dotted line), we find that the number of tokens *decreases*, whereas the vocabulary size *increases*. If we downsample based on tokens, both the number of sentences and the vocabulary size increase.

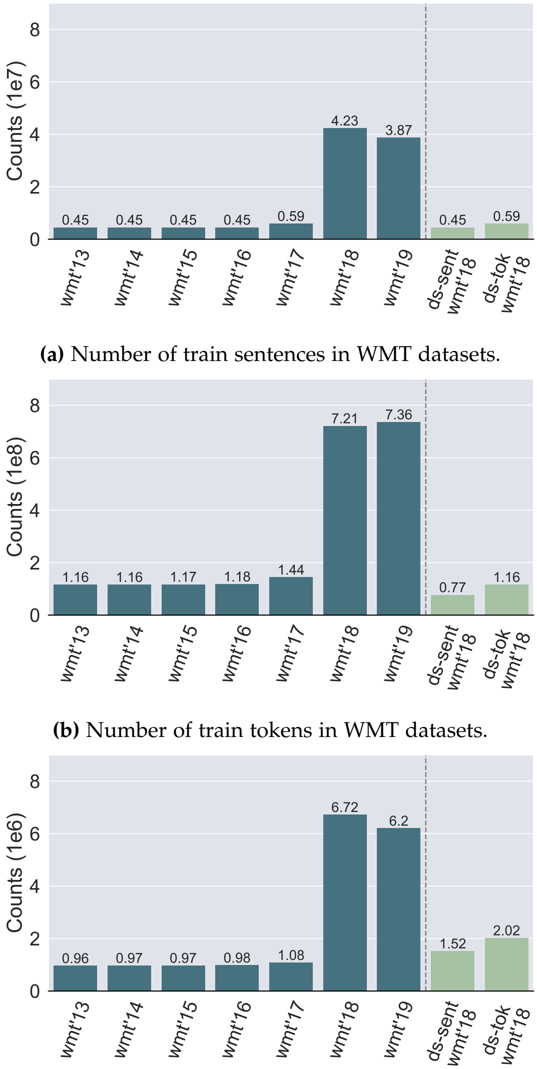
We also qualitatively inspect the vocabulary distributions. In Appendix 5.A, Figure 5.B.1 we plot the 100 most frequent words in each dataset that we compare. We find that there are quite a few differences, especially in the second half of the plot.

Effect of Downsampling on Model Training

Next, we investigate the effect of downsampling on model training. To this end, we train and evaluate transformer sequence-to-sequence models (Vaswani et al., 2017) in different data settings. We use the Flax transformer code,[13] and only adapt the data pipeline to be able to work with our downsampled datasets. We train these models on 8 A100 GPUs on the standard WMT14 and WMT18 train sets, and on our two downsampled datasets (token- and sentence-based). We test on the WMT14 and WMT18 test sets (i.e., newstest data (Barrault et al., 2020b)). We report the scores in Table 5.3.

A few observations stand out. First, the models trained on downsampled versions of the WMT18 score lower on the WMT18 test set than the model trained on the original WMT18 dataset. This is as expected, if we assume that the additional WMT18 data would lead to better results. We also find that training on WMT14 and testing on WMT18 leads to higher scores than testing on the WMT14 test set. This is remarkable, but in line with earlier findings (Edunov et al., 2018). Finally, we observe that the models trained on the downsampled WMT18 datasets perform worse on the WMT14 test set than the models trained on the WMT14 dataset itself, in contrast to our findings for the POS-tagging experiments. For the MT experiments, having a richer vocabulary does not seem to help performance. We hypothesize that this can be explained

[13] `https://github.com/google/flax/tree/main/examples/wmt`, 213M parameters



(a) Number of train sentences in WMT datasets.

(b) Number of train tokens in WMT datasets.

(c) Vocabulary size of WMT datasets.

Figure 5.2: Statistics of WMT datasets (ds = downsampled, sent = sentence-based, tok = token-based). The dotted line separates the original WMT datasets from the downsampled datasets.

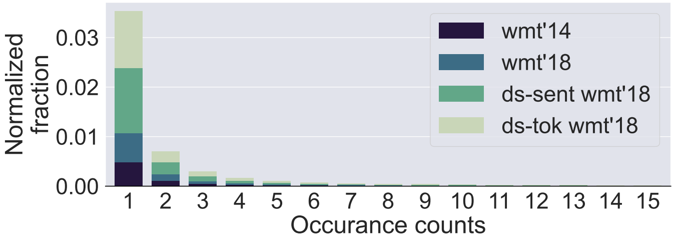


Figure 5.3: How many words occur N times in the different WMT datasets, normalized.

by the quality of the WMT18 datasets, i.e., the second issue that we raised in Section 5.1. As shown in Figure 5.2, the amount of data increased heavily in 2017, mostly driven by the inclusion of the Paracrawl data source (Bañón et al., 2020). This data source is known to be noisy, and people have worked on filtering it (e.g., Junczys-Dowmunt, 2018; Aulamo et al., 2020; Zhang et al., 2020a).

To add to the data quality investigation, we count how many words occur N times in each dataset, normalized by the total number of words in the dataset. If a dataset contains many words that only occur once, this indicates that it contains more noise (such as links) than datasets with fewer single occurring words. We plot the results in Figure 5.3. WMT18 contains more single occurring words than WMT14, an indicator that the average quality of the WMT18 dataset is indeed lower, negatively impacting our downsampled experiments.

Summarizing, for our MT experiments we find that downsampling also increases the vocabulary size, in line with our hypothesis and with our findings for the POS-tagging experiments. We also find that the downsampled datasets did not increase the translation performance, which can be explained by the lower quality of the high resource data.

5.4 DISCUSSION

We found evidence for both issues that we raised regarding simulating a low-resource scenario by randomly downsampling from a high-resource language. A downsampled dataset is likely a poor proxy for a low-resource scenario. On

Table 5.3: BLEU scores for MT models trained and tested on different (downsampled) train and test sets.

	WMT	Train 14	18	ds-sent-18	ds-tok-18
Test	**14**	32.62	32.12	29.37	30.23
	18	41.49	39.70	37.66	38.20

the one hand, such a downsampled dataset can cover a wider vocabulary range than a low-resource dataset, resulting in higher scores. On the other hand, the high-resource dataset might be less carefully constructed than is needed for a low-resource scenario, causing noise in the downsample, eventually leading to lower scores. In this section we reflect on these findings. We hope that our work serves as additional evidence for the proxy fallacy. Being aware of this fallacy puts individual researchers and the field as a whole in a better position.

The best strategy for low-resource investigations is to use truly low-resource data, whenever possible. There are many examples that do this, or in which downsampled high-resource data is only used for additional experiments (e.g., Kann et al., 2020; Kumar et al., 2021; Adelani et al., 2021). For situations in which using truly low-resource data is not an option, for example because the required data is simply not available, we first want to echo Hedderich et al. (2020). They show that large improvements can be obtained by only labeling a few data points. We believe that we can use recommendations from active learning and curriculum learning to choose which data points are best to label, and hope to experiment with this question in future work. If labeling additional data points is not an option, and one is truly bound to simulating a low-resource scenario by downsampling from a high-resource dataset, one needs to be aware of the biases that we found in this work. A downsampled dataset is likely not a good reflection of the low-resource setting, which can result in scores that are either too high (because of the richness of the data) or rather too low (because the high-resource data may be of insufficient quality).

5.5 LIMITATIONS

Throughout this work we flagged some of the limitations of our approach. In this section we summarize these in more detail, to help future investigations.

The datasets. In this work we concentrated on corpora from two data sources: the UD and the WMT. Although this is a good start and these datasets are a good fit for our investigation, we hope that future work investigates different corpora, to get an even better understanding of the effect of uniform downsampling.

The tasks. The same holds for the types of tasks that we chose. Although we believe POS-tagging and MT to be a good start, future work should investigate different tasks to be able to form a more general understanding.

5.6 ETHICAL STATEMENT

In this work we developed an understanding of the effect of simulating a low-resource language by downsampling uniformly from a high-resource language. By pointing out biases that may occur, we hope to have raised awareness for this issue, making follow-up work on low-resource languages more inclusive. However, there are around 7,000 languages world-wide, of which we have only been able to cover a few.

5.7 CONCLUSION

In this chapter we answered the fourth research question of this thesis, as we investigated the validity of simulating a low-resource scenario by downsampling from a high-resource dataset. We argued that this process might be a poor proxy for a truly low-resource setting, for two reasons: (i) a high--resource dataset might be much richer in content than a low-resource dataset, and (ii) the high-resource dataset might be of lower quality than a low-resource dataset that was carefully crafted. We empirically studied this on two well-known NLP tasks: POS-tagging and machine translation. Our investigation showed that uniform downsampling is indeed a poor proxy in these two scenarios, and we found evidence for both hypothesized reasons. As such, our work serves as a warning for work in low-resource domains. Our work also serves as a starting point to formalize best practices to grow datasets, and to more reliable simulations of low- to high-resource settings. This is also related to questions regarding external and ecological validity (Andrade, 2018). In future work, we plan to expand our analysis to more tasks and more languages.

In the next chapter we continue in a new direction, as we leave the user-centered focus somewhat behind us. Instead, we take a more linguistically inspired approach to language modeling.

CHAPTER APPENDIX

5.A ADDITIONAL PLOTS POS-TAGGING EXPERIMENTS

In this section we give additional results for the POS-tagging experiments (Section 5.3.1). We give additional results for downsampling based on number of sentences in Figure 5.A.1 and Figure 5.A.2.

5.A.1 *Additional micro F1-Scores for POS-tagging Experiments*

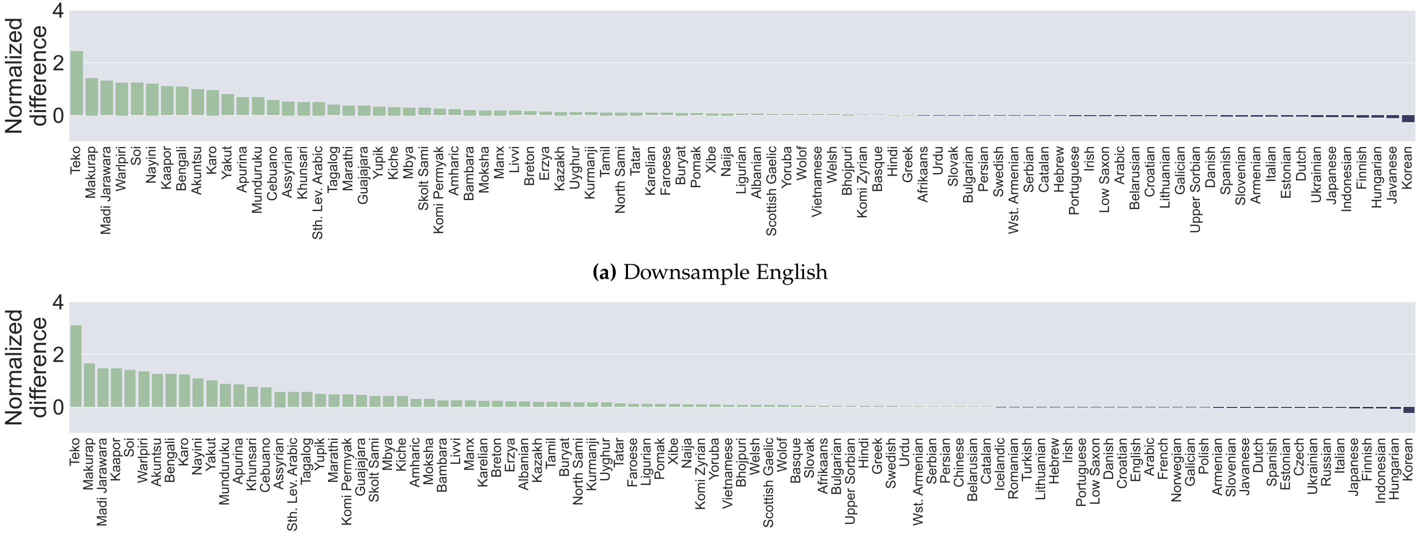
In Table 5.A.1 we give additional Micro F1 scores for the POS-tagging modeling performance with the downsampled datasets.

5.B ADDITIONAL PLOTS MT EXPERIMENTS

In this section we give additional results for the MT experiments (Section 5.3.2). In Figure 5.B.1 we show the plots of a qualitative inspection of the words in the different WMT datasets (downsampled and original).

Table 5.A.1: POS-tagging scores for different vocabulary sizes, and different word embeddings. Micro F1-scores. Sents = Sentences. Toks = Tokens.

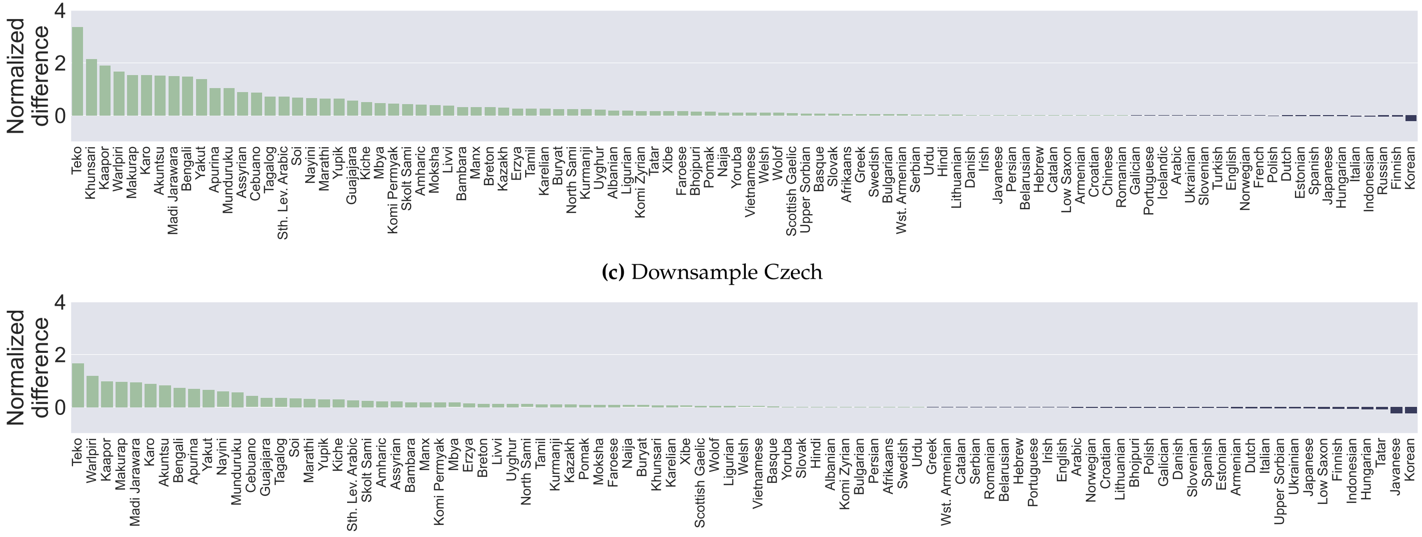
			Micro F1		
Vocab size	**Nr Sents**	**Nr Toks**	**Word2Vec**	**Glove**	**BERT**
1,000	1,000	7,235.75±174.485	0.189±0.013	0.581±0.021	0.801±0.007
2,000	1,000	11,252.0±227.885	0.208±0.015	**0.624±0.017**	0.853±0.008
3,000	1,000	14,867.2±292.534	**0.215±0.005**	0.622±0.030	**0.880±0.015**



(a) Downsample English

(b) Downsample German

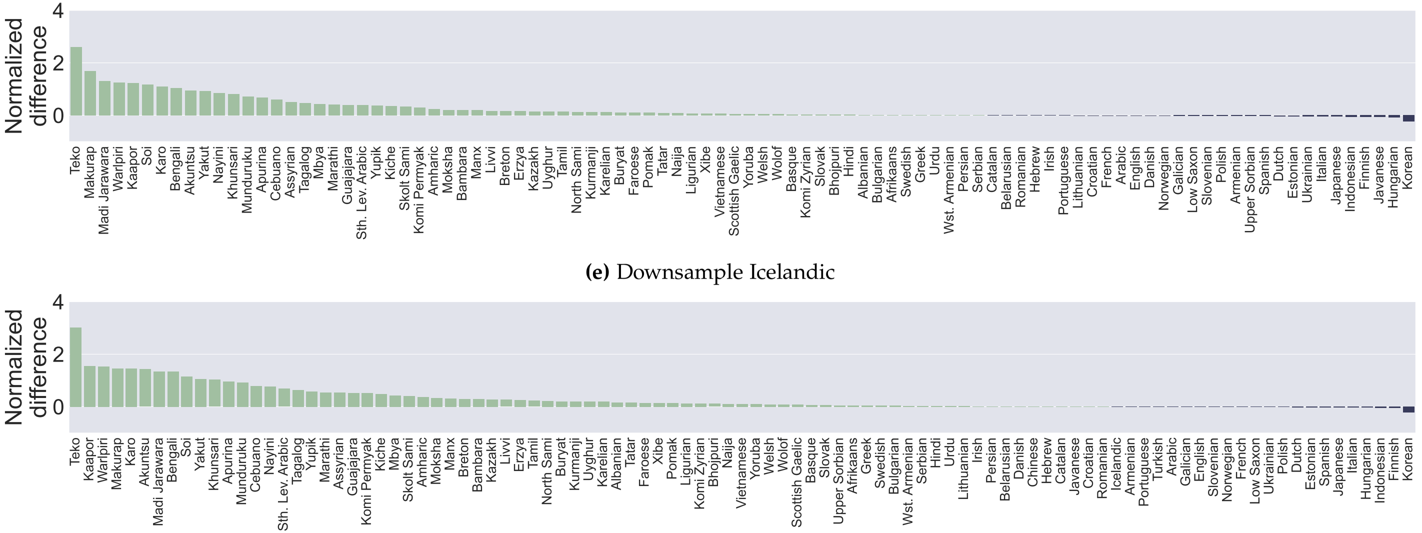
Figure 5.A.1: *Figure continues on next page.* Effect of downsampling on vocabulary size. Downsampling based on number of sentences. Each plot describes a different language that we are downsampling. The x-axis describes the language that we use as reference.



(c) Downsample Czech

(d) Downsample French

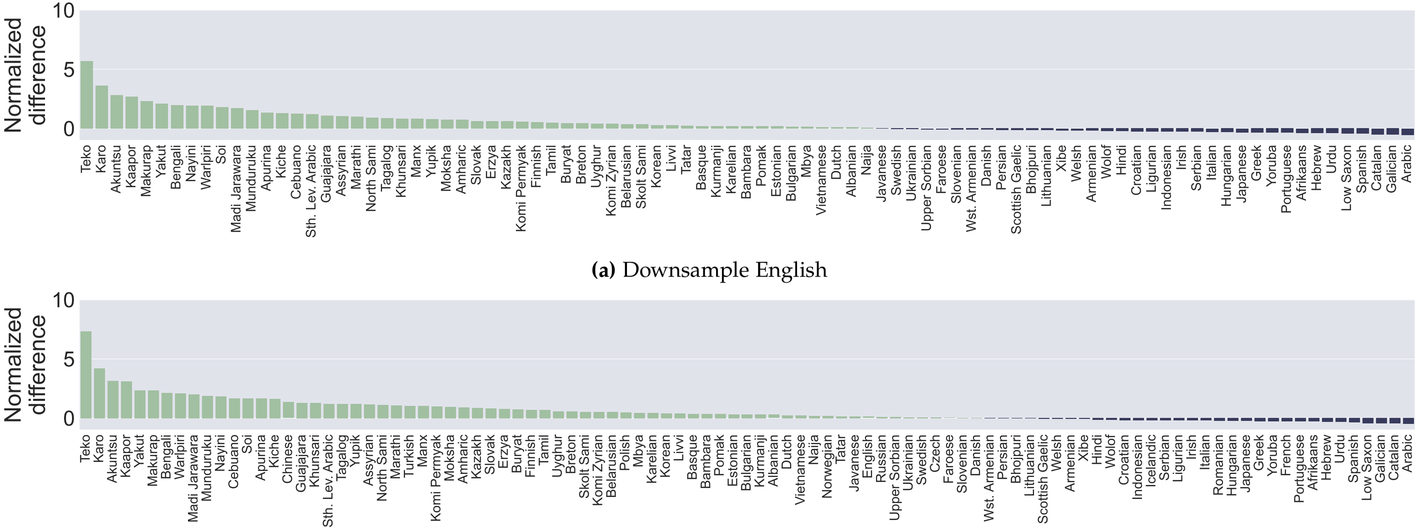
Figure 5.A.1: *Figure continued from previous page and continues on next page.* Effect of downsampling on vocabulary size. Downsampling based on number of sentences. Each plot describes a different language that we are downsampling. The x-axis describes the language that we use as reference.



(e) Downsample Icelandic

(f) Downsample Russian

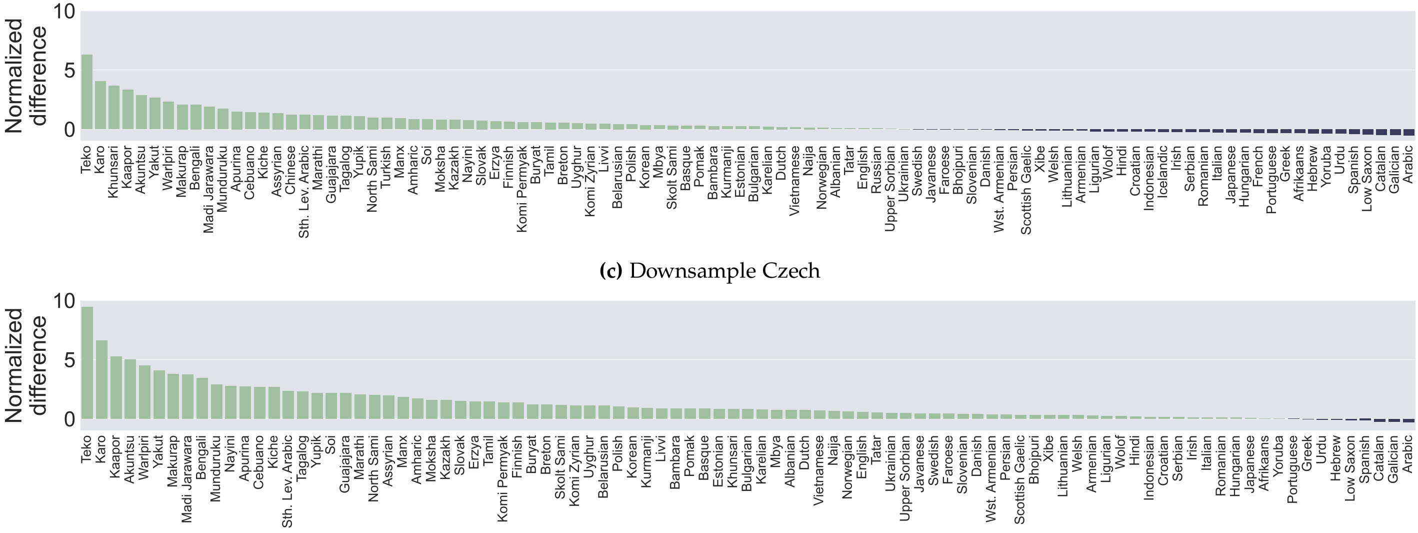
Figure 5.A.1: *Figure continued from previous page.* Effect of downsampling on vocabulary size. Downsampling based on number of sentences. Each plot describes a different language that we are downsampling. The x-axis describes the language that we use as reference.



(a) Downsample English

(b) Downsample German

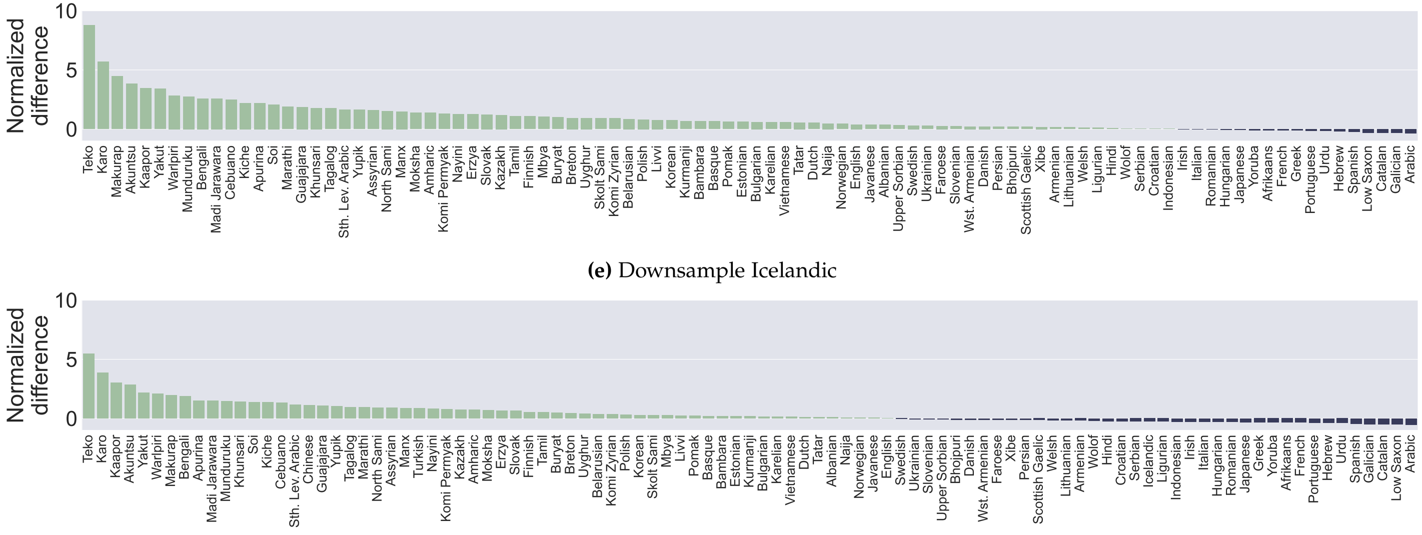
Figure 5.A.2: *Figure continues on next page.* Effect of downsampling on number of tokens. Downsampling based on number of sentences. Each plot describes a different language that we are downsampling. The x-axis describes the language that we use as reference.



(c) Downsample Czech

(d) Downsample French

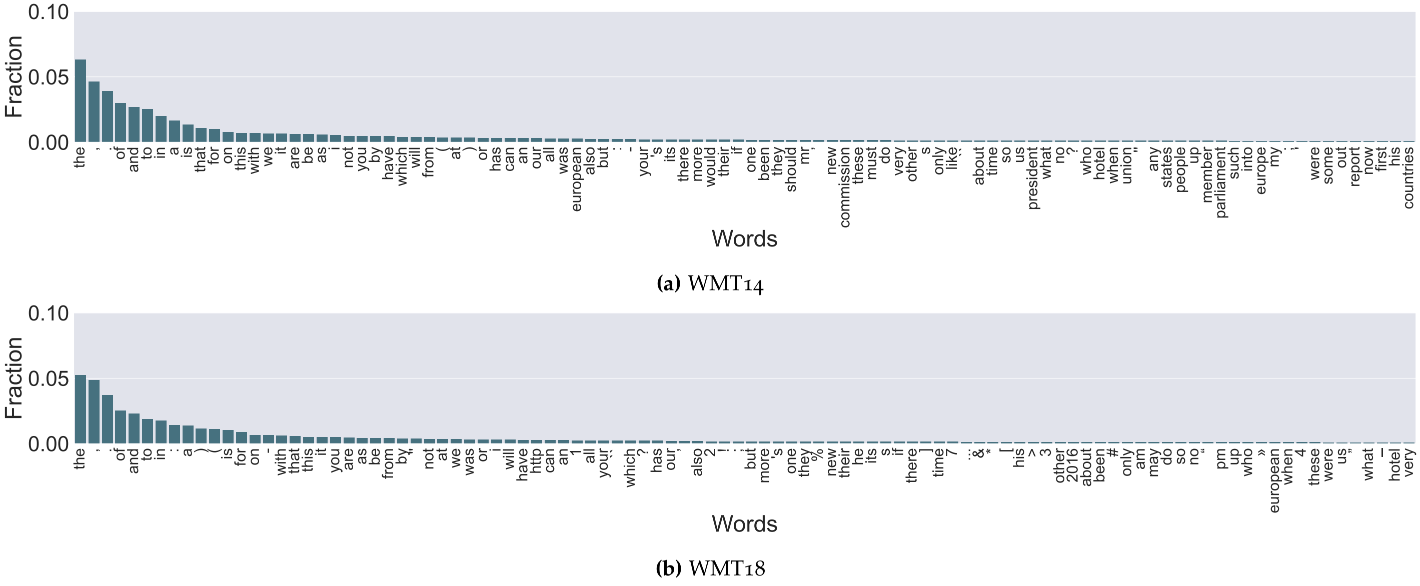
Figure 5.A.2: *Figure continued from previous page and continues on next page.* Effect of downsampling on number of tokens. Downsampling based on number of sentences. Each plot describes a different language that we are downsampling. The x-axis describes the language that we use as reference.



(e) Downsample Icelandic

(f) Downsample Russian

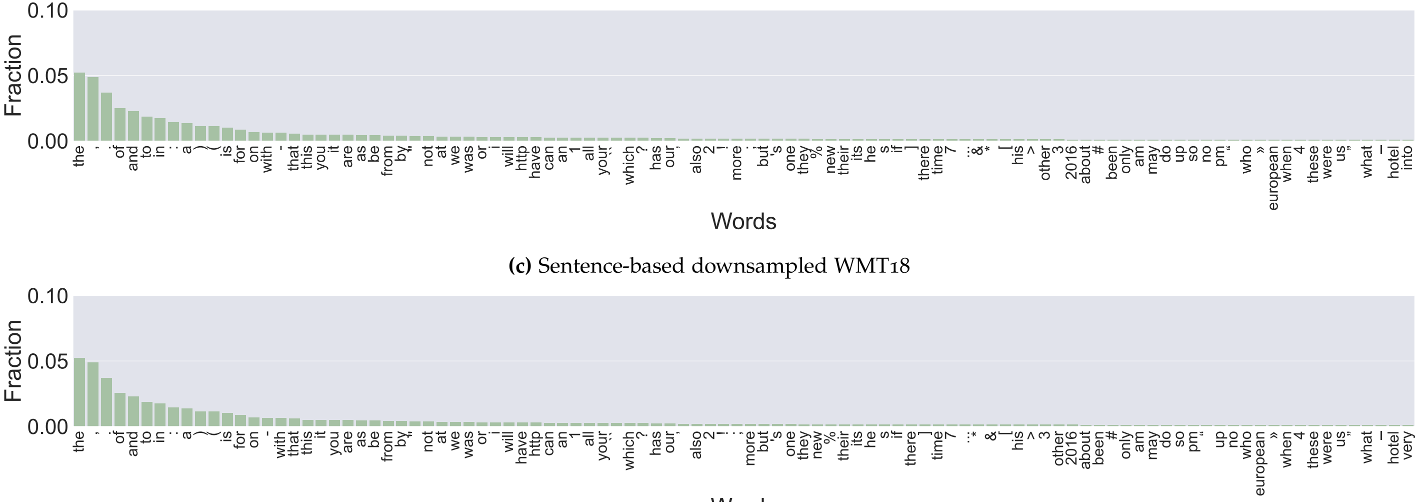
Figure 5.A.2: *Figure continued from previous page.* Effect of downsampling on number of tokens. Downsampling based on number of sentences. Each plot describes a different language that we are downsampling. The x-axis describes the language that we use as reference.



(a) WMT14

(b) WMT18

Figure 5.B.1: *Figure continues on next page.* Top 100 words in the train sets of WMT14, WMT18 and downsampled WMT18.



(c) Sentence-based downsampled WMT18

(d) Token-based downsampled WMT18

Figure 5.B.1: *Figure continued from previous page.* Top 100 words in the training sets of WMT14, WMT18 and downsampled WMT18.

6 TOWARDS INTERACTIVE LANGUAGE MODELING

An important motivation of this thesis is to design and develop NLP models in line with users' needs. We have also been partially inspired by ideas from cognitive science and linguistics, in particular in Chapter 4. In this chapter[1] we continue in this direction, while focussing less on the user aspects. Specifically, we take a look at language modeling. Although large language models perform extremely well, they are trained on very large amounts of data and their training regime appears unnatural from the perspective of human language acquisition. The latter is much more interactive in nature. Fascinated by this observation, we explore the role that interaction can play in *artificial* language modeling, as we answer the fifth research question of this thesis:

Research Question 5: *How can we make artificial language modeling more human-like by taking a more interactive approach?*

Throughout this chapter, we also refer to this interactive approach to language modeling as *interactive language modeling*. This chapter is exploratory in nature. We first define the objective of interactive language modeling more explicitly and propose a teacher-student framework for the purpose of interactive language modeling. Next, we present a road map in which we detail the steps that need to be taken towards a fully interactive approach to language modeling, for each of the components in this framework. We then lead by example

[1] This chapter is based on (ter Hoeve et al., 2021). A similar version was also presented as (ter Hoeve et al., 2022b).

and take the first steps on this road map, which show the initial feasibility of our proposal. Our aim is to start a larger discussion and research agenda on interactive language modeling with this work.

6.1 INTRODUCTION

Interaction between children and more advanced language interlocutors (such as caregivers) plays an important role in many theories and studies on human language acquisition (e.g., Bruner, 1985; Clark, 2018). For example, although culturally dependent (Shneidman and Goldin-Meadow, 2012) and with the precise effects still up for discussion (Cristia et al., 2019), caregivers can communicate with their children in child directed speech. In turn, children can for example experiment with the meaning of words, to elicit a response from their caregivers (Gillis and Schaerlaekens, 2000).

Despite the importance of interaction in human language acquisition, interaction plays little to no role in artificial language modeling. This is remarkable, as language modeling also has the objective to learn human language, albeit with artificial models. Instead, current state-of-the-art language models (LMs) take large amounts of text as input, and are tasked to predict the next or masked words (e.g., Devlin et al., 2019; Brown et al., 2020). The learning signal only comes from a cross-entropy loss that indicates whether a prediction was correct. Although this setup has shown to be effective, from the perspective of human language acquisition it appears unnatural — children clearly do not learn language this way. This motivates us to explore ways in which interaction can play a role in artificial language modeling.

Specifically, in this chapter we explore a teacher-student setup for interactive language modeling. Figure 6.1 depicts a high level overview. In this setup we distinguish four main parts: *the teacher*, whose role is inspired by the caregiver in the human language acquisition, *the student*, who resembles the child, *the interaction* between the teacher and the student, and *the environment* that they both share (such as the language that needs to be learned by the student). The student and the teacher can interact with each other, and with the environment. We motivate and detail our setup further in Section 6.3.

An interactive approach to language modeling is not only interesting from the perspective of human language acquisition. Explicitly allowing for inter-

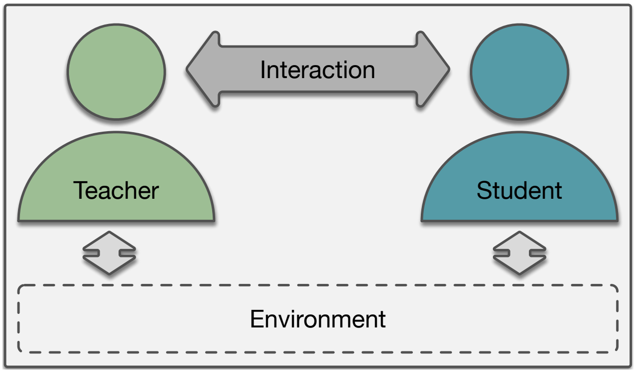


Figure 6.1: Teacher-Student setup for interactive language modeling.

action also has the potential to make language modeling more efficient and versatile. For example, a teacher can adapt its input to a student based on the specific feedback signals it receives from the student, and a teacher that is fluent in one domain can teach the specifics of that domain to a student trained on another domain, and vice versa. Moreover, an interactive approach to language modeling has the potential to impact downstream applications, for example in foreign language teaching apps where a student can be replaced by a human.

We structure the contributions in this chapter as follows:

- We define the objective of interactive language modeling;
- We present a road map that details the steps that need to be taken towards this objective;
- We take the first steps on this road map, which show the initial feasibility of our approach.

With these contributions we aim to start a larger research agenda on interactive language modeling.

6.2 RELATED WORK

In this section we describe a number of different learning strategies to train machine learning models that are particularly related to the current work.

6.2.1 Interactive Language Learning in NLP

Recently, a number of studies have focused on interactive language learning. Stein et al. (2021) learn logical semantic representations in an interactive way. Nikolaus and Fourtassi (2021) propose a proof of concept to model perception and production based learning of semantic knowledge acquisition in children. Kiseleva et al. (2022a) and Kiseleva et al. (2022b) take an interactive approach to language *understanding* in a recent NeurIPS challenge. To the best of our knowledge, none of these earlier works have focused specifically on language modeling.[2]

6.2.2 Curriculum Learning

Curriculum learning (CL) (Bengio et al., 2009) is an approach to learning in which data samples are presented in a meaningful order — typically in order of complexity — motivated by the idea that humans learn in a similar way. Bengio et al. show the effectiveness of CL on a number of tasks, among which a classical approach to language modeling. More recently, a number of studies have shown the effectiveness of CL for (fine-tuning) LMs (Xu et al., 2020a; Zhang et al., 2021b), although other studies have shown that not all intuitive curricula are also effective (Liu et al., 2019). Matiisen et al. (2020) propose a teacher-student framework for automatic CL for the addition of decimal numbers and navigation in Minecraft.

[2] Since the work in this chapter was published, InstructGPT (Ouyang et al., 2022) was released, a language model that is trained with human feedback to follow instructions. InstructGPT is a fine-tuned version of GPT-3 (Brown et al., 2020). Briefly, InstructGPT is trained with *reinformcement learning from human feedback* (Christiano et al., 2017), an iterative process in which human labelers rank outputs of a pre-trained language model. These scores are used to train a reward model. The original language model is then fine-tuned based on the scores of the reward model. Next, human labelers again rate the output of the now more fine-tuned language model, and the fine-tuning process starts again. At the time of writing this thesis, ChatGPT (`https://openai.com/blog/chatgpt/`) was released. ChatGPT is very similar to InstructGPT, with minor changes in the fine-tuning loop. InstructGPT and ChatGPT are different from the interactive approach to language modeling proposed in this chapter, as their training requires fine-tuning already pre-trained language models. However, InstructGPT and chatGPT show the potential of an interactive approach.

6.2.3 Active Learning

In active learning (AL) (Cohn et al., 1996) a learner (the model to be trained) actively selects which data it can most effectively be trained on. That is, where CL is often more associated with choosing a teaching strategy, AL is rather focused on the student side. AL is often used to efficiently label data in a low resource setting (e.g., Reichart et al., 2008; Ein-Dor et al., 2020).

6.2.4 Continual Learning

In continual learning, or life-long learning, the aim is to train a model in an online fashion, i.e., on a continuous stream of data, whilst avoiding *catastrophic forgetting* (McCloskey and Cohen, 1989; French, 1999). This makes models versatile to an ever-changing world. Some recent work has focused on types of continual learning for large LMs (e.g., Lazaridou et al., 2021; Jin et al., 2022). We envision interactive language modeling to play an important role in life-long learning in the future.

6.3 A ROAD MAP TOWARDS INTERACTIVE LANGUAGE MODELING

In this section we present a road map towards interactive language modeling. Before we can do this, we first need to define our objective for an interactive modeling framework in more detail:

Our objective is to build an automated teacher-student loop for language modeling that attains good performance in the student for a fixed (low) number of bits transmitted in the interactions.

We propose a teacher-student loop as this format closely resembles caregiver-child interactions. In Section 6.1 and Figure 6.1 we already introduced a high level overview of this setup and its four main components: (1) *the teacher*, (2) *the student*, (3) *the interaction* and (4) *the environment*. Generally, in this setup teachers transmit language data to their students, according to a certain budget ("a (low) fixed number of bits"). Having this budget forces the teacher to actively choose a learning strategy, as just sending all data that is available to the teacher would not be allowed. Students have the objective to learn the language and they send a signal back that informs their teacher of their performance, e.g., a score on an exam. This interaction takes place in an environment, e.g., a common language.

In Table 6.1 we present the road map that we envision towards interactive language modeling. This road map works as follows. For each of the four aforementioned components we detail steps that need to be taken. We also add a fifth component: the evaluation of the setup. Each component has different aspects (bold-faced in Table 6.1). For example, for the *teacher* we can focus on how it can access the data that it can transmit to the student, which we call "ways of speaking" in Table 6.1. Another aspect of the teacher side focuses on what we call the "degree of awareness", which entails different ways in which the teacher can remember different aspects of the teaching loop. In a similar fashion we fill in the remaining components in the table. We focus on text as a single modality and acknowledge grounded interactive language modeling as an interesting future research direction.

On our road map there are multiple ways to reach the destination. For example, one can focus on taking a few steps for each of the components, or to take many steps for only one or a few of the components. Moreover, although mostly structured in increasing degrees of complexity, this does not always hold for all individual steps in the table. For example, zooming in on the "degrees of awareness" for the teacher again, one could imagine an example where a teacher does not have an explicit memory buffer of what it sent to the student before, but does have an explicit way of remembering what the student's fine-grained capabilities are, as well as the other way around.

In the remainder of this work we take the first steps on the road map. We focus on the teacher side, i.e., learning the correct didactic approach.

Table 6.1: *Table spans over multiple pages.* Road map to interactive language modeling. We detail each of the steps we need to take for each of the components in the interactive language modeling setup. The steps that we take in this chapter are indicated by *. Note that there is not a single correct road through this road map. One could decide to fully focus on one of the components, or to take (multiple) steps in multiple components.

Teacher	**Student**
Ways of speaking • Select data from bin;* • Generate data with own language model. **Degrees of awareness** • (No*) memory buffer of what has been sent to the student and being able to act on it (see *Interaction* cell); • (No*) explicit way of remembering what the student's fine-grained capabilities are and being able to act on it (see *Interaction* cell).	**Ways of speaking** • Generate language data in a standard LM fashion;* • Actively experiment with language generation to elicit direct feedback from the teacher (see also *Interaction* cell). **Degrees of using the teacher data** • Use all data received from the teacher; • Actively select data that is useful; • Actively know when to stop training (for example to avoid overfitting).

Interaction	**Environment**
Teacher side • Send all data at once;* • Send data in batches, based on student feedback (see below). Batches can be as small as single utterances, after which the student sends an utterance back, like in real human-to-human interaction (see below); • Send (mid-term) exams. **Student side** • Send a single average exam score back to the teacher;* • Send a fine-grained exam score back, e.g., – score per item on the exam set; – (average) scores of different components (tasks) of the exam(s) • Ask for feedback, for example by actively experimenting with language generation for the teacher to judge ('generate own exam').	**Language** • Artificial languages, in increasing level of difficulty in terms of complexity, e.g., – random language;* – different types of structures;* – different vocabulary sizes; • Subset of human language, e.g., in terms of – semantics (e.g., different domains) – syntax (e.g., different grammatical structures) – pragmatics • Unrestricted human language. **Task** • *Teacher:* Learn to select or generate the optimal data such that the student performs well on the exam set (see cell below);* • *Teacher:* Learn to adapt to different types of students, e.g., – architectural differences – different prior knowledge (be aware of catastrophic forgetting in neural networks) • *Student:* Learn to adapt to different types of teachers (didactic strategies).

Evaluation / Exam

Teacher

- Accuracy in selecting the optimal teaching protocol.*

Student (Exam / Feedback for teacher)

- General performance, measured in perplexity;*
- Performance on specific tasks, such as
  - Subset of the data known to the teacher (e.g., specific domain or (grammatical) structure)
  - BLIMP (Warstadt et al., 2020);
  - BIG-Bench (Srivastava et al., 2022) (`https://github.com/google/BIG-bench`).
- Scores either as an average* or more fine-grained (see *Interaction* cell).

6.4 TAKING THE FIRST STEPS ON THE ROAD MAP

Figure 6.2 shows how we adapt the general setup from Figure 6.1 to take the first steps on the road map. Here we describe each modification per component: *the teacher*, *the student*, *the interaction*, *the environment* and *the exam* that the student takes. The main focus of this work is on the teacher's side, and thus we keep the remaining components relatively simple.

6.4.1 *The Teacher*

The role of the teacher is to transmit language data that will optimally help the student to learn the language. Figure 6.2 shows that we train the teacher to do this in a number of time steps. At each of these steps a teacher samples data from a larger language data set according to a fixed budget. We discuss the specifics of the sampling function below. To reduce the variance in the teacher's learning process we repeat this process for multiple students, i.e., a teacher selects N "lessons" for N students. Due to the stochasticity of the

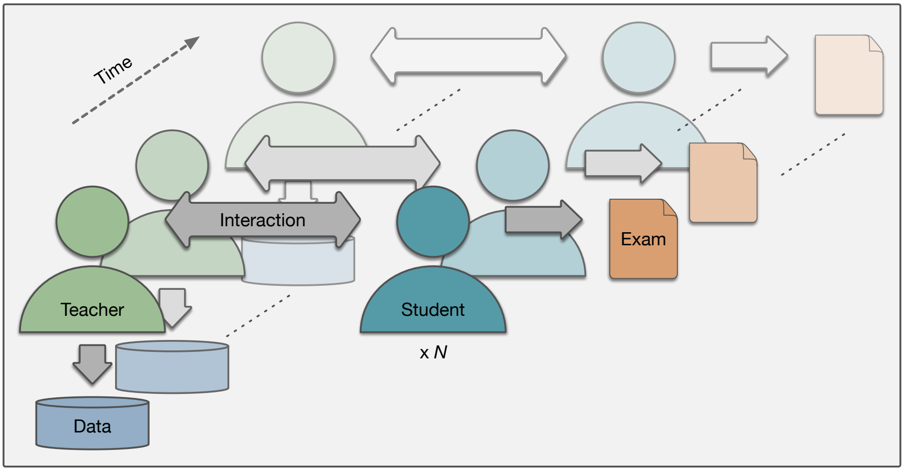


Figure 6.2: Teacher-student loop as used in this work.

sampling process, each student has the potential to be trained on a slightly different part of the data. Because we use a multiprocessing setup we can train multiple students on a single GPU. Hence, using multiple students does not drastically increase the computational cost.

Knowing the Language

The teacher is modeled as a native speaker of the language that it needs to teach. We represent the teacher's language understanding with a pre-trained causal transformer LM (Vaswani et al., 2017). We pre-train this model on a *different* subset of the data than the teacher can select from for the students, and thus we ensure that we measure whether a teacher can teach a language as a whole, and not only a particular subset that it was trained on itself.

Selecting the Data

We use REINFORCE (Williams, 1992) with entropy regularization (Mnih et al., 2016) to learn the teacher's didactic approach.[3] We want to optimize the teacher's policy such that it learns to select the optimal data to train the student on, given a predefined budget. The policy is a one-layer feed forward

[3] We also experimented with gradient-free optimization approaches such as the ones implemented in Nevergrad (Rapin and Teytaud, 2018), but found REINFORCE to be more flexible in learning the different tasks in our setup, and therefore it is a better fit for our needs.

neural network that outputs a score for each sentence, i.e., the teacher's policy network takes a sentence embedding as input, based on the pre-trained transformer LM that we use to represent the teacher's language understanding. An action is modeled as selecting k sentences from the larger data set, where k is a predefined teacher budget. We use the GumbelTopK trick (Vieira, 2014; Kool et al., 2019) to sample k sentences without replacement, based on the teacher policy's output scores. We compute the log probabilities (needed to compute the loss) for each sample by adding the log probabilities of each element in the sample. We refer to Appendix 6.A for more details.

6.4.2 The Student

As the teacher is the main focus of our work, we choose to keep the student side simple. We represent the student as a causal transformer LM that we train on the data that it receives from the teacher.

6.4.3 The Interaction

Following Table 6.1, the teacher sends all selected data to the student at once. The student uses this data to train its LM and takes an exam after a predefined number of updates. The average exam score is sent back to the teacher as feedback. We use the student's last model checkpoint to compute the scores (as opposed to the best checkpoint on a validation set), to ensure that the learning signal for the teacher is restricted to the student's performance on the exam set, i.e., we do not expect teachers to reverse the learning process of the students (just like caregivers cannot do this for their children).

6.4.4 The Environment

Following Table 6.1, we design a number of artificial languages to test our approach on (see Section 6.5 for details). Using artificial languages is a well-tested approach to study the behavior of neural networks (e.g., Batali, 1994; Wiles and Elman, 1995; Rodriguez et al., 1999; Gers and Schmidhuber, 2001; Rodriguez, 2001; Hupkes et al., 2018; Lake and Baroni, 2018; Saxton et al., 2019; Hupkes et al., 2020; Rodríguez Luna et al., 2020; Wal et al., 2020; Chaabouni et al., 2021; Dagan et al., 2021). Using artificial languages gives us the control we need to design our experiments in such a way that we can correctly interpret the results.

6.4.5 The Exam

The exam is a held-out set over which we compute the student's perplexity. The details of the exam are task dependent and we discuss these next.

6.5 EXPERIMENTAL DETAILS

Having defined the steps that we are taking on the road map in this work, we now test the validity of this setup with two tasks. Here, we describe and motivate these tasks in Section 6.5.1 and Section 6.5.2, and give the training details in Section 6.5.3.

6.5.1 Task 1 – Teaching Different Domains

For this task we design a language consisting of two strictly separated vocabularies, loosely representing two different domains in natural language. Specifically, $V_1 = \{a, b, c, d, e, f, g, h, i, j\}$, and $V_2 = \{k, l, m, n, o, p, q, r, s, t\}$. We construct sentences by randomly sampling from these sets. Sentences consist either of tokens only from V_1 or of tokens only from V_2. Sentences have an equal length of 10 tokens each. Half of the data set that the teacher can choose from consists of V_1 sentences, the other half consists of V_2 sentences. The teacher's LM is trained on a similarly constructed data set, yet consisting of different sentences. The student's exam set consists of sentences from only one of the vocabularies, V_1 in our case. These are different sentences than in the training set, i.e., the teacher cannot simply sample the exam set to train the student. Hence, the optimal teaching strategy is to present the student with sentences from the exam vocabulary. We confirm this in our baseline experiments that we present in Section 6.5.4.

6.5.2 Task 2 – Teaching Different Structures

For this task we do not use different vocabularies, but different sentence structures. All our sentences are constructed with V_1 and are between 2 and 10 tokens long. We use two different structures: single repetitions and double repetitions. In the case of the single repetitions two identical tokens never occur next to each other, whereas in the case of double repetitions tokens are sampled in pairs:

Structure 1 - Single repetitions: $(xy)^n$

Structure 2 - Double repetitions: (xx) or $(xxyy)^n$

The data set that the teacher can sample from consists for 20% of sentences with Structure 1 and for 80% of sentences of Structure 2. The exam set consists of sentences with Structure 1. We opt for this way of splitting the data, as we found that a student performs quite well when trained on data consisting half of Structure 1 and half of Structure 2. Having an unequal split thus allows us to make sure that we can appropriately distinguish a learned didactic approach from a random one. For this task the optimal teaching strategy is to select sentences with the exam structure, as we confirm with our baseline experiments that we present in Section 6.5.4.

6.5.3 Training Details

The teacher LM is trained on 100 unique sentences till convergence. The dataset the teacher can sample from for the student consists of 100 different unique sentences. The exam consists of 10 unique sentences and we set the teacher budget to 10 as well. In follow-up work we hope to experiment with different dataset sizes, as well as investigate the effect of different budgets. We run our experiments with five different random seeds and report the averages and standard deviations. We use the negative perplexity of the student on the exam as reward for the teacher. We experiment with two sentence embeddings for the teacher: average word embeddings and the average of the last hidden layer. We train students for a predefined number of steps that we determine by inspecting the loss and perplexity curves of training an LM once before the actual experiments. We base the threshold on when a student LM starts to overfit, so that a teacher can get clear feedback signals. We set this value to 400 for Task 1 and 300 for Task 2. Automatically determining when the student stops training is an important avenue for future work (Table 6.1). We use Fairseq's (Ott et al., 2019) `transformer_lm`[4] for the implementation of the transformer LMs. We use up to four GPUs with 32 GB RAM per experiment. The exact number depends on the number of students per teacher, as we can fit up to 6 students on a single GPU due to our multiprocessing implementation.

6.5.4 Baseline experiments

We run three baseline experiments with three different didactic strategies: an *oracle*, *random*, and *worst case* strategy. We run the baselines for five different random seeds. In each experiment, we randomly select data according to the teacher budget. We do this five times and each time train a student LM with the selected data. The difference between baselines is the type of data that can be selected. For the oracle baseline we only select sentences that consist of the exam vocabulary (Task 1) or structure (Task 2). For the random baseline we randomly select sentences. For the worst case baseline all sentences that we select are from a different vocabulary or structure than the exam sentences.

6.6 RESULTS

In this section we give the results of our experiments on both tasks.

6.6.1 Task 1 – Different Domains

Baseline Results

In Table 6.2 we present the results for the baseline experiments for Task 1. We report the averages and standard deviations of the perplexity on the exam set and the fraction of training sentences that consisted of the exam vocabulary. For space reasons, we report the results for two seeds per baseline: the seed with the best average perplexity and the worst. The results for all five seeds are given in Appendix 6.B. There we also present scores for the *n*-gram overlap between the selected training set and the exam set. The results are as expected:

[4] `https://fairseq.readthedocs.io/en/latest/command_line_tools.html`

Table 6.2: Baseline results Task 1. Averages and standard deviations reported based on five runs per seed.

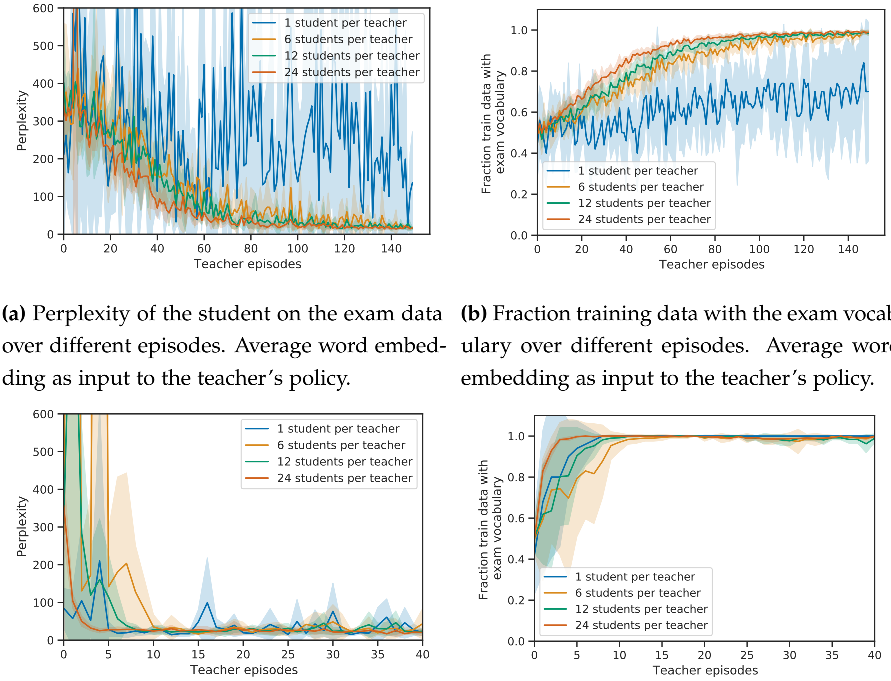
Type	**Seed**	**Avg**	**Avg train**
		Perplexity	**from test**
Random	best	$160.9_{\pm 217.7}$	$0.54_{\pm 0.16}$
	worst	$742.5_{\pm 159.8}$	$0.50_{\pm 0.17}$
Oracle	best	$14.99_{\pm 5.364}$	$1.00_{\pm 0.00}$
	worst	$68.95_{\pm 87.49}$	$1.00_{\pm 0.00}$
Worst case	best	$4.78e4_{\pm 2.67e4}$	$0.00_{\pm 0.00}$
	worst	$8.46e4_{\pm 4.69e4}$	$0.00_{\pm 0.00}$

The oracle baseline gives the best results, followed by the random and worst case baseline respectively.

Results of Training the Teacher

In Figure 6.3 we present the results for Task 1 for different numbers of students per teacher.[5] The teacher's didactic strategy correctly converges to the oracle baseline. There is a clear difference between different sentence embeddings (Section 6.4.1). Both embedding types are converging, but the average hidden layer embeddings are clearly superior. We investigate this further by plotting the t-SNE embeddings (van der Maaten and Hinton, 2008) of the different sentence embeddings in Figure 6.4. To prepare for Task 2, we also plot the embeddings of Task 2. The hidden layer sentence embeddings result in the clearest separation between sentences from different vocabularies or structures. Especially for Task 2, where we use the same vocabulary, this is unsurprising. From now on we opt for these sentence embeddings. Based on the results for Task 1 we opt for 12 students per teacher as a good trade-off between computational cost and convergence stability for Task 2.

[5] We present plots for the n-gram overlap in Appendix 6.D.



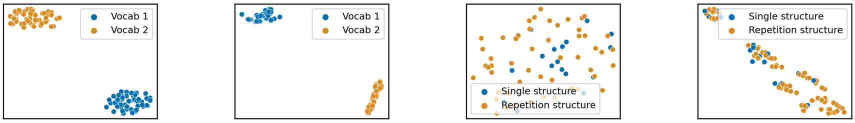
(a) Perplexity of the student on the exam data over different episodes. Average word embedding as input to the teacher's policy.

(b) Fraction training data with the exam vocabulary over different episodes. Average word embedding as input to the teacher's policy.

(c) Perplexity of the student on the exam data over different episodes. Average last hidden layer as input to the teacher's policy.

(d) Fraction training data with the exam vocabulary over different episodes. Average last hidden layer as input to the teacher's policy.

Figure 6.3: Results Task 1 – Different domains. Plots for different numbers of students per teacher. Results per setting reported as average and standard deviation over five random seeds. x-axis of lower plots bound to 40 as the teacher had already converged by then.



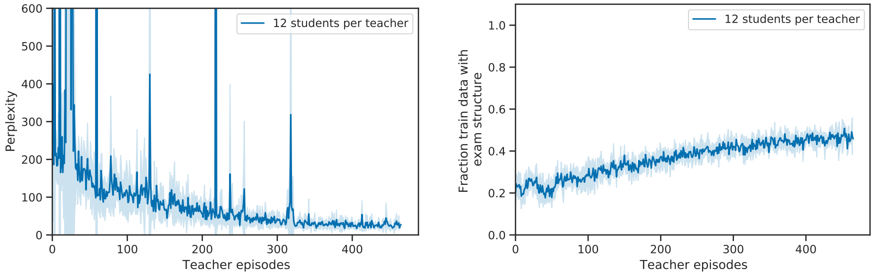
(a) Task 1 - Different vocabularies. Sentence embedding is average word embeddings.

(b) Task1 - Different vocabularies. Sentence embedding is average last hidden layer.

(c) Task 2 - Different structures. Sentence embedding is average word embeddings.

(d) Task 2 - Different structures. Sentence embedding is average last hidden layer.

Figure 6.4: T-SNE plots for different sentence representations for different tasks.



(a) Perplexity of the student on the exam data over different episodes.

(b) Fraction training data with the exam structure over different episodes.

Figure 6.5: Results Task 2 – Plots for 12 students per teacher. Results per setting reported as average and standard deviation over five random seeds.

Table 6.3: Baseline results Task 2. Averages and standard deviations reported based on five runs per seed.

Type	**Seed**	**Avg Perplexity**	**Avg train from test**
Random	best	119.0±56.48	0.18±0.04
	worst	342.1±241.4	0.12±0.08
Oracle	best	6.821±0.619	1.00±0.00
	worst	9.431±3.057	1.00±0.00
Worst Case	best	299.6±124.2	0.00±0.00
	worst	595.3±297.9	0.00±0.00

6.6.2 Task 2 – Different Structures

Baseline Results

We present the baseline results for Task 2 in Table 6.3. Again we report the results for the best and the worst seed. Full results are available in Appendix 6.C. Similar to the results for Task 1, we confirm that the oracle baseline performs strongest, followed by the random and worst case baseline respectively.

Results of Training the Teacher

In Figure 6.5 we present the results for Task 2.[6] Again we see that the teacher learns to gradually converge to the oracle teaching strategy, although convergence is less fast than for Task 1; we do not achieve full convergence in the number of training episodes that we run these experiments for. We postulate that this can be explained by the differences we found in Figure 6.4. The differences in sentence embeddings between the two different structures are clearly less apparent than between the sentences from two vocabularies. This indicates the importance of good sentence embeddings for future work. Moreover, as stated in Section 6.6.2, we found that transmitting roughly 50% of Structure 1 and 50% of Structure 2 also already leads to good performance. Therefore, the teacher likely needs to learn from a less distinct learning signal than in Task 1.

6.7 IMPLICATIONS AND OUTLOOK

We took the first steps on our proposed road map. Here we want to share our learnings and the limitations of the current setup to help future research to take the next steps on the road map.

6.7.1 The Importance of Designing Experiments with Interpretable Outcomes

We designed our experiments such that we knew the teacher's oracle strategy, which allowed us to properly test our setup. However, in designing our experiments we found that finding such settings is non-trivial. For example, in a task that contains a language with multiple structures, a student might unexpectedly learn information from structure 1 that also proves useful for structure 2. This might be acceptable if one's only objective is to obtain a good performance. However, in our case it is critical to be able to know that a teacher is "right for the right reasons", which motivated our choices for the tasks and languages.

[6] We present plots for the n-gram overlap in Appendix 6.E.

6.7.2 The Teacher's Budget

Following the objective as defined in Section 6.3, we designed our experiments in such a way that the teacher was given a budget that limits the amount of data it can send to the student. As mentioned in Section 6.5.3, we confirmed that the student's learning converges with this budget. In follow-up work we plan to investigate the importance of different budgets in more detail. One interesting direction is to give the teacher a flexible budget, i.e., such that a teacher could decide to stop training if it deems it no longer necessary for the student.

6.7.3 Computational Complexity

Apart from the multiprocessing setup that allows us to train multiple students on a single GPU, we did not yet focus on the computational complexity of our approach. In the current setup many student language models need to be trained for a single teacher. In our case we deem this justifiable as we are just at the start of the road map. Moreover, once a teacher model is trained, it can be used for many different purposes. However, in future work we hope to focus on decreasing the computational complexity of our approach. One promising avenue to do this is by optimizing the learning process of the student.

6.8 ETHICAL IMPACT STATEMENT

At this point we use artificial language data only, for which we do not see any direct negative implications. As we move towards using real data sets, it is necessary to be aware of potential biases with these data sets. One needs to ensure that the data is not biased towards any (protected) group to avoid any harm. Currently, much of the NLP research focuses on English as its language of interest. Our approach is not bound to any language in particular and can even be used to improve language learning in a low resource setting. Once the models achieve human-like performance and are used for downstream tasks and applications it is necessary to explicitly state that language is produced by an artificial language model. However, as with all language models, misuse can still happen, and it is our responsibility as a research community, amongst others, to spend effort on making users aware of these possibilities.

6.9 CONCLUSION

In this chapter we were inspired by the observation that human language acquisition is much more interactive in nature than the training regimes of large language models. Therefore, we explored the space of interactive language modeling in this chapter, answering the fifth research question of this thesis. We defined our objective for interactive language modeling, which makes use of a student-teacher framework. In this framework, the teacher is inspired by the caregiver and the student resembles the child in the human language acquisition. The teacher and student can interact with each other, and with the environment. We presented a road map that details the steps towards interactive language modeling for each of the components of the teacher-student loop. We led by example and took the first steps on this road map, leading to a tangible proof of concept of our proposal. With this work, we aim to inspire a broader discussion and research agenda around interactive language learning. In future work, we plan to take the next steps on the road map. We are especially interested in taking the steps that require a more explicit and elaborate form of interaction between the teacher, student, and their environment.

CHAPTER APPENDIX

In this chapter appendix we give more details about how we compute the probability of our Top-K sample, as discussed in Section 6.4.1 and additional results for the baseline experiments.

6.A COMPUTING THE PROBABILITY OF A TOP-K SAMPLE

Our objective is to find the (log) probability of sampling the subset $(i_1, \ldots, i_K)$ from $\{1, \ldots, N\}$ *without* replacement from the categorical probability $(p_1, \ldots, p_N)$.

Let us first consider sampling K elements from the $\{1, \ldots, N\}$ *with* replacement. In that case

$$p(i_1, \ldots, i_K) = \prod_{k=1}^{K} p_{i_k}. \tag{6.1}$$

If we allow for all possible permutations of observing $(i_1, \ldots, i_K)$ we get

$$p(i_1, \ldots, i_K) = C \prod_{k=1}^{K} p_{i_k}, \tag{6.2}$$

where $C = K!$.

To go from sampling *with* replacement, to sampling *without* replacement, we consider event A = "all sampled elements $(i_1, \ldots, i_K)$ are unique". Then

$$\begin{aligned} p_{\text{w/o replacement}}(i_1, \ldots, i_K) = \\ p_{\text{w/ replacement}}(i_1, \ldots, i_K | A). \end{aligned} \tag{6.3}$$

Applying Bayes Rule gives us:

$$p_{\text{w/o replacement}}(i_1, \ldots, i_K) = \frac{p_{\text{w/ replacement}}(A|i_1, ..., i_K) p_{\text{w/ replacement}}(i_1, \ldots, i_K)}{p_{\text{w/ replacement}}(A)}. \tag{6.4}$$

As in our case all samples in $(i_1, \ldots, i_K)$ are unique we know that

$$p_{\text{w/ replacement}}(A|i_1, \ldots, i_K) = 1. \tag{6.5}$$

Combining this with Equation 6.2 gives us

$$p_{\text{w/o replacement}}(i_1, \ldots, i_K) = \frac{C \prod_{k=1}^{K} p_{i_k}}{p(A)}, \tag{6.6}$$

and thus

$$p_{\text{w/o replacement}}(i_1, \ldots, i_K) \propto \prod_{k=1}^{K} p_{i_k}, \tag{6.7}$$

and

$$\log p_{\text{w/o replacement}}(i_1, \ldots, i_K) \propto \sum_{k=1}^{K} \log p_{i_k}. \tag{6.8}$$

From an implementation perspective this boils down to the following steps:

1. We compute the scores per sentence.
2. We sample K sentences without replacement, using the GumbelTopK trick.
3. We compute the log probabilities for each score: $\log \text{softmax}(scores)$.
4. We compute the log probability of our sample by adding the log probabilities of the elements in our sample, according to Equation 6.8.

6.A.1 *Comparison to Prior Work*

Our problem of sampling K sentences as a single action is similar to the problem formulation of using reinforcement learning for extractive summarization to optimize for Rouge (Lin, 2004) directly. In this setting K sentences need to be selected from a document. This results in a very large search space. Narayan et al. (2018b) limit the search space by first selecting n sentences that have a high Rouge score. Then all possible summaries are made with these n sentences. These summaries are ranked according to their Rouge scores and the top K sentences are taken as action. This approach has the disadvantage that it limits the search space heuristically, which does not guarantee that the best summary is found. Dong et al. (2018) frame the problem as a contextual bandit problem, which allows them to sample from the true action space. We choose our approach as it is intuitive, simple and effective.

6.B ADDITIONAL RESULTS BASELINE EXPERIMENTS TASK 1

In Table 6.B.1 we present the results for our baseline runs on all five seeds.

6.C ADDITIONAL RESULTS BASELINE EXPERIMENTS TASK 2

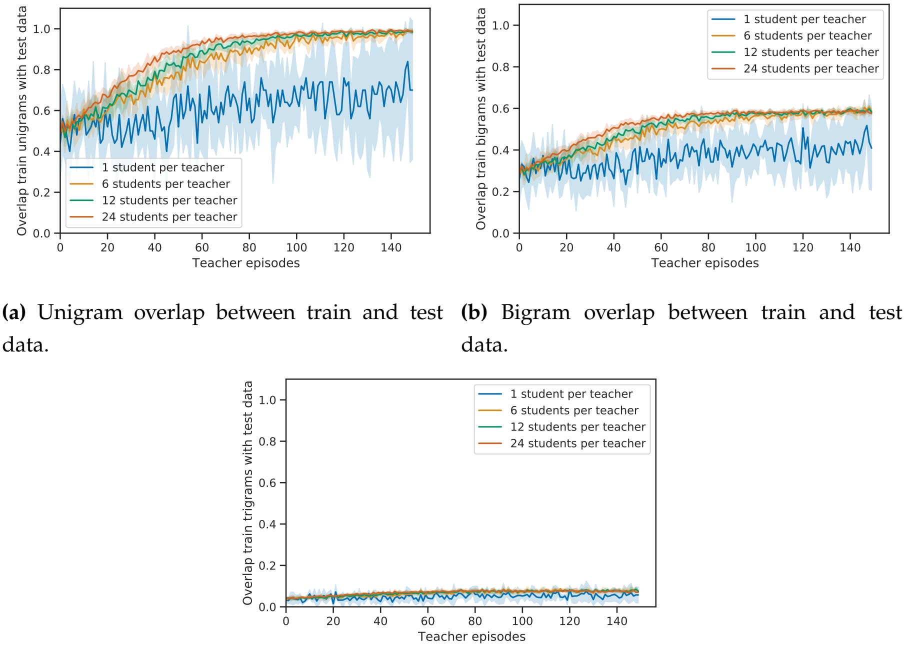
In Table 6.C.1 we present the results for our baseline runs on all five seeds.

6.D ADDITIONAL N-GRAM PLOTS TASK 1

In this section we present the plots for the n-gram overlap for Task 1 in Figures 6.D.1 and 6.D.2.

6.E ADDITIONAL N-GRAM PLOTS TASK 2

In this section we present the plots for the n-gram overlap for Task 2 in Figure 6.E.1.

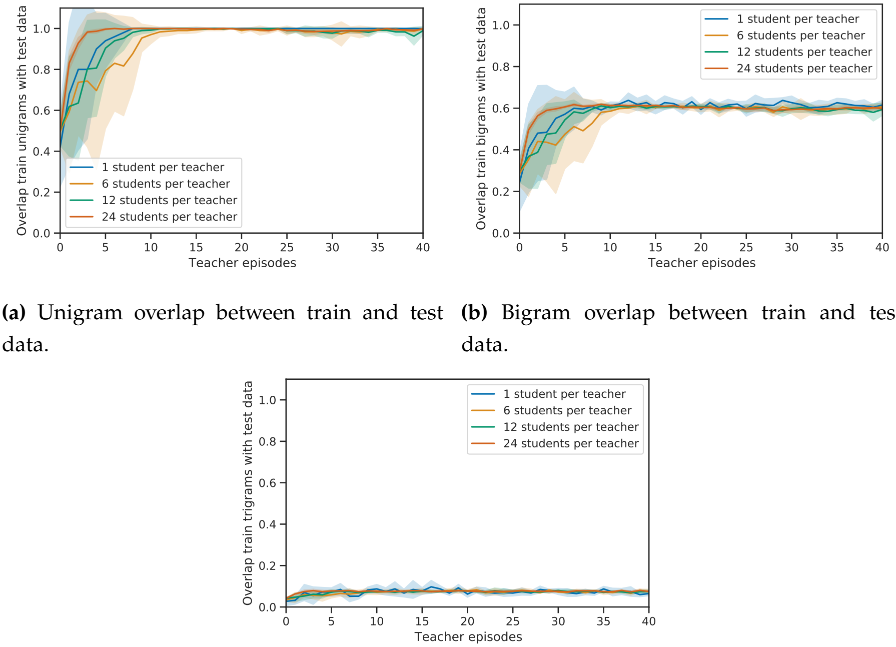


(a) Unigram overlap between train and test data.

(b) Bigram overlap between train and test data.

(c) Trigram overlap between train and test data.

Figure 6.D.1: Additional results Task 1 – Different domains. Plots for different numbers of students per teacher. Results per setting reported as average and standard deviation over five random seeds. Average word embedding as sentence embeddings.



(a) Unigram overlap between train and test data.

(b) Bigram overlap between train and test data.

(c) Trigram overlap between train and test data.

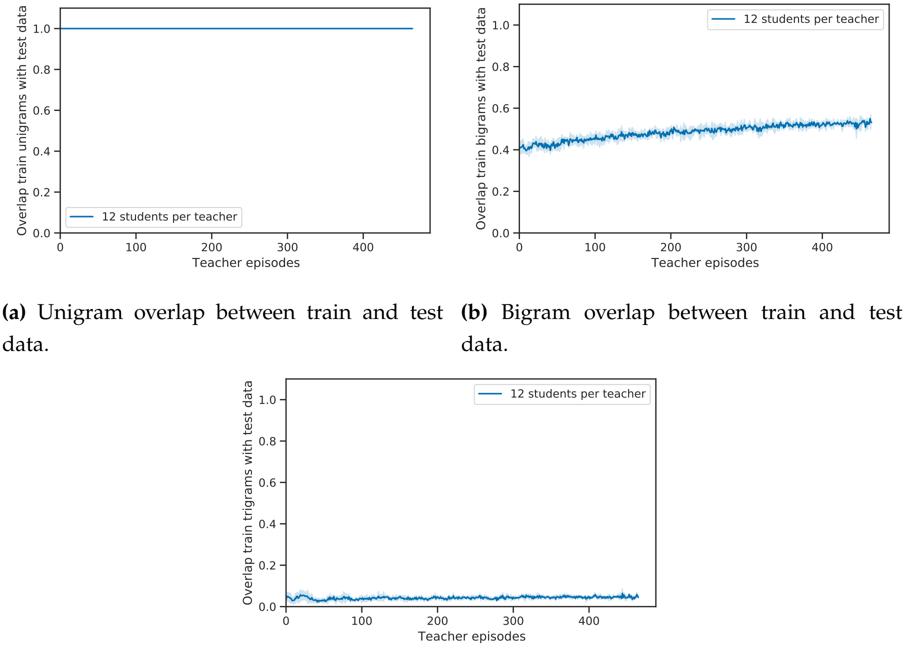
Figure 6.D.2: Additional results Task 1 – Different domains. Plots for different numbers of students per teacher. Results per setting reported as average and standard deviation over five random seeds. Average hidden layer embedding as sentence embeddings.

Table 6.B.1: Baseline results for Task 1. Different domains. Averages and standard deviations reported based on five runs per seed.

	Seed	**Avg Perplexity**	**Avg train from test**	**Avg 1-gram overlap**	**Avg 2-gram overlap**	**Avg 3-gram overlap**
Random	6639	$193.9_{\pm 100.3}$	$0.46_{\pm 0.14}$	$0.46_{\pm 0.14}$	$0.278_{\pm 0.07}$	$0.023_{\pm 0.009}$
	7519	$683.1_{\pm 634.3}$	$0.52_{\pm 0.15}$	$0.52_{\pm 0.15}$	$0.291_{\pm 0.10}$	$0.030_{\pm 0.010}$
	1007	$742.5_{\pm 159.8}$	$0.50_{\pm 0.17}$	$0.50_{\pm 0.17}$	$0.298_{\pm 0.10}$	$0.035_{\pm 0.014}$
	4520	$160.9_{\pm 217.7}$	$0.54_{\pm 0.16}$	$0.54_{\pm 0.16}$	$0.327_{\pm 0.09}$	$0.035_{\pm 0.025}$
	4527	$307.1_{\pm 295.1}$	$0.58_{\pm 0.17}$	$0.58_{\pm 0.17}$	$0.349_{\pm 0.10}$	$0.035_{\pm 0.014}$
Oracle	6639	$14.99_{\pm 5.364}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.551_{\pm 0.06}$	$0.072_{\pm 0.029}$
	7519	$44.37_{\pm 58.94}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.611_{\pm 0.02}$	$0.085_{\pm 0.017}$
	1007	$68.95_{\pm 87.49}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.598_{\pm 0.02}$	$0.077_{\pm 0.025}$
	4520	$15.65_{\pm 4.616}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.578_{\pm 0.02}$	$0.087_{\pm 0.028}$
	4527	$23.66_{\pm 21.44}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.624_{\pm 0.02}$	$0.095_{\pm 0.019}$
Worst case	6639	$8.46e4_{\pm 4.69e4}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$
	7519	$7.03e4_{\pm 3.73e4}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$
	1007	$8.17e4_{\pm 4.26e4}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$
	4520	$4.78e4_{\pm 2.67e4}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$
	4527	$6.69e4_{\pm 1.98e4}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$

Table 6.C.1: Baseline results for Task 2. Different structures. Averages and standard deviations reported based on five runs per seed.

	Seed	Avg Perplexity	Avg train from test	Avg 1-gram overlap	Avg 2-gram overlap	Avg 3-gram overlap
Random	6639	119.0±56.48	0.18±0.04	1.00±0.00	0.401±0.033	0.030±0.020
	7519	162.8±201.9	0.24±0.05	1.00±0.00	0.408±0.044	0.035±0.038
	1007	234.1±192.0	0.24±0.12	1.00±0.00	0.414±0.034	0.034±0.020
	4520	161.7±190.6	0.22±0.04	1.00±0.00	0.410±0.023	0.038±0.033
	4527	342.1±241.4	0.12±0.08	1.00±0.00	0.348±0.024	0.013±0.017
Oracle	6639	6.973±1.534	1.00±0.00	1.00±0.00	0.720±0.044	0.151±0.022
	7519	7.626±2.298	1.00±0.00	1.00±0.00	0.682±0.056	0.177±0.033
	1007	7.895±1.106	1.00±0.00	1.00±0.00	0.726±0.045	0.207±0.025
	4520	6.821±0.619	1.00±0.00	1.00±0.00	0.740±0.073	0.197±0.054
	4527	9.431±3.057	1.00±0.00	1.00±0.00	0.700±0.056	0.174±0.017
Worst case	6639	595.3±297.9	0.00±0.00	1.00±0.00	0.326±0.026	0.00±0.00
	7519	317.2±235.8	0.00±0.00	1.00±0.00	0.311±0.018	0.00±0.00
	1007	508.1±155.7	0.00±0.00	1.00±0.00	0.345±0.017	0.00±0.00
	4520	299.6±124.2	0.00±0.00	1.00±0.00	0.310±0.027	0.00±0.00
	4527	432.8±72.05	0.00±0.00	1.00±0.00	0.330±0.035	0.00±0.00



(a) Unigram overlap between train and test data.

(b) Bigram overlap between train and test data.

(c) Trigram overlap between train and test data.

Figure 6.E.1: Additional results Task 2 – Different structures. Results per setting reported as average and standard deviation over five random seeds.

7 CONCLUSION

In this thesis we have taken a human-centered approach to NLP. We have visited a variety of tasks, and provided insights into users' needs for these tasks. We have also taken inspiration from cognitive science and linguistics to develop models and systems for these tasks. We have shown how a human-centered approach can help us (i) understand model behavior and capabilities, (ii) identify where and how modeling can be improved, and (iii) make sure models are in line with users' needs. Throughout the thesis we found that there are still opportunities to more adequately model a human-centered approach to NLP, and thus this thesis also aims to be the start of new directions in this area.

In this final chapter of the thesis we revisit the main research questions that we raised in Chapter 1. We summarize our findings for these questions in Section 7.1. We conclude this chapter, and this thesis, with directions for future work in Section 7.2.

7.1 SUMMARY OF FINDINGS

Research Question 1: *What does document-centered assistance look like, and how can we model it?*

In Chapter 2 and in (ter Hoeve et al., 2020) we found that users' information needs in the context of document-centered assistance is different from their information needs in other areas of question-answering or digital assistance. In a document-centered scenario users ask questions that are directly grounded in the process of writing or consuming the document, such as *"Does the document already contain information about topic X?"*. In order to model this new scenario, we collected a human-labeled dataset that we called Document Question-Answering, which contains questions and answers grounded in this document-centered scenario. We trained passage ranking and answer selection models on this data and showed that these models perform well on our task, but that there is also still room for improvement.

Research Question 2: *What makes a good and useful summary for users of automatically generated summaries?*

In Chapter 3 and in (ter Hoeve et al., 2022d) we contributed a survey methodology that can be used to answer this question for many user groups. We focused on university students, as they are heavy users of pre-made summaries, for example during exam preparations. We centered our survey around the three context factors for automatic summarization (Spärck Jones, 1998): (i) *input*, (ii) *purpose*, and (iii) *output factors*. We found that survey participants indicated many different needs for a pre-made summary, many of which are different from the focus of automatic summarization research at the time of writing. We also contributed an evaluation methodology to measure the usefulness of a generated summary.

Research Question 3: *How can we fulfill users' request for summaries that include graphical elements?*

Motivated by our findings for the second research question, we proceeded to generate summaries with graphical elements in Chapter 4 and in (ter Hoeve et al., 2022c). We called our task *summarization with graphical elements*. In formulating our task, we were also inspired by the cognitive science literature on how humans read written texts. By means of a user study, we confirmed that a critical mass of people is interested in our proposed summaries. Next, we collected a high quality, human-labeled test set for our task, which we called GraphelSums. Finally, we proposed a number of baseline methods for the task, ranging from heuristically labeling the data, to training and fine-tuning neural models on weakly labeled training data. Although the results are promising, none of these baseline methods achieve satisfactory performance yet, indicating the difficulty of the task. Hence, this chapter opens the door to many future research directions.

Research Question 4: *How are low-resource investigations in NLP biased by high-resource approaches?*

In Chapter 5 and in (ter Hoeve et al., 2022a) we expanded our focus from English to a wider variety of languages. Specifically, we were interested in languages that do not have as many available resources. We observed that research on these lower resource languages is often grounded in high-resource approaches. For example, low-resource scenarios are frequently simulated by downsampling from a high-resource dataset. We empirically investigated the validity of this approach on two well-known NLP tasks: POS-tagging and machine translation. We showed that this type of downsampling introduces a bias in the dataset statistics of the downsample. Next, we showed that this results in a biased view regarding the model performance on these two tasks. Scores are either too high (because of the richness of the data) or rather too low (because the high-resource data may be of insufficient quality). Being aware of this bias puts individual researchers and the field as a whole in a better position.

Research Question 5: *How can we make artificial language modeling more human-like by taking a more interactive approach?*

In Chapter 6 and in (ter Hoeve et al., 2021) we were motivated by the observation that, from the perspective of human language acquisition, large language models are trained in an unnatural fashion. Human language acquisition is much more interactive in nature. Fascinated by this observation, we explored the possibilities of using interaction more actively in artificial language modeling, which we called *interactive language modeling*. To model such an interactive approach, we suggested using a teacher-student framework, in which the teacher is loosely inspired by a caregiver, and the student by a child. The teacher and the student can interact with each other, and with the environment. Next, we proposed a road map towards interactive language modeling, which includes steps for each of the components in the student-teacher framework. We took the first steps on this road map, by which we showed the initial feasibility of our approach. This work was exploratory in nature. In future work we plan to take the next steps on the road map, especially focusing on expanding the interactive nature of the approach.

7.2 FUTURE WORK

An important learning from our journey is that taking a human-centered approach opens the door to many new research directions, of which we have discussed multiple throughout this thesis. We believe that these are all important avenues to continue working on. In this section we discuss opportunities for future work that we have not explicitly addressed yet.

Different Tasks, Users, and Languages

Throughout this thesis we have touched upon multiple tasks and domains, yet there are still many possibilities for a user-centered focus that we have not covered yet.

We believe that there is an increasing need for human-centered NLP in the context of large language modeling, and its downstream tasks. In this thesis we were motivated to make language modeling more human-like, but we strongly believe there are also many challenges regarding the societal impact of these models. As model performance increases (e.g., Devlin et al., 2019; Brown et al., 2020; Chowdhery et al., 2022), these models will increasingly be used for user facing implementations (e.g., Chung et al., 2022; Vaithilingam et al., 2022). This requires additional efforts from a human-centered NLP perspective to ensure that these models are well in line with users' needs. Fortunately, we see an increase in recent work in this area (e.g., Bender et al., 2021; Crisan et al., 2022; Ouyang et al., 2022; Weidinger et al., 2022).

This naturally brings us to another important aspect — the variety of users that are investigated. In this thesis we have focused our user studies and surveys on students from Dutch universities, and English speaking, U.S. based crowd workers. It is essential to expand the variety of users for a complete human-centered investigation.

One of these facets is the language that we focus on. With English as the default language within NLP, we are in the undesirable situation that research on any other language is often seen as "language specific" (Bender, 2011), enforcing an English language bias in NLP. In this thesis we investigated this bias in more detail in Chapter 5. We view detailed human-centered NLP research beyond English and a few other higher resource languages as a vital avenue for future work.

One important way to achieve the aforementioned directions is to explicitly include evaluation metrics that measure these aspects. We addressed this partially in Chapter 3, where we proposed an evaluation methodology for usefulness. We advocate for a holistic evaluation methodology, in which metrics like fairness and diversity are considered as important as metrics like accuracy and F1-score. In our view, the former are still too often seen as a side product, or as a "nice to have".

The Role of Cognitive Science and Linguistics

The role of cognitive science and linguistics in NLP research has become particularly interesting with the increased performance of large language models, which was an important motivation in Chapter 6. Besides the question regarding whether we should make the training regime more human-like, we believe that there are still important questions to answer that will improve our understanding of these models and their performance. For example, what are the requirements for a model to master language as humans do? Currently, large improvements are achieved by scaling (e.g., Devlin et al., 2019; Brown et al., 2020; Chowdhery et al., 2022). However, whether these models are trustworthy enough to be released to the general public, is still subject to discussion (Bender et al., 2021; Markov et al., 2022). One important argument involves the harmfulness of these models when their performance is not in line with societal needs, which is related to the previous section on user-centered NLP. Another important question is whether models need to achieve human-level language understanding before we could fully trust and use them (Bender and Koller, 2020). For example, is human-level language understanding needed to avoid models to hallucinate (e.g., Ji et al., 2022)? We believe that insights from cognitive science and linguistics are important to efficiently improve model performance on these aspects.

BIBLIOGRAPHY

Adelani, David Ifeoluwa, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei (2021). "MasakhaNER: Named Entity Recognition for African Languages." In: *Transactions of the Association for Computational Linguistics* 9.

Agić, Željko, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard (2016). "Multilingual Projection for Parsing Truly Low-Resource Languages." In: *Transactions of the Association for Computational Linguistics* 4.

Agić, Željko and Ivan Vulić (July 2019). "JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics.

Ai, Qingyao, Susan T Dumais, Nick Craswell, and Dan Liebling (2017). "Characterizing email search using large-scale behavioral logs and surveys." In: *Proceedings of the 26th International Conference on World Wide Web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Akhbardeh, Farhad, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri (Nov. 2021). "Findings of the 2021 Conference on Machine Translation (WMT21)." In: *Proceedings of the Sixth Conference on Machine Translation*. Online: Association for Computational Linguistics.

Amhag, Lisbeth, Lisa Hellström, and Martin Stigmar (2019). "Teacher educators' use of digital tools and needs for digital competence in higher education." In: *Journal of Digital Learning in Teacher Education* 35.4.

Amplayo, Reinald Kim and Mirella Lapata (July 2020). "Unsupervised Opinion Summarization with Noising and Denoising." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Andrade, Chittaranjan (2018). "Internal, external, and ecological validity in research design, conduct, and evaluation." In: *Indian journal of psychological medicine* 40.5, pp. 498–499.

Angelidis, Stefanos, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata (2021). "Extractive Opinion Summarization in Quantized Transformer Spaces." In: *Transactions of the Association for Computational Linguistics* 9.

Araabi, Ali and Christof Monz (Dec. 2020). "Optimizing Transformer for Low-Resource Neural Machine Translation." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Ariannezhad, Mozhdeh (2022). "Understanding and Learning from User Behavior for Recommendation in Multi-channel Retail." In: *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II*. Vol. 13186. Lecture Notes in Computer Science. Springer.

Ariannezhad, Mozhdeh, Mohamed Yahya, Edgar Meij, Sebastian Schelter, and Maarten de Rijke (2022). "Understanding Financial Information Seeking Behavior from User Interactions with Company Filings." In: *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*. ACM.

Aries, Abdelkrime, Djamel Eddine Zegour, and Walid-Khaled Hidouci (2019). "Automatic text summarization: What has been done and what has to be done." In: *CoRR* abs/1904.00688. arXiv: 1904.00688.

Ashok, Vikas, Yevgen Borodin, Yury Puzis, and IV Ramakrishnan (2015). "Capti-speak: a speech-enabled web screen reader." In: *Proceedings of the 12th Web for All Conference*. ACM.

Aulamo, Mikko, Sami Virpioja, and Jörg Tiedemann (July 2020). "OpusFilter: A Configurable Parallel Corpus Filtering Toolbox." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics.

Baan, Joris, Maartje ter Hoeve, Marlies van der Wees, Anne Schuth, and Maarten de Rijke (2019a). "Do Transformer Attention Heads Provide Transparency in Abstractive Summarization?" In: *CoRR* abs/1907.00570. arXiv: 1907.00570.

Baan, Joris, Maartje ter Hoeve, Marlies van der Wees, Anne Schuth, and Maarten de Rijke (2019b). "Understanding Multi-Head Attention in Abstractive Summarization." In: *CoRR* abs/1911.03898. arXiv: 1911.03898.

Bai, Yu, Yang Gao, and Heyan Huang (Aug. 2021). "Cross-Lingual Abstractive Summarization with Limited Parallel Resources." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics.

Baker, Anne E, Jan Don, and Kees Hengeveld (2012). *Taal en taalwetenschap*. John Wiley & Sons.

Balasuriya, Saminda Sundeepa, Laurianne Sitbon, Jinglan Zhang, and Khairi Anuar (2021). "Summary and Prejudice: Online Reading Preferences of Users with Intellectual Disability."

In: *CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia, March 14-19, 2021*. ACM.

Bañón, Marta, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza (July 2020). "ParaCrawl: Web-Scale Acquisition of Parallel Corpora." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Barrault, Loïc, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri (Nov. 2020a). "Findings of the 2020 Conference on Machine Translation (WMT20)." In: *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics.

Barrault, Loïc, Magdalena Biesialska, Marta R. Costa-jussà, Fethi Bougares, and Olivier Galibert (Nov. 2020b). "Findings of the First Shared Task on Lifelong Learning Machine Translation." In: *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics.

Barrault, Loïc, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri (Aug. 2019). "Findings of the 2019 Conference on Machine Translation (WMT19)." In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics.

Batali, John (1994). "Artificial evolution of syntactic aptitude." In: *Proceedings from the Sixteenth Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum Associates Hillsdale, NJ.

Bender, Emily M (2011). "On achieving and evaluating language-independence in NLP." In: *Linguistic Issues in Language Technology* 6.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*. Ed. by Madeleine Clare Elish, William Isaac, and Richard S. Zemel. ACM.

Bender, Emily M. and Alexander Koller (July 2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston (2009). "Curriculum learning." In: *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*. Vol. 382. ACM International Conference Proceeding Series. ACM.

Bennett, Paul N., Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui (2012). "Modeling the impact of short- and long-term behavior on search personalization." In: *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*. ACM.

Bickel, Balthasar and Johanna Nichols (2013). "Inflectional Synthesis of the Verb." In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Bird, Steven (May 2022). "Local Languages, Third Spaces, and other High-Resource Scenarios." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics.

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia (Aug. 2013). "Findings of the 2013 Workshop on Statistical Machine Translation." In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics.

Bojar, Ondřej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna (June 2014). "Findings of the 2014 Workshop on Statistical Machine Translation." In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA: Association for Computational Linguistics.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi (Sept. 2017). "Findings of the 2017 Conference on Machine Translation (WMT17)." In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri (Aug. 2016). "Findings of the 2016 Conference on Machine Translation." In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi (Sept. 2015). "Findings of the 2015 Workshop on Statistical Machine Translation." In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics.

Bojar, Ondřej, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz (Oct. 2018). "Findings of the 2018 Conference on Machine Translation (WMT18)." In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics.

Bordes, Antoine, Nicolas Usunier, Sumit Chopra, and Jason Weston (2015). "Large-scale Simple Question Answering with Memory Networks." In: *CoRR* abs/1506.02075. arXiv: `1506.02075`.

Borlund, Pia (2003). "The IIR evaluation model: a framework for evaluation of interactive information retrieval systems." In: *Information Research* 8.3.

Borlund, Pia (2016). “A study of the use of simulated work task situations in interactive information retrieval evaluations: A meta-evaluation.” In: *J. Documentation* 72.3.

Bota, Horatiu, Adam Fourney, Susan T Dumais, Tomasz L Religa, and Robert Rounthwaite (2018). “Characterizing Search Behavior in Productivity Software.” In: *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. ACM.

Bražinskas, Arthur, Mirella Lapata, and Ivan Titov (Nov. 2021). “Learning Opinion Summarizers by Selecting Informative Reviews.” In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). “Language Models are Few-Shot Learners.” In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Bruner, Jerome (1985). “Child’s talk: Learning to use language.” In: *Child Language Teaching and Therapy* 1.1.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder (June 2007). “(Meta-) Evaluation of Machine Translation.” In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder (June 2008). “Further Meta-Evaluation of Machine Translation.” In: *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, Ohio: Association for Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan (July 2010). “Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation.” In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Uppsala, Sweden: Association for Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia (June 2012). “Findings of the 2012 Workshop on Statistical Machine Translation.” In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada: Association for Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder (Mar. 2009). “Findings of the 2009 Workshop on Statistical Machine Translation.” In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece: Association for Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar Zaidan (July 2011). “Findings of the 2011 Workshop on Statistical Machine Translation.” In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics.

Cambazoglu, Berkant Barla, Valeria Bolotova-Baranova, Falk Scholer, Mark Sanderson, Leila Tavakoli, and W. Bruce Croft (2021). "Quantifying Human-Perceived Answer Utility in Non-factoid Question Answering." In: *CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia, March 14-19, 2021*. ACM.

Cao, Meng, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung (Nov. 2020). "Factual Error Correction for Abstractive Summarization Models." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics.

Cao, Ziqiang, Wenjie Li, Sujian Li, and Furu Wei (July 2018). "Retrieve, Rerank and Rewrite: Soft Template Based Neural Summarization." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics.

Cardenas, Ronald, Matthias Galle, and Shay B. Cohen (2021). "Unsupervised Extractive Summarization by Human Memory Simulation." In: *CoRR* abs/2104.08392. arXiv: `2104.08392`.

Carroll, David W (2008). *Psychology of language, Fifth Edition*. Thomson Brooks.

Caswell, Isaac, Theresa Breiner, Daan van Esch, and Ankur Bapna (Dec. 2020). "Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Chaabouni, Rahma, Roberto Dessì, and Eugene Kharitonov (2021). "Can Transformers Jump Around Right in Natural Language? Assessing Performance Transfer from SCAN." In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2021, Punta Cana, Dominican Republic, November 11, 2021*. Association for Computational Linguistics.

Chan, Yee Seng and Dan Roth (June 2011). "Exploiting Syntactico-Semantic Structures for Relation Extraction." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics.

Chang, Ernie, Hui-Syuan Yeh, and Vera Demberg (Apr. 2021). "Does the Order of Training Samples Matter? Improving Neural Data-to-Text Generation with Curriculum Learning." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics.

Chaudhary, Aditi, Antonios Anastasopoulos, Zaid Sheikh, and Graham Neubig (2021). "Reducing Confusion in Active Learning for Part-Of-Speech Tagging." In: *Transactions of the Association for Computational Linguistics* 9.

Chen, Danqi, Adam Fisch, Jason Weston, and Antoine Bordes (July 2017). "Reading Wikipedia to Answer Open-Domain Questions." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics.

Cheng, Jianpeng and Mirella Lapata (Aug. 2016). "Neural Summarization by Extracting Sentences and Words." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics.

Chopra, Sumit, Michael Auli, and Alexander M. Rush (June 2016). "Abstractive Sentence Summarization with Attentive Recurrent Neural Networks." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics.

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel (2022). "PaLM: Scaling Language Modeling with Pathways." In: *CoRR* abs/2204.02311. arXiv: 2204.02311.

Christiano, Paul F., Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei (2017). "Deep Reinforcement Learning from Human Preferences." In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett.

Chronopoulou, Alexandra, Dario Stojanovski, and Alexander Fraser (Nov. 2020). "Reusing a Pretrained Language Model on Languages with Limited Corpora for Unsupervised NMT." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics.

Chung, John Joon Young, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang (2022). "TaleBrush: Sketching Stories with Generative Pretrained Language Models." In: *CHI Conference on Human Factors in Computing Systems*.

Clark, Eve V (2018). "Conversation and language acquisition: A pragmatic approach." In: *Language Learning and Development* 14.3.

Clark, Herbert H and S Haviland (1977). "Comprehension and the Given-New Contract." In: *R. O. Freedly (Ed.) Discourse Production and Comprehension*.

Clark, Herbert H and Susan E Haviland (1974). "Psychological processes as linguistic explanation." In: *Explaining linguistic phenomena*.

Cohan, Arman, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian (June 2018). "A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics.

Cohn, David A., Zoubin Ghahramani, and Michael I. Jordan (1996). "Active Learning with Statistical Models." In: *J. Artif. Intell. Res.* 4.

Costa-jussà, Marta R., James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang (2022). "No Language Left Behind: Scaling Human-Centered Machine Translation." In: *CoRR* abs/2207.04672. arXiv: `2207.04672`.

Crisan, Anamaria, Margaret Drouhard, Jesse Vig, and Nazneen Rajani (2022). "Interactive Model Cards: A Human-Centered Approach to Model Documentation." In: *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM.

Cristia, Alejandrina, Emmanuel Dupoux, Nan Bernstein Ratner, and Melanie Soderstrom (2019). "Segmentability differences between child-directed and adult-directed speech: A systematic test with an ecologically valid corpus." In: *Open Mind* 3.

Dagan, Gautier, Dieuwke Hupkes, and Elia Bruni (Apr. 2021). "Co-evolution of language and agents in referential games." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics.

Daniel, Jeanne E., Willie Brink, Ryan Eloff, and Charles Copley (July 2019). "Towards Automating Healthcare Question Answering in a Noisy Multilingual Low-Resource Setting." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics.

Dehghani, Mostafa, Hosein Azarbonyad, Jaap Kamps, and Maarten de Rijke (2019). "Learning to Transform, Combine, and Reason in Open-Domain Question Answering." In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. WSDM '19. Melbourne VIC, Australia: ACM. ISBN: 978-1-4503-5940-5.

Dehouck, Mathieu and Carlos Gómez-Rodríguez (Dec. 2020). "Data Augmentation via Subtree Swapping for Dependency Parsing of Low-Resource Languages." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Deng, Yang, Wenxuan Zhang, and Wai Lam (Nov. 2020). "Multi-hop Inference for Question-driven Summarization." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics.

Di Geronimo, Linda, Maria Husmann, and Moira C Norrie (2016). "Surveying personal device ecosystems with cross-device applications in mind." In: *Proceedings of the 5th ACM International Symposium on Pervasive Displays*. ACM.

Ding, Bosheng, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao (Nov. 2020). "DAGA: Data Augmentation with a Generation Approach for Low-resource Tagging Tasks." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics.

Dong, Yue, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung (2018). "BanditSum: Extractive Summarization as a Contextual Bandit." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics.

Dorr, Bonnie, Christof Monz, Stacy President, Richard Schwartz, and David Zajic (June 2005). "A Methodology for Extrinsic Evaluation of Text Summarization: Does ROUGE Correlate?" In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics.

DUC (2003). *DUC 2003: Documents, Tasks, and Measures*. `https://duc.nist.gov/duc2003/tasks.html`.

Dunn, Matthew, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho (2017). "SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine." In: *CoRR* abs/1704.05179. arXiv: `1704.05179`.

Durmus, Esin, He He, and Mona Diab (July 2020). "FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier (2018). "Understanding Back-Translation at Scale." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics.

Ein-Dor, Liat, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim (Nov. 2020). "Active Learning for BERT: An Empirical Study." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics.

Fadaee, Marzieh, Arianna Bisazza, and Christof Monz (July 2017). "Data Augmentation for Low-Resource Neural Machine Translation." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics.

Falke, Tobias and Iryna Gurevych (Sept. 2017). "Bringing Structure into Summaries: Crowdsourcing a Benchmark Corpus of Concept Maps." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics.

Feigenblat, Guy, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov (Nov. 2021). "TWEETSUMM - A Dialog Summarization Dataset for Customer Service." In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Fourney, Adam and Susan T Dumais (2016). "Automatic identification and contextual reformulation of implicit system-related queries." In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM.

French, Robert M (1999). "Catastrophic forgetting in connectionist networks." In: *Trends in cognitive sciences* 3.4.

Gan, Wee Chung and Hwee Tou Ng (July 2019). "Improving the Robustness of Question Answering Systems to Question Paraphrasing." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics.

Gehrmann, Sebastian, Yuntian Deng, and Alexander Rush (2018). "Bottom-Up Abstractive Summarization." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics.

Gers, Felix A. and Jürgen Schmidhuber (2001). "LSTM recurrent networks learn simple context-free and context-sensitive languages." In: *IEEE Trans. Neural Networks* 12.6.

Gillis, Steven and Annemarie Schaerlaekens (2000). *Kindertaalverwerving: Een handboek voor het Nederlands*.

Goodrich, Ben, Vinay Rao, Peter J. Liu, and Mohammad Saleh (2019). "Assessing The Factual Accuracy of Generated Text." In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*. ACM.

Grbovic, Mihajlo, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, and Narayan Bhamidipati (2015). "Context-and content-aware embeddings for query rewriting in sponsored search." In: *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. ACM.

Gu, Jiatao, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho (2018). "Meta-Learning for Low-Resource Neural Machine Translation." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics.

Haddow, Barry, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch (Sept. 2022). "Survey of Low-Resource Machine Translation." In: *Computational Linguistics* 48.3.

Hämäläinen, Mika (2021). "Endangered Languages are not Low-Resourced!" In: *CoRR* abs/2103.09567. arXiv: `2103.09567`.

Hashim, Harwati (2018). "Application of technology in the digital era education." In: *International Journal of Research in Counseling and Education* 2.1.

Haviland, Susan E and Herbert H Clark (1974). "What's new? Acquiring new information as a process in comprehension." In: *Journal of verbal learning and verbal behavior* 13.5.

Hedderich, Michael A., David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow (Nov. 2020). "Transfer Learning and Distant Supervision for Multilingual Transformer Models: A Study on African Languages." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics.

Hedderich, Michael A., Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow (June 2021). "A Survey on Recent Approaches for Natural Language Processing in Low-Resource

Scenarios." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics.

Hendriksen, Mariya, Ernst Kuiper, Pim Nauts, Sebastian Schelter, and Maarten de Rijke (2020). "Analyzing and Predicting Purchase Intent in E-commerce: Anonymous vs. Identified Customers." In: *SIGIR Workshop On eCommerce*.

Hermann, Karl Moritz, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom (2015). "Teaching Machines to Read and Comprehend." In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*.

Hirsch, Eran, Alon Eirew, Ori Shapira, Avi Caciularu, Arie Cattan, Ori Ernst, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, and Ido Dagan (Nov. 2021). "iFacetSum: Coreference-based Interactive Faceted Summarization for Multi-Document Exploration." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Hu, Baotian, Qingcai Chen, and Fangze Zhu (Sept. 2015). "LCSTS: A Large Scale Chinese Short Text Summarization Dataset." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics.

Huang, Jin, Harrie Oosterhuis, and Maarten de Rijke (2022). "It Is Different When Items Are Older: Debiasing Recommendations When Selection Bias and User Preferences Are Dynamic." In: *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*. ACM.

Hupkes, Dieuwke, Verna Dankers, Mathijs Mul, and Elia Bruni (2020). "Compositionality Decomposed: How do Neural Networks Generalise?" In: *J. Artif. Intell. Res.* 67.

Hupkes, Dieuwke, Sara Veldhoen, and Willem H. Zuidema (2018). "Visualisation and 'Diagnostic Classifiers' Reveal How Recurrent and Recursive Neural Networks Process Hierarchical Structure." In: *J. Artif. Intell. Res.* 61.

Irvine, Ann and Chris Callison-Burch (June 2014). "Hallucinating Phrase Translations for Low Resource MT." In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan: Association for Computational Linguistics.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung (2022). "Survey of hallucination in natural language generation." In: *ACM Computing Surveys*.

Jin, Xisen, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren (July 2022). "Lifelong Pretraining: Continually Adapting Language Models to Emerging Corpora." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics.

Jokela, Tero, Jarno Ojala, and Thomas Olsson (2015). "A diary study on combining multiple information devices in everyday activities and tasks." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM.

Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy (2020). "SpanBERT: Improving Pre-training by Representing and Predicting Spans." In: *Transactions of the Association for Computational Linguistics* 8.

Joshi, Mandar, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer (July 2017). "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics.

Ju, Jiaxin, Ming Liu, Huan Yee Koh, Yuan Jin, Lan Du, and Shirui Pan (Nov. 2021). "Leveraging Information Bottleneck for Scientific Document Summarization." In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Junczys-Dowmunt, Marcin (Oct. 2018). "Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora." In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics.

Jurafsky, D. and J.H. Martin (2014). *Speech and Language Processing. Second Edition.* Pearson Education Limited. ISBN: 9781292025438.

Kalchbrenner, Nal and Phil Blunsom (Oct. 2013). "Recurrent Continuous Translation Models." In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics.

Kann, Katharina, Ophélie Lacroix, and Anders Søgaard (2020). "Weakly Supervised POS Taggers Perform Poorly on *Truly* Low-Resource Languages." In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press.

Karlson, Amy K, Shamsi T Iqbal, Brian Meyers, Gonzalo Ramos, Kathy Lee, and John C Tang (2010). "Mobile taskflow in context: a screenshot study of smartphone usage." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.

Kintsch, Walter and Teun A. van Dijk (1978). "Toward a model of text comprehension and production." In: *Psychological review* 85.5.

Kiseleva, Julia, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, et al. (2022a). "Interactive Grounded Language Understanding in a Collaborative Environment: IGLU 2021." In: *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR.

Kiseleva, Julia, Alexey Skrynnik, Artem Zholus, Shrestha Mohanty, Negar Arabzadeh, Marc-Alexandre Côté, Mohammad Aliannejadi, Milagro Teruel, Ziming Li, Mikhail S. Burtsev, Maartje ter Hoeve, Zoya Volovikova, Aleksandr I. Panov, Yuxuan Sun, Kavya Srinet, Arthur Szlam, and Ahmed Hassan Awadallah (2022b). "IGLU 2022: Interactive Grounded Language Understanding in a Collaborative Environment at NeurIPS 2022." In: *CoRR* abs/2205.13771. arXiv: `2205.13771`.

Kitaev, Nikita and Dan Klein (July 2018). "Constituency Parsing with a Self-Attentive Encoder." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics.

Kočiský, Tomáš, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette (2018). "The NarrativeQA Reading Comprehension Challenge." In: *Transactions of the Association for Computational Linguistics* 6.

Koehn, Philipp (2020). *Neural machine translation*. Cambridge University Press.

Koehn, Philipp and Christof Monz (June 2006). "Manual and Automatic Evaluation of Machine Translation between European Languages." In: *Proceedings on the Workshop on Statistical Machine Translation*. New York City: Association for Computational Linguistics.

Kool, Wouter, Herke van Hoof, and Max Welling (2019). "Stochastic Beams and Where To Find Them: The Gumbel-Top-k Trick for Sampling Sequences Without Replacement." In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Vol. 97. Proceedings of Machine Learning Research. PMLR.

Koupaee, Mahnaz and William Yang Wang (2018). "WikiHow: A Large Scale Text Summarization Dataset." In: *CoRR* abs/1810.09305. arXiv: 1810.09305.

Kratzwald, Bernhard, Anna Eigenmann, and Stefan Feuerriegel (July 2019). "RankQA: Neural Question Answering with Answer Re-Ranking." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics.

Kreutzer, Julia, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi (2022). "Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets." In: *Transactions of the Association for Computational Linguistics* 10.

Kumar, Sachin, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov (Aug. 2021). "Machine Translation into Low-resource Language Varieties." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics.

Kwiatkowski, Tom, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov (2019). "Natural Questions: A Benchmark for Question Answering Research." In: *Transactions of the Association for Computational Linguistics* 7.

Ladhak, Faisal, Esin Durmus, Claire Cardie, and Kathleen McKeown (Nov. 2020). "WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics.

Lake, Brenden M. and Marco Baroni (2018). "Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks." In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Vol. 80. Proceedings of Machine Learning Research. PMLR.

Lazaridou, Angeliki, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Sebastian Ruder, Dani Yogatama, Kris Cao, Tomás Kociský, Susannah Young, and Phil Blunsom (2021). "Pitfalls of Static Language Modelling." In: *CoRR* abs/2102.01951. arXiv: `2102.01951`.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (July 2020). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Lewis, Patrick, Ludovic Denoyer, and Sebastian Riedel (July 2019). "Unsupervised Question Answering by Cloze Translation." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics.

Li, Manling, Tengfei Ma, Mo Yu, Lingfei Wu, Tian Gao, Heng Ji, and Kathleen McKeown (Nov. 2021). "Timeline Summarization based on Event Graph Compression via Time-Aware Optimal Transport." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Li, Shen, João Graça, and Ben Taskar (July 2012). "Wiki-ly Supervised Part-of-Speech Tagging." In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics.

Li, Yuan (2019). "Probabilistic models for aggregating crowdsourced annotations." PhD thesis.

Lin, Chin-Yew (July 2004). "ROUGE: A Package for Automatic Evaluation of Summaries." In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics.

Litvak, Marina and Natalia Vanetik (2017). "Query-based summarization using MDL principle." In: *Proceedings of the Workshop on Summarization and Summary Evaluation Across Source Types and Genres, MultiLing@EACL 2017, Valencia, Spain, April 3, 2017*. Association for Computational Linguistics.

Liu, Junpeng, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang (Nov. 2021a). "Topic-Aware Contrastive Learning for Abstractive Dialogue Summarization." In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Liu, Linlin, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao (Aug. 2021b). "MulDA: A Multilingual Data Augmentation Framework for Low-Resource Cross-Lingual NER." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics.

Liu, Peter J., Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer (2018). "Generating Wikipedia by Summarizing Long Sequences." In: *6th*

International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Liu, Yang and Mirella Lapata (Nov. 2019). "Text Summarization with Pretrained Encoders." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach." In: *CoRR* abs/1907.11692. arXiv: 1907.11692.

Lo, Victor Ei-Wen and Paul A Green (2013). "Development and evaluation of automotive speech interfaces: useful information from the human factors and the related literature." In: *International Journal of Vehicular Technology* 2013.

Luan, Yi, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi (June 2019). "A general framework for information extraction using dynamic span graphs." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics.

Luccioni, Alexandra and Joseph Viviano (Aug. 2021). "What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics.

Luckin, Rosemary, Brett Bligh, Andrew Manches, Shaaron Ainsworth, Charles Crook, and Richard Noss (2012). *Decoding learning: The proof, promise and potential of digital education*. Nesta.

Ma, Shuming, Pengcheng Yang, Tianyu Liu, Peng Li, Jie Zhou, and Xu Sun (July 2019). "Key Fact as Pivot: A Two-Stage Model for Low Resource Table-to-Text Generation." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics.

Mani, Inderjeet (2001a). *Automatic summarization*. Vol. 3. John Benjamins Publishing.

Mani, Inderjeet (2001b). "Summarization Evaluation: An Overview." In: *Proceedings of the Third Second Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, NTCIR-2, Tokyo, Japan, March 7-9, 2001*. National Institute of Informatics (NII).

Markov, Todor, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng (2022). "A Holistic Approach to Undesired Content Detection in the Real World." In: *CoRR* abs/2208.03274. arXiv: 2208.03274.

Marneffe, Marie-Catherine de, Christopher D. Manning, Joakim Nivre, and Daniel Zeman (June 2021). "Universal Dependencies." In: *Computational Linguistics* 47.2.

Martelaro, Nikolas, Jaime Teevan, and Shamsi T Iqbal (2019). "An Exploration of Speech-Based Productivity Support in the Car." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

Matiisen, Tambet, Avital Oliver, Taco Cohen, and John Schulman (2020). "Teacher-Student Curriculum Learning." In: *IEEE Trans. Neural Networks Learn. Syst.* 31.9.

Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald (July 2020). "On Faithfulness and Factuality in Abstractive Summarization." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

McCloskey, Michael and Neal J Cohen (1989). "Catastrophic interference in connectionist networks: The sequential learning problem." In: *Psychology of learning and motivation*. Vol. 24. Elsevier.

Meng, Rui, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He (Aug. 2021). "Bringing Structure into Summaries: a Faceted Summarization Dataset for Long Scientific Documents." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics.

Microsoft (2019). *Voice report. From answers to action: customer adoption of voice technology and digital assistants*. `https://advertiseonbing-blob.azureedge.net/blob/bingads/media/insight/whitepapers/2019/04%20apr/voice-report/bingads_2019_voicereport.pdf`. Accessed: 2019-12-04.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013). "Distributed representations of words and phrases and their compositionality." In: *Advances in neural information processing systems* 26.

Mnih, Volodymyr, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu (2016). "Asynchronous Methods for Deep Reinforcement Learning." In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org.

Móro, Róbert and Mária Bieliková (2012). "Personalized Text Summarization Based on Important Terms Identification." In: *23rd International Workshop on Database and Expert Systems Applications, DEXA 2012, Vienna, Austria, September 3-7, 2012*. IEEE Computer Society.

Nadeau, David and Satoshi Sekine (2007). "A Survey of Named Entity Recognition and Classification." In: *Lingvisticae Investigationes* 30.1.

Nallapati, Ramesh, Feifei Zhai, and Bowen Zhou (2017). "SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents." In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. AAAI Press.

Narayan, Shashi, Shay B. Cohen, and Mirella Lapata (2018a). "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics.

Narayan, Shashi, Shay B. Cohen, and Mirella Lapata (June 2018b). "Ranking Sentences for Extractive Summarization with Reinforcement Learning." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics.

Nema, Preksha, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran (July 2017). "Diversity driven attention model for query-based abstractive summarization." In: *Proceedings of* *the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics.

Nenkova, Ani and Rebecca Passonneau (2004). "Evaluating Content Selection in Summarization: The Pyramid Method." In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Boston, Massachusetts, USA: Association for Computational Linguistics.

Nguyen, Tri, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng (2016). "MS MARCO: A Human-Generated MAchine Reading COmprehension Dataset." In.

Nikolaus, Mitja and Abdellah Fourtassi (Nov. 2021). "Modeling the Interaction Between Perception-Based and Production-Based Learning in Children's Early Acquisition of Semantic Knowledge." In: *Proceedings of the 25th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli (June 2019). "fairseq: A Fast, Extensible Toolkit for Sequence Modeling." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe (2022). "Training language models to follow instructions with human feedback." In: *CoRR* abs/2203.02155. arXiv: 2203.02155.

Papalampidi, Pinelopi, Frank Keller, Lea Frermann, and Mirella Lapata (July 2020). "Screenplay Summarization Using Latent Narrative Structure." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Park, Cheonbok, Yunwon Tae, TaeHee Kim, Soyoung Yang, Mohammad Azam Khan, Lucy Park, and Jaegul Choo (Aug. 2021). "Unsupervised Neural Machine Translation for Low-Resource Domains via Meta-Learning." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics.

Parton, Kristen, Kathleen R. McKeown, Bob Coyne, Mona T. Diab, Ralph Grishman, Dilek Hakkani-Tür, Mary Harper, Heng Ji, Wei Yun Ma, Adam Meyers, Sara Stolbach, Ang Sun, Gokhan Tur, Wei Xu, and Sibel Yaman (Aug. 2009). "Who, What, When, Where, Why? Comparing Multiple Approaches to the Cross-Lingual 5W Task." In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore: Association for Computational Linguistics.

Paulus, Romain, Caiming Xiong, and Richard Socher (2018). "A Deep Reinforced Model for Abstractive Summarization." In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Pires, Telmo, Eva Schlinger, and Dan Garrette (July 2019). “How Multilingual is Multilingual BERT?” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics.

Plank, Barbara and Željko Agić (2018). “Distant Supervision from Disparate Sources for Low-Resource Part-of-Speech Tagging.” In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics.

Plank, Barbara, Anders Søgaard, and Yoav Goldberg (Aug. 2016). “Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss.” In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics.

Platanios, Emmanouil Antonios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell (2018). “Contextual Parameter Generation for Universal Neural Machine Translation.” In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics.

Qian, Yujie, Enrico Santus, Zhijing Jin, Jiang Guo, and Regina Barzilay (June 2019). “GraphIE: A Graph-Based Framework for Information Extraction.” In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics.

Rae, Jack W., Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d’Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving (2021). “Scaling Language Models: Methods, Analysis & Insights from Training Gopher.” In: *CoRR* abs/2112.11446. arXiv: `2112.11446`.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” In: *J. Mach. Learn. Res.* 21.

Rajpurkar, Pranav, Robin Jia, and Percy Liang (July 2018). “Know What You Don’t Know: Unanswerable Questions for SQuAD.” In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (Nov. 2016). "SQuAD: 100,000+ Questions for Machine Comprehension of Text." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics.

Rao Vijjini, Anvesh, Kaveri Anuranjana, and Radhika Mamidi (Apr. 2021). "Analyzing Curriculum Learning for Sentiment Analysis along Task Difficulty, Pacing and Visualization Axes." In: *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Online: Association for Computational Linguistics.

Rapin, J. and O. Teytaud (2018). *Nevergrad - A gradient-free optimization platform*. `https://GitHub.com/FacebookResearch/Nevergrad`.

Ratner, Alexander, Stephen H. Bach, Henry R. Ehrenberg, Jason A. Fries, Sen Wu, and Christopher Ré (2020). "Snorkel: rapid training data creation with weak supervision." In: *VLDB J.* 29.2-3.

Reddy, Siva, Danqi Chen, and Christopher D. Manning (2019). "CoQA: A Conversational Question Answering Challenge." In: *Transactions of the Association for Computational Linguistics* 7.

Reder, Lynne M and John R Anderson (1980). "A comparison of texts and their summaries: Memorial consequences." In: *Journal of Verbal Learning and Verbal Behavior* 19.2.

Reichart, Roi, Katrin Tomanek, Udo Hahn, and Ari Rappoport (June 2008). "Multi-Task Active Learning for Linguistic Annotations." In: *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics.

Riccardi, Giuseppe, Frédéric Béchet, Morena Danieli, Benoît Favre, Robert J. Gaizauskas, Udo Kruschwitz, and Massimo Poesio (2015). "The SENSEI Project: Making Sense of Human Conversations." In: *Future and Emergent Trends in Language Technology - First International Workshop, FETLT 2015, Seville, Spain, November 19-20, 2015, Revised Selected Papers*. Vol. 9577. Lecture Notes in Computer Science. Springer.

Robertson, Stephen, Hugo Zaragoza, et al. (2009). "The probabilistic relevance framework: BM25 and beyond." In: *Foundations and Trends® in Information Retrieval* 3.4.

Rodriguez, Paul (2001). "Simple recurrent networks learn context-free and context-sensitive languages by counting." In: *Neural computation* 13.9.

Rodriguez, Paul, Janet Wiles, and Jeffrey L Elman (1999). "A recurrent neural network that learns to count." In: *Connection Science* 11.1.

Rodríguez Luna, Diana, Edoardo Maria Ponti, Dieuwke Hupkes, and Elia Bruni (Nov. 2020). "Internal and external pressures on language emergence: least effort, object constancy and frequency." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics.

Rush, Alexander M., Sumit Chopra, and Jason Weston (Sept. 2015). "A Neural Attention Model for Abstractive Sentence Summarization." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics.

Sandhaus, Evan (2008). "The New York Times annotated corpus." In: *Linguistic Data Consortium, Philadelphia* 6.12.

Sankar, Chinnadhurai and Sujith Ravi (2018). "Modeling non-goal oriented dialog with discrete attributes." In: *NeurIPS Workshop on Conversational AI:"Today's Practice and Tomorrow's Potential*.

Saxton, David, Edward Grefenstette, Felix Hill, and Pushmeet Kohli (2019). "Analysing Mathematical Reasoning Abilities of Neural Models." In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

See, Abigail, Peter J. Liu, and Christopher D. Manning (July 2017). "Get To The Point: Summarization with Pointer-Generator Networks." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics.

Shneidman, Laura A and Susan Goldin-Meadow (2012). "Language input and acquisition in a Mayan village: How important is directed speech?" In: *Developmental science* 15.5.

Siro, Clemencia, Mohammad Aliannejadi, and Maarten de Rijke (2022). "Understanding User Satisfaction with Task-oriented Dialogue Systems." In: *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*. ACM.

Spärck Jones, Karen (1998). "Automatic summarizing: factors and directions." In: *Advances in automatic text summarization*. 1. MIT press Cambridge, Mass, USA.

Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. (2022). "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models." In: *CoRR* abs/2206.04615. arXiv: 2206.04615.

Stein, Katharina, Leonie Harter, and Luisa Geiger (Aug. 2021). "SHAPELURN: An Interactive Language Learning Game with Logical Inference." In: *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*. Online: Association for Computational Linguistics.

Tan, Jiwei, Xiaojun Wan, and Jianguo Xiao (July 2017). "Abstractive Document Summarization with a Graph-Based Attentional Neural Model." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics.

ter Hoeve, Maartje, David Grangier, and Natalie Schluter (2022a). "High-Resource Methodological Bias in Low-Resource Investigations." In: *CoRR* abs/2211.07534. arXiv: 2211.07534.

ter Hoeve, Maartje, Mathieu Heruer, Daan Odijk, Anne Schuth, and Maarten de Rijke (2017). "Do News Consumers Want Explanations for Personalized News Rankings." In: *FATREC Workshop on Responsible Recommendation Proceedings*.

ter Hoeve, Maartje, Evgeny Kharitonov, Dieuwke Hupkes, and Emmanuel Dupoux (2021). "Towards Interactive Language Modeling." In: *ACL, Workshop on Semiparametric Methods in NLP* abs/2112.11911. arXiv: 2112.11911.

ter Hoeve, Maartje, Evgeny Kharitonov, Dieuwke Hupkes, and Emmanuel Dupoux (2022b). "Towards Interactive Language Modeling." In: *NeurIPS, Second Workshop on Interactive Learning for Natural Language Processing*.

ter Hoeve, Maartje, Julia Kiseleva, and Maarten de Rijke (2022c). "Summarization with Graphical Elements." In: *CoRR* abs/2204.07551. arXiv: `2204.07551`.

ter Hoeve, Maartje, Julia Kiseleva, and Maarten de Rijke (July 2022d). "What Makes a Good and Useful Summary? Incorporating Users in Automatic Summarization Research." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics.

ter Hoeve, Maartje, Robert Sim, Elnaz Nouri, Adam Fourney, Maarten de Rijke, and Ryen W. White (2020). "Conversations with Documents: An Exploration of Document-Centered Assistance." In: *CHIIR '20: Conference on Human Information Interaction and Retrieval, Vancouver, BC, Canada, March 14-18, 2020*. ACM.

Trischler, Adam, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman (Aug. 2017). "NewsQA: A Machine Comprehension Dataset." In: *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Vancouver, Canada: Association for Computational Linguistics.

Tsai, Peter (2018). *Data snapshot: AI Chatbots and Intelligent Assistants in the Workplace*. `https://community.spiceworks.com/blog/2964-data-snapshot-ai-chatbots-and-intelligent-assistants-in-the-workplace`. Accessed: 2019-10-04. Spiceworks, Inc.

Vaithilingam, Priyan, Tianyi Zhang, and Elena L Glassman (2022). "Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models." In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*.

Vakkari, Pertti (2020). "The Usefulness of Search Results: A Systematization of Types and Predictors." In: *CHIIR '20: Conference on Human Information Interaction and Retrieval, Vancouver, BC, Canada, March 14-18, 2020*. ACM.

van der Maaten, Laurens and Geoffrey Hinton (2008). "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need." In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*.

Vieira, Tim (2014). *Gumbel-max trick and weighted reservoir sampling*.

Völske, Michael, Martin Potthast, Shahbaz Syed, and Benno Stein (Sept. 2017). "TL;DR: Mining Reddit to Learn Automatic Summarization." In: *Proceedings of the Workshop on New Frontiers in Summarization*. Copenhagen, Denmark: Association for Computational Linguistics.

Vtyurina, Alexandra, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen White (2019). "VERSE: Bridging Screen Readers and Voice Assistants for Enhanced Eyes-Free Web Search." In: *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility. ACM*.

Wadden, David, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi (Nov. 2019). "Entity, Relation, and Event Extraction with Contextualized Span Representations." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th Interna-*

tional Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics.

Wal, Oskar van der, Silvan de Boer, Elia Bruni, and Dieuwke Hupkes (Nov. 2020). "The Grammar of Emergent Languages." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics.

Wang, Alex, Kyunghyun Cho, and Mike Lewis (July 2020). "Asking and Answering Questions to Evaluate the Factual Consistency of Summaries." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Wang, Lu, Hema Raghavan, Vittorio Castelli, Radu Florian, and Claire Cardie (2016). "A Sentence Compression Based Framework to Query-Focused Multi-Document Summarization." In: *CoRR* abs/1606.07548. arXiv: `1606.07548`.

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman (2020). "BLiMP: The Benchmark of Linguistic Minimal Pairs for English." In: *Transactions of the Association for Computational Linguistics* 8.

Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William S. Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel (2022). "Taxonomy of Risks posed by Language Models." In: *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM.

Wiles, Janet and Jeff Elman (1995). "Learning to count without a counter: A case study of dynamics and activation landscapes in recurrent networks." In: *Proceedings of the seventeenth annual conference of the cognitive science society*. s 482. Erlbaum Hillsdale, NJ.

Williams, Alex C, Harmanpreet Kaur, Shamsi Iqbal, Ryen W White, Jaime Teevan, and Adam Fourney (2019). "Mercury: Empowering Programmers' Mobile Work Practices with Microproductivity." In: *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology. ACM Press, New Orleans, Louisana, USA. https://doi. org/10.1145/3332165.3347932*.

Williams, Ronald J. (1992). "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning." In: *Mach. Learn.* 8.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics.

Wu, Zeqiu, Rik Koncel-Kedziorski, Mari Ostendorf, and Hannaneh Hajishirzi (2020). "Extracting Summary Knowledge Graphs from Long Documents." In: *CoRR* abs/2009.09162. arXiv: `2009.09162`.

Xie, Yuexiang, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding (2021). "Factual Consistency Evaluation for Text Summarization via Counterfactual Estimation." In: *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*. Association for Computational Linguistics.

Xu, Benfeng, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang (July 2020a). "Curriculum Learning for Natural Language Understanding." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Xu, Jiacheng, Zhe Gan, Yu Cheng, and Jingjing Liu (July 2020b). "Discourse-Aware Neural Extractive Text Summarization." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Xu, Weijia, Yuwei Yin, Shuming Ma, Dongdong Zhang, and Haoyang Huang (Nov. 2021). "Improving Multilingual Neural Machine Translation with Auxiliary Source Languages." In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Xu, Yang, Yu Hong, Huibin Ruan, Jianmin Yao, Min Zhang, and Guodong Zhou (2018). "Using active learning to expand training data for implicit discourse relation recognition." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (June 2021). "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics.

Yan, Rui and Dongyan Zhao (2018). "Coupled context modeling for deep chit-chat: towards conversations between human and computer." In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM.

Yang, Yi, Wen-tau Yih, and Christopher Meek (Sept. 2015). "WikiQA: A Challenge Dataset for Open-Domain Question Answering." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics.

Yang, Zhilin, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning (2018). "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics.

Yasunaga, Michihiro, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev (Aug. 2017). "Graph-based Neural Multi-Document Summarization." In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Vancouver, Canada: Association for Computational Linguistics.

Yu, Yi, Adam Jatowt, Antoine Doucet, Kazunari Sugiyama, and Masatoshi Yoshikawa (Aug. 2021). "Multi-TimeLine Summarization (MTLS): Improving Timeline Summarization by Generating Multiple Summaries." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics.

Zhang, Boliang, Ajay Nagesh, and Kevin Knight (July 2020a). "Parallel Corpus Filtering via Pre-trained Language Models." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Zhang, Meng, Liangyou Li, and Qun Liu (Aug. 2021a). "Two Parents, One Child: Dual Transfer for Low-Resource Neural Machine Translation." In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics.

Zhang, Shuo, Zhuyun Dai, Krisztian Balog, and Jamie Callan (2020b). "Summarizing and Exploring Tabular Data in Conversational Search." In: *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*. ACM.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2020c). "BERTScore: Evaluating Text Generation with BERT." In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhang, Wei, Wei Wei, Wen Wang, Lingling Jin, and Zheng Cao (2021b). "Reducing BERT Computation by Padding Removal and Curriculum Learning." In: *2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE.

Zhang, Wei Vivian, Xiaofei He, Benjamin Rey, and Rosie Jones (2007). "Query rewriting using active learning for sponsored search." In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM.

Zhu, Yi, Benjamin Heinzerling, Ivan Vulić, Michael Strube, Roi Reichart, and Anna Korhonen (Nov. 2019). "On the Importance of Subword Information for Morphological Tasks in Truly Low-Resource Languages." In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics.

SUMMARY

In this thesis we take a human-centered approach to natural language processing (NLP). That is, throughout this thesis, we center the design and development of natural language technology around humans. We are motivated from two angles, roughly summarized as: (i) who are the users of NLP systems, and what are their needs?, and (ii) how can we use our knowledge of human language processing and acquisition in designing and developing these systems?

We show that a human-centered approach to NLP helps us understand model behavior and capabilities, identify where and how modeling can be improved, and make sure models are in line with users' needs. As we proceed, we find that there are still many opportunities to more adequately model this human-centered perspective. Hence, this thesis is the start of a variety of new research directions in human-centered language technology — we propose new tasks, data, and (evaluation) methodologies.

In Chapter 2 we investigate the needs of users of a digital assistant in a document consumption scenario. By means of a survey we explore the space of questions that users ask in this scenario, and the answers that they expect. Next, we collect a human-labeled dataset that we use to train baseline models for the task. We find that these models perform well on the task, but that there is also still a lot of room for improvement.

In Chapter 3 we move from the space of digital assistance and question-answering to automatic text summarization. We are motivated by the observation that the users of automatically generated summaries are often ignored in earlier work on automatic summarization. By means of a survey amongst users of pre-made summaries, i.e., summaries that are pre-written by someone else, we show that current research on automatic summarization is not always in line with users' needs. Our survey can be reused to investigate other user groups with minor modifications. Finally, we contribute an evaluation methodology to investigate the usefulness of a generated summary.

Amongst others, participants in our survey indicated a need for summaries that contain a variety of graphical elements, such as arrows and colored text. In Chapter 4 we follow up on this request as we propose a new task: *summarization with graphical elements*. In designing this task, we are also motivated by how humans process written texts. We show that a critical mass of people finds our proposed summaries useful. We then continue with a data collection step, in which we collect a human-labeled test set. In the final step of this work we implement baseline methods that show that our task is feasible, yet challenging.

So far we have focused on English as the language of our modeling efforts, which is limited from a user-centered perspective. In Chapter 5 we expand our focus to a multitude of languages. Specifically, we are motivated by the observation that work on low-resource languages is often biased by methodologies used for high-resource scenarios. For example, a prominent approach to simulate a low-resource scenario is by randomly downsampling from a high-resource dataset. In this chapter we empirically show that this approach introduces bias in the context of part-of-speech tagging and machine translation, which leads to a biased view of how well these systems perform in real low-resource scenarios.

Finally, we leave the user-centered approach somewhat behind us, and seek to build on insights about human language acquisition. Specifically, we focus on artificial language modeling. State-of-the-art language models perform increasingly well. However, their training regime is unnatural from the perspective of human language acquisition, which is much more interactive. Motivated by this observation, we explore a more interactive approach to language modeling in Chapter 6. We propose a road map towards interactive language modeling and take the first steps on this road map that show the initial feasibility of our approach.

SAMENVATTING

In dit proefschrift hanteren we een mensgerichte benadering van natuurlijke taalverwerking. Dat wil zeggen dat we de mens centraal stellen bij het ontwerpen en het ontwikkelen van natuurlijke taaltechnologie. We zijn hierbij gemotiveerd vanuit twee invalshoeken, grofweg samengevat als: (i) wie zijn de gebruikers van systemen voor natuurlijke taalverwerking en wat zijn hun behoeften?, en (ii) hoe kunnen we onze kennis van menselijke taalverwerking en taalverwerving gebruiken bij het ontwerpen en ontwikkelen van deze systemen?

We tonen aan dat een mensgerichte benadering van natuurlijke taalverwerking ons helpt om het gedrag en de mogelijkheden van modellen voor natuurlijke taalverwerking te begrijpen, om te identificeren waar en hoe modellering kan worden verbeterd, en om ervoor te zorgen dat de modellen in overeenstemming zijn met de behoeften van gebruikers. We ontdekken dat er nog steeds veel mogelijkheden zijn om dit mensgerichte perspectief beter te modelleren. Dit proefschrift is dan ook het begin van een scala aan nieuwe onderzoeksrichtingen in mensgerichte taaltechnologie — we introduceren nieuwe taken, data en (evaluatie)methodologieën.

In Hoofdstuk 2 onderzoeken we de behoeften van gebruikers van een digitale assistent in een documentconsumptiescenario. Door middel van een survey verkennen we de vragen die gebruikers stellen in dit scenario, en de antwoorden die ze verwachten. Vervolgens verzamelen we een door mensen gelabelde dataset die we gebruiken om baselinemodellen voor de taak te trainen. We vinden dat deze modellen goed presteren op de taak, maar dat er ook nog veel ruimte voor verbetering is.

In Hoofdstuk 3 focussen we op het automatisch samenvatten van tekst. We zijn gemotiveerd vanuit de observatie dat de gebruikers van automatisch gegenereerde samenvattingen vaak worden genegeerd in eerder onderzoek naar automatische samenvattingen. Door middel van een survey onder gebruikers van vooraf gemaakte samenvattingen, d.w.z., samenvattingen die vooraf door iemand anders zijn geschreven, tonen we aan dat huidig onderzoek naar automatisch samenvatten niet altijd aansluit bij de behoeften van gebruikers. Onze survey kan met kleine aanpassingen worden hergebruikt om andere gebruikersgroepen te onderzoeken. Tot slot dragen we een evaluatiemethodiek bij om het nut van een gegenereerde samenvatting te onderzoeken.

Deelnemers aan ons onderzoek gaven onder meer aan behoefte te hebben aan samenvattingen die verschillende grafische elementen bevatten, zoals pijlen en gekleurde tekst. In Hoofdstuk 4 bouwen we hierop voort en stellen we een nieuwe taak voor: *samenvatten met grafische elementen*. Bij het ontwerpen van deze taak zijn we ook gemotiveerd door hoe mensen geschreven teksten verwerken. We laten zien dat een kritische hoeveelheid aan mensen onze voorgestelde samenvattingen nuttig vindt. In de volgende stap verzamelen we een door mensen gelabelde testset. In de laatste stap van dit werk implementeren we baselinemethodes die laten zien dat onze taak haalbaar maar uitdagend is.

Tot nu toe hebben we ons gericht op Engels als taal in onze modelleringsinspanningen. Vanuit een gebruikersgericht perspectief is dit beperkt. In Hoofdstuk 5 breiden we onze focus uit naar een groter aantal talen. We zijn met name gemotiveerd door de observatie dat het werken aan talen met weinig middelen vaak vertekend is door de methodologieën die gebruikt worden voor scenario's met veel middelen. Een prominente benadering voor het simuleren van een scenario met weinig middelen is bijvoorbeeld het willekeurig downsamplen van een grote dataset. In dit hoofdstuk laten we empirisch zien dat deze benadering bias introduceert in de context van *part-of-speech tagging* en machinaal vertalen, wat leidt tot een vertekend beeld van hoe goed deze systemen echt presteren in een scenario met weinig middelen.

Ten slotte laten we de gebruikersgerichte benadering enigszins achter ons en proberen we voort te bouwen op inzichten over menselijke taalverwerving. We richten ons specifiek op kunstmatige taalmodellering. Moderne taalmodellen presteren steeds beter. Echter, hun trainingsregime is onnatuurlijk vanuit het perspectief van menselijke taalverwerving, dat veel interactiever is. Gemotiveerd door deze observatie onderzoeken we een meer interactieve benadering van taalmodellering in Hoofdstuk 6. We stellen een roadmap voor naar interactieve taalmodellering en zetten de eerste stappen op deze roadmap die de initiële haalbaarheid van onze aanpak aantonen.