WN-Salience: A Corpus of News Articles with Entity Salience Annotations

Chuan Wu[1,2], **Evangelos Kanoulas**[2,3], **Maarten de Rijke**[2], **Wei Lu**[1]
[1]School of Information Management, Wuhan University, Wuhan, China
[2]Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[3]Amsterdam Business School, University of Amsterdam, Amsterdam, The Netherlands
wuchuan114@gmail.com, e.kanoulas@uva.nl, derijke@uva.nl, weilu@whu.edu.cn

Abstract

Entities can be found in various text genres, ranging from tweets and web pages to user queries submitted to web search engines. Existing research either considers all entities in the text equally important, or heuristics are used to measure their salience. We believe that a key reason for the relatively limited work on entity salience is the lack of appropriate datasets. To support research on entity salience, we present a new dataset, the WikiNews Salience dataset (WN-Salience), which can be used to benchmark tasks such as entity salience detection and salient entity linking. WN-Salience is built on top of Wikinews, a Wikimedia project whose mission is to present reliable news articles. Entities in Wikinews articles are identified by the authors of the articles and are linked to Wikinews categories when they are salient or to Wikipedia pages otherwise. The dataset is built automatically, and consists of approximately 7,000 news articles, and 90,000 in-text entity annotations. We compare the WN-Salience dataset against existing datasets on the task and analyze their differences. Furthermore, we conduct experiments on entity salience detection; the results demonstrate that WN-Salience is a challenging testbed that is complementary to existing ones.

Keywords: entity salience, salience detection, Wikinews

1. Introduction

Text modeling has traditionally made no distinction between different terms in the text. Examples include bag of words representation, language models, and term weighting methods. Research on knowledge extraction and text semantics has shifted some of the attention towards utterances that represent real world entities, while recent work on entity linking (Shen et al., 2015) has made it possible to take entities into consideration in various downstream applications, such as information retrieval (Dalton et al., 2014; Raviv et al., 2016).

Various corpora annotated with entities have been built for entity related research, such as FACC1 (Gabrilovich et al., 2013). However, these corpora make no distinction between salient and non-salient entities, despite the fact that only few entities are central to a document. For instance, in the Web domain, fewer than 5% of the entities on a web page are salient to the page (Gamon et al., 2013). Many existing publications have recognized the importance of understanding entity salience (Fetahu et al., 2015; Tran et al., 2015; Xiong et al., 2018; Ponza et al., 2019; Wu et al., 2019). For example, automatically suggesting news pages for populating Wikipedia requires determining whether a news article should be referenced by an entity, considering several aspects of the article, including entity salience, relative authority, and novelty of the article (Fetahu et al., 2015). In general, there is growing interest in understanding entity salience, demonstrated by research on entity salience detection (Gamon et al., 2013; Dunietz and Gillick, 2014). Therefore, it is very important to be able to quantify the salience of an entity.

To facilitate research involving entity salience, datasets with both entity annotations and salience labels are necessary. Ideally, one would like to have human annotators labeling salient entities in documents. Unfortunately, this is not scalable due to the high volume of documents that need to be annotated and the cost of human labor. At the same time, with the rise of deep learning algorithms datasets should consist of tens of thousands of annotations to allow learning.

A small number of datasets (Gamon et al., 2013; Dunietz and Gillick, 2014) have been developed, to facilitate research on entity salience. However, existing datasets suffer from several limitations: (1) computational errors in entity annotations, (2) strong assumptions in collecting entity salience labels, and (3) noise in entity salience labeling. For example, in the NYT-salience dataset (Gamon et al., 2013), entities in documents are identified by applying an NP extractor, a co-reference resolver, and an entity resolver, which might propagate mistakes to the final annotations. Gamon et al. (2013) assume a soft labeling approach: if users click on a web page link after they issue an entity query, the entity is likely to be salient in the web page. It is also believed that heuristic design is a difficult proposition (Gamon et al., 2013).

To address the aforementioned limitations, we propose a method to extract a new dataset by collecting news articles from Wikinews,[1] and build a new dataset referred to as *WN-Salience*. Wikinews is a free-content news source wiki, where anyone can write news articles. In each article, text fragments referring to entities are linked by the article authors to Wikipedia pages corresponding to the respective entity or Wikinews categories. Though Wikinews itself is multi-lingual, without loss of generality, we focus on English language news articles only, given the popularity and the number of articles in the language. We believe that our method can be applied to other languages as well.

Our method is based on the following observation. Authors are highly advised to link news articles to Wikinews categories, to allow effective information organization in Wikinews, and do so only when a category is strongly related to the written article. Therefore, the categories can be

[1]https://en.wikinews.org/wiki/Main_Page

viewed as salience annotations and entities corresponding to these categories as salient entities.

To illustrate the utility of the developed WN-Salience dataset, we conduct experiments on entity salience detection. By applying simple algorithms, we confirm the effectiveness of positional features in entity salience detection found in (Dunietz and Gillick, 2014), but also demonstrate the inferiority of other hand crafted features found discriminative in the literature, which shows that this dataset is challenging and likely orthogonal in some aspects to existing datasets. The dataset is available on GitHub.[2] We follow the license policy of Wikinews and publish the dataset under a free license.[3]

The main contributions of this work are summarized as follows:

1. We propose a method for extracting human-annotated entity salience labels using Wikinews categories and in-text entity annotations.
2. We develop a new dataset for research around entity salience.
3. We analyze our dataset and compare it with previous datasets.
4. We conduct experiments to demonstrate the utility of the dataset.

2. Related Work

2.1. Notion of salience

A recent definition of entity salience is given in (Gamon et al., 2013). Gamon et al. (2013) first declare that a thing that has a Wikipedia page associated with it to be an entity and then present a notion of entity salience using two assumptions, i.e., local scoping and invariable perception. Local scoping indicates that the salience of an entity in a document can be solely determined by the document itself, while invariable perception means that entity salience can be assessed independently from the interests of readers, and independently from the prior importance of the entity as it exists outside of the document. Another notion of entity salience is more empirical: salient entities are those that human readers deem most relevant to the document (Dunietz and Gillick, 2014).

Even though they are reasonable, the two assumptions above are not easy to handle in practice. In this work, we adopt an assumption similar to the empirical definition of entity salience: salient entities are those that authors of articles deem most relevant to the document. Given an article, there might be hundreds or thousands readers, while there can only be one or few writers. Instead of considering salience from the perspective of readers, we adopt the opinion of writers. Two advantages of the assumption are the following: first, the potential inconsistency between different readers is avoided; and second, it is easier to capture authors' opinion on salience than that of readers, which makes it more convenient to collect explicit salience labels.

[2]https://github.com/researchdatasets/wn-salience-dataset

[3]https://creativecommons.org/licenses/by/2.5/

2.2. Existing datasets

Gamon et al. (2013) propose to identify salient entities in web pages by using a soft labeling approach based on behavioral signals from web users as a proxy for salience. The assumption is that when a user issues an entity query and clicks on an URL on the returned results page, the entity is salient in the corresponding web page. For pages that receive enough traffic, reliable user click statistics can be obtained and used to derive entity salience labels. As a result, a dataset called Microsoft Document Aboutness (MDA), was constructed. A major limitation of the dataset is that it is not publicly available. Furthermore, it is also hard to reproduce a similar dataset without access to large scale web search log data. Another limitation of the approach is that the assumption relies on the behavior of web users, which is known t come with bias, e.g., position bias (Craswell et al., 2008) or domain bias (Ieong et al., 2012).

The New York Times salience (NYT-Salience) benchmark collection introduced by Dunietz and Gillick (2014) is built on top of the New York Times corpus (Sandhaus, 2008). To build the NYT-Salience corpus, two steps were taken, recognizing entities and assigning salience labels. Given a document and its abstract, a standard NLP pipeline was first run to identify entities both in the abstract and in the text of the news article; then, entities in the abstract were aligned with entities in the document. Entities in the document that also appear in the abstract are considered salient. Two limitations lie in NYT-Salience. First, entities are identified by a multi-step NLP pipeline, which might lead to errors in entity annotations. Second, the dataset is only partially available. The NYT-Salience dataset does not provide the underlying textual content along with the annotations due to copyright restrictions.

The Reuters-128 Salience dataset is a corpus built on top of Reuters-128 (Röder et al., 2014), an English corpus built for evaluating NER systems, which contains 128 news articles in economy. The entity salience labels are obtained by crowdsourcing (Dojchinovski et al., 2016). The key limitation of the dataset is its small size, which does not allow for the development of supervised learning algorithms. In addition, the entity annotation process used might suffer from errors introduced by entity linking tools. Finally, entities in the dataset are uniquely identified by Wikipedia titles, DBpedia urls and others. Ideally, it is expected that all entities come from the same knowledge base. If entities are identified by entities in different knowledge bases, then many additional processing steps are needed whenever it is necessary to refer to information in knowledge bases.

The Wikinews dataset (Trani et al., 2018) is constructed for salient entity linking, which combines the task of entity linking and entity salience detection. Since Wikinews is a collection of news articles with entity annotations, the creators created entity salience labels and used them for salient entity linking. The entity salience labels are collected using a crowdsourcing platform. The dataset creators define entity salience using a 4-grade metric, i.e., *top relevant*, *highly relevant*, *partially relevant* and *not relevant*. To deal with subjectivity in the assignment of salience scores, the salience scores from multiple annotators are averaged. Though also extracted from Wikinews, this dataset is differ-

Table 1: Comparison of existing datasets on entity salience.

Dataset	Entity Annotations	Salience Labels	Size of Corpus
MDA dataset	proprietary NER pipeline	soft labeling	∼50,000
nyt-salience	proprietary NLP pipeline	heuristic rules	100,976
Reuters-128	human labeling	crowdsourcing	128
Wikinews	human labeling	crowdsourcing	604
WN-Salience	human labeling	automatic deriviation	∼7,000

ent from our dataset. First, Trani et al. (2018) use graded scores to measure salience. Second, we exploit the category information to induct entity salience labels automatically, while they rely on annotators from crowdsourcing platform.

2.3. Summary

Here, we summarize existing datasets involving entity salience and present the comparison in Table 1. In terms of entity annotation, manual entity annotation is preferred over entities tagged by entity recognition pipelines. For salience labels, human annotated salience labels are considered to be more reliable. However, human annotated salience labels rely on crowdsourcing, which is usually very expensive. Therefore, it is preferred to derive salience labels using automated methods.

As we can see, existing datasets suffer from either less preferred entity annotations (MDA dataset and nyt-salience) or the limitation of expensive salience label collection method (Reuters-128 and Wikinews). By making use of entity annotations in Wikinews articles and categories assigned to articles by writers, our dataset is able to use human annotated entity annotations and collect salience labels using automated methods. In this way, we avoid either limitation. As for the size of corpus, our dataset is of moderate size compared to existing datasets.

3. Wikinews and Annotations

Wikinews is a Wikipedia project, the mission of which is to present reliable, unbiased and relevant news.[4] News articles in Wikinews are written by volunteers, who can write or edit a page by expanding it, correcting facts and so on. There are various types of article in Wikinews, such as original reporting,[5] interviews,[6] daily summaries[7] and so on. For example, interview articles usually start with background descriptions of interviews, followed by conversations between interviewers and the interviewees.

In this work, we mainly focus on two types of article in Wikinews, i.e., synthesis articles and original reporting. Synthesis articles are written by collecting media reports from many other sources (always fully cited), synthesizing them into a single article. Bias is stripped out and a neutral point of view is presented. Original reporting articles are first-hand news reports written by Wikinews contributors on-the-spot of news events.[8] The reason we only focus on these two types of article is that they are usually the most typical and popular types, that are also frequently observed on the web.

A typical Wikinews article consists of a title, body content with **in-text annotations**, related news, sources, and **Wikinews categories**. In the rest of the work, we will use the example Wikinews article, entitled "*Koreas hold joint training session for Olympics*."[9] Among all the elements of a Wikinews article, Wikinews categories and in-text annotations within the body content are the important ones for constructing our dataset; they are introduced below.

3.1. Wikinews categories

In Wikinews, every article needs to be listed under one or more categories, so that articles under a particular category can be easily found. The process of selecting appropriate categories is guided by the following principle provided by Wikinews: "*Typically, both a "location" category (where did the news event take place?) and a "topic" category (what is the event about?) is required.*"[10] For example, an article about a computer science conference in Brussels might have the following categories: *Computer Science*, *Brussels*, and *Belgium*. Such a set of categories can be seen at the bottom of every Wikinews article.

3.2. In-text annotations

Wikinews encourages authors to add wikilinks when textual fragments (i.e., entity mentions) are referring to entries in other Wiki site, such as categories in Wikinews, and article pages in Wikipedia. These wikilinks are considered in-text annotations.

Wikinews articles typically contain two types of in-text annotation, Wikinews category annotations, and Wikipedia entity annotations, as shown in Figure 1. Wikinews category annotations are links to Wikinews categories. For example, in the example article, the entity mention *Kim Jong-un* is representing an entity and has corresponding Wikinews category *Kim Jong-un*.[11] As a result, a wikilink is added to refer to the Wikinews category *Kim Jong-un*. Wikipedia entity annotations are links to Wikipedia entities. For example, the text fragment *National Assembly*

[4] https://en.wikinews.org/wiki/Main_Page

[5] https://en.wikinews.org/wiki/Wikinews:Original_reporting

[6] https://en.wikinews.org/wiki/Category:Interview

[7] https://en.wikinews.org/wiki/Category:Wikinews_Shorts

[8] https://en.wikinews.org/wiki/Wikinews:Introduction

[9] https://en.wikinews.org/wiki/Koreas_hold_joint_training_session_for_Olympics?dpl_id=2833718

[10] https://en.wikinews.org/wiki/Wikinews:Writing_an_article

[11] https://en.wikinews.org/wiki/Category:Kim_Jong-un

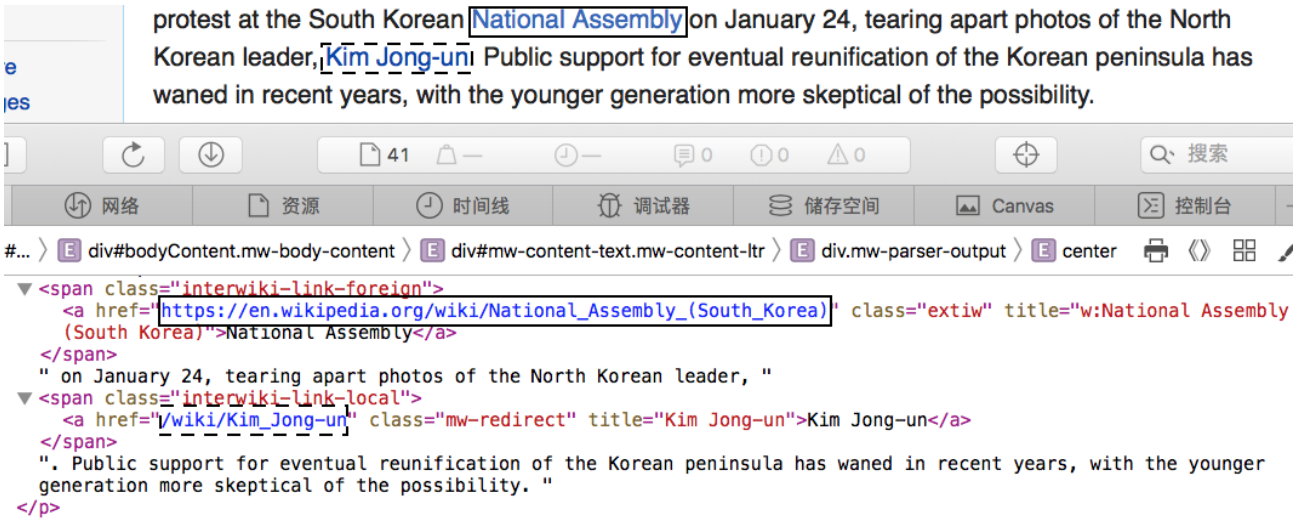

Figure 1: Examples of Wikinews category annotation (dash line box) and Wikipedia entity annotation (solid line box).

in the example article can be linked to the corresponding Wikipedia page *National_Assembly_(South_Korea)*.[12]
We observe that even though many Wikinews categories correspond to Wikipedia entities, authors annotate entity mentions by Wikinews categories first, and by Wikipedia pages only when Wikinews categories are not available.

4. Entity Salience Hypothesis

In this section, we present our entity salience hypothesis, which is used to induce salience labels in our datasets. Based on how Wikinews categories are annotated and how Wikinews category pages are organized, we propose the following hypothesis: *an entity is salient if the Wikinews category that corresponds to the entity is also labeled as a category of the article*. In contrast, if an entity in an article is labeled as a category that is not included in the set of the article categories, or if it is labeled as a Wikipedia page, it is not salient in the article.
To illustrate the above hypothesis, we examine the example article mentioned in Section 3.. In the example article, categories such as *North Korea*, *South Korea*, *Olympics*, *Ice Hockey*, *Kim Jong-un*, and *Moon Jae-in* are labeled as categories by the author of the article. Based on the main content of this article, we can observe that the two countries and the two presidents represent the "*main characters*" of the story presented, while *Olympics* and *Ice Hockey* serve as the topic explaining the reason why the characters connect with each other in this article. And it is clear that the category entities labeled here are salient entities in the article.
On the other hand, we can see that category entities that are not annotated as a category of the article are not salient entities. For example, categories such as *Seoul* are not labeled as a category of the example article. *Seoul* appears when the article mentions the historical fact that the 1988 Summer Olympics happens in Seoul, and this fact is not related to the main story of article. Therefore, it is not a salient entity of the article, and is not labeled as a category of the article.
Note that some categories of articles might not appear in the body content of articles. Since our focus is the salience of entities in documents, we do not consider entities that do not appear in documents, even though they might be helpful for document understanding. We preserve all categories of articles in our dataset, including the categories that are simple dates.

5. The WN Salience Dataset

In this section, we first describe the dataset extraction process, including the categories collection and the articles collection process. Then, we show some basic statistics of the dataset, and analyze entity salience within and across documents.

5.1. Dataset collection

We collect raw web pages from Wikinews, and parse them using jsoup.[13] Given the elements in Wikinews articles, we extract the following fields: *title*, *date*, *body content*, *categories*. Note that we keep the paragraph structure of articles to facilitate possible scenarios where paragraph information is needed. For each paragraph, we extract the main text and the annotations. The information in each annotation includes mention text, the corresponding entity (Wikipedia title or Wikinews category), position in the paragraph (begin offset and end offset).
On the basis of our aforementioned entity salience hypothesis, we include in each annotation a binary entity salience label (1 for salient entities, 0 otherwise). Since our focus is to extract a dataset for entity salience related tasks, we focus on articles that have at least one salient entity. The collection process consists of two steps, i.e., collecting categories and collecting articles under selected categories.

Collecting categories. In Wikinews, categories are organized in a hierarchy, where each category belongs to at least one parent category. The root category of the Wikinews category hierarchy is *Internal Wikinews organization*, which

[12] https://en.wikipedia.org/wiki/National_Assembly_(South_Korea)
[13] https://jsoup.org

Table 2: Statistics of WN Salience. The numbers on the lower part are document-wise. Document length and paragraph length are counted in terms of words.

	Train set	Test set
# of articles	5928	1040
Avg. doc length	335	679
Avg. paragraph length	50	78
Avg. # of paragraphs	6.7	8.7
Avg. # of unique entities	12.5	14.2
Avg. # of annotations	13.0	19.2
Avg. # of categories	11.9	15.0

belongs to itself. If we start from *Internal Wikinews organization*, and iterate over subcategories of each category, we are able to iterate over all categories.

Instead of iterating over all categories and parsing all articles, we consider a category as a target category if it satisfy the following criterion: the Wikinews category has a corresponding Wikipedia page. The reason for this is that we want to have a unified representation of salient entities, the Wikipedia entity unique identifier. Imagine an extreme case, where the only salient entity is a Wikinews category and the Wikinews category does not have corresponding Wikipedia entity. Then the salient entity would be just a unique identifier and does not have connection to any knowledge base. This is undesirable because: (1) there is no guarantee that all Wikinews categories are entities; (2) in existing datasets involving entity salience, all (salient) entities are knowledge base entities, either Freebase entities or Wikipedia entities; and (3) it would prevent research that involves entity salience and knowledge bases.

Note that there are also categories that are irrelevant to our purpose. For example, news articles whose titles start with *Wikinews interviews* are very different documents compared to ordinary news report. Other examples include *Wikinews Shorts*, *Original reporting*, *Translated news*, *Photo essays*, *Published*, *Archived* and so on. These categories are not meaningful categories in terms of representing some real world entity. Instead, they are either for the purpose of website organization (e.g., Published and Archived), or for the purpose of guiding the writing of authors (e.g., Photo essays). However, no filtering is needed for these categories because they usually do not have corresponding Wikipedia pages. In the end, 4,214 categories are found, out of which 1,813 categories have corresponding Wikipedia pages.

Collecting articles. We iterate over the collected category pages and obtain the articles within each category. We iterate over all articles in all categories and obtain 11,005 articles. Then we select articles that have at least one salient entity, which means that at least one category of an article is an entity that appears in article body. In the end, we obtain 6,968 articles, which constitute the WN-Salience dataset.

5.2. Dataset statistics

To facilitate supervised methods, we divide all articles into a training set and a test set. Temporal splitting is an intuitive

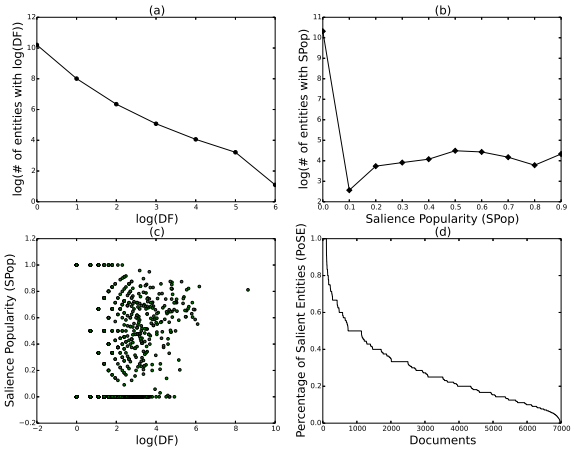

Figure 2: Analysis of the WN-Salience dataset.

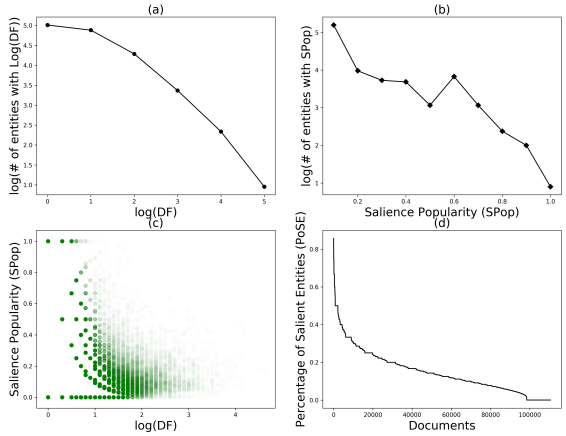

Figure 3: Analysis results of NYT-Salience dataset.

way to construct a training and a test set. In previous work, temporal splitting by year was used. However, we observe that basic statistics show major differences between news articles in different years. Therefore, we choose to split the dataset on a monthly basis, i.e., all articles up to a threshold month are placed in the training set, while the remaining articles are placed in the test set. We set the threshold month to September. Basic statistics of our dataset are shown in Table 2.

5.3. Dataset analysis

In order to have an intuitive understanding of the statistics of our dataset (WN-Salience), we perform an analysis of how document frequency and salience popularity of entities are distributed. For the purpose of comparison, we also present a similar analysis results of the NYT Salience dataset.

Entity document frequency (DF). We present the distribution of log document frequency of entities in Figure 2 (a) and Figure 3 (a). Since the document frequency of entities varies a lot from high frequency entities to low frequency entities (power law distribution), we focus on the scale of the document frequency of entities. Specifically, we put entities whose log document frequency under the same scale

into the same group, and present the log of the number of entities in each group. As shown in the results, the statistics of WN-Salience is similar to that of NYT Salience.

Entity salience popularity (SPop). The salience popularity of entity e is defined as SDF_e/DF_e, where SDF_e is the number of documents where e is salient and $DF(e)$ is the document frequency of e. We count the log number of entities whose salience popularity range from $[sp, sp+0.1]$, where $sp \in [0, 0.9]$. The results are shown in Figure 2 (b) and Figure 3 (b). In both datasets, the SPop of many entities are zero, which indicates that entity salience is skewed towards few entities. More entities in NYT Salience dataset shows moderate salience percentage (0.3 to 0.7) compared to that of WN-Salience. This indicates that it might be more difficult to identify salient entities in WN-Salience compared to NYT Salience.

DF vs. SPop. To see how document frequency and salience percentage of entities correlate with each other in our dataset, we represent each entity as a two dimensional point in a figure, where the two dimension are its DF and SPop. The results are shown in Figure 2 (c) and Figure 3 (c). Entities tend to be evenly distributed in WN-Salience and skewed towards bottom-left in NYT Salience. This shows that with the increase of document frequency, the SPop of entities in NYT Salience is very likely to decrease, while that in WN-Salience can still remain high.

Percentage of salient entities (PoSE) of documents. The percentage of salient entities in document d is defined as S_d/E_d, where S_d is the number of salient entities in d, while E_d is the total number of entities in d. We count PoSE in each entity and rank them in descent order. The results are shown in Figure 2 (d) and Figure 3 (d). As we can see, the PoSE of most entities are lower than 5%, which conforms with the observation in (Gamon et al., 2013) that fewer than 5% entities on a web page are salient to the web page.

6. Experiments

6.1. Research questions

We address the following research questions:

RQ1 How consistent is salience annotation between our dataset and the Wikinews dataset proposed by Trani et al. (2018)?

RQ2 Does the small number of existing Wikinews categories affect the quality of salience labels?

RQ3 How do baseline methods on entity salience detection perform on our dataset?

6.2. Comparative analysis between datasets

An existing dataset with salience labels proposed by Trani et al. (2018), referred to as the *SELWikinews* dataset, has also been extracted from Wikinews. Given the same origin, we are able to perform a comparative analysis between SELWikinews and our dataset. To make the comparison possible, article matching and entity alignment are necessary. In particular, we first identify a common set of articles by title matching, i.e., only articles with the exactly same title are selected. Then, we match the entities across the datasets. Entities in SELWikinews dataset are represented as entity id in Wikipedia, while in our dataset, entities are represented by their Wikipedia title. We process the 2018.07.20 Wikipedia dump to extract the mapping from entity id to its Wikipedia title, so that we can match entities between the two datasets.

Table 3: The results of comparing salience annotations in WN-Salience dataset against that in SELWikinews. Each row presents the results under different threshold of salience score.

	WN-Salience			
Threshold	P	R	F1	Acc.
3.0	0.0433	0.8750	0.0825	0.4166
2.0	0.4031	0.8556	0.5480	0.5971
1.0	0.9784	0.6079	0.7499	0.6046

After extracting a common set of articles and making entities comparable, we perform salience label matching to validate annotation consistency. The salience score in SELWikinews ranges from 0.0 to 3.0, while in our datasets, we have binary salience labels, indicating whether an entity is salient or not. We propose to use simple rules to flatten the salience scores in SELWikinews to binary labels: if the salience score of an entity is above a predefined threshold value, the entity is salient and it is not salient otherwise. Then, we use the salience labels derived from SELWikinews as ground truth, and those in our datasets as predictions. We choose binary evaluation metrics over the salience labels, including precision, recall, F1, and accuracy in our experiments. We use three threshold values, i.e., 1.0, 2.0 and 3.0, to see the results for different levels of saliency.

The article title matching identified 243 articles that exist in both datasets. The results for different thresholds are shown in Table 3. Since the individual salience score given by annotators in SELWikinews range from 0.0 to 3.0, and the final score is the average score of multiple annotators, we consider 2.0 as a reasonable threshold for the flattening process. The results for the other two thresholds are given for comparison.

As we can see, our dataset has a reasonable accuracy, which is around 0.6. The high recall and moderate precision indicate the fact that we are more aggressive at assigning salience labels to entities. This can be either due to the fact that (1) human annotators in SELWikinews are more cautious in annotating salient entities (low precision), or (2) article writers tend to annotate more salient entities (high recall). Therefore, we consider our dataset as complementary to existing datasets given its different method of salience annotation.

6.3. Risk of missing salient entities

Table 4 provides statistics about the in-text annotation of entity mentions. We define as category entities, those entity mentions that have both a corresponding Wikinews category and a Wikipedia page. As one can observe from this table, when an entity mention is a category entity, then

Table 4: In-text annotation statistics. CE stands for category entity.

# of Wikipedia entity annotation	49,556
# of Wikinews category annotation	19,534
# of CE as Wikipedia entity annotation	2,002
# of CE as Wikinews category annotation	15,968
# of other annotations	3,086

the chance of the writer annotating it as a Wikinews category is about 89%, while the chance of annotating it as a Wikipedia page is 11%. This is rather important, given that only entities annotated as Wikinews categories can be considered for salience. What is worrying, however, is that if we consider all annotations, 70% of those are Wikipedia page annotations. This means that there is a large number of entity mentions for which there is no corresponding Wikinews category, and hence they are annotated as Wikipedia pages. This also means that these 49,556 entity mentions will never be considered for salience.

Since not all entities have corresponding Wikinews categories, there might be a risk of missing salient entities. We refer to this risk as *low recall risk* (LRR), since it might lead to lower recall than it should be. We investigate this issue by measuring the impact of LRR. In particular, we extract subsets with decreasing LRR and present ESD results of the subsets. In principle, if all entities are category entities, LRR does not exist, since all entities will be considered for salience. The higher the ratio of category entity in articles is, the lower LRR is. To measure LRR, we define the ratio of category entity in an article as follows:

$$ce\text{-}ratio = \frac{N_{ce}}{N_{ce} + N_{nc} + N_{pe}}$$

where $N_{i,ce}$, $N_{i,nc}$ and $N_{i,pe}$ represents the number of category entity annotations, WN category annotations and WP entity annotations in the i-th article. Note that WN category annotations represent categories that does not have a corresponding Wikipedia page. We extract subsets of WN Salience by specifying $ce\text{-}ratio$ ranging from 0.5 to 0.8 and compare against SELWikinews dataset.

After extracting WN-Salience subsets under different $ce\text{-}ratio$, we compare each subset against SELWikinews as was done in Section 6.2.. The results are shown in Table 5. As we can see, the value of all metrics of these subsets are quite close and there is no clear winner between subsets under different $ce\text{-}ratio$. Therefore, we assume that for the LRR risk can be neglected for our dataset.

6.4. Application: Entity salience detection

Since the focus of this work is to introduce a new dataset for tasks involving entity salience, we run simple algorithms to showcase the use of our dataset. We choose to evaluate on the task of entity salience detection over WN-Salience.

We follow the work of Dunietz and Gillick (2014). In particular, we use some hand-crafted features to train a binary classifier to identify whether an entity is salient in a document. Because of the difference between our dataset and their dataset (NYT-Salience), we do not follow all their implementations. We use Naive Bayes as our classifier.

Table 5: WN-Salience subsets with different ce-ratio, and their comparison against SELWikinews. The results of WN Salience is using threshold 2.0 to convert graded scores in SELWikinews to binary labels.

ce-ratio	**# of docs**	**P**	**R**	**F1**
0.5	111	0.4021	0.8519	0.5463
0.6	69	0.3974	0.8564	0.5429
0.7	39	0.3808	0.8505	0.5260
0.8	19	0.4091	0.8333	0.5488
WN Salience	243	0.4031	0.8556	0.5480

We consider three kind of features, i.e., positional features, count features and entity centrality features. Positional features are investigated here because they achieve reasonable performance. Since count of head word is actually ambiguous, we use entity frequency in articles as count features. Following Dunietz and Gillick (2014), we also applying the function $f(x) = \text{round}(\log(k(x+1)))$ to count features, and k is set to 10 in our experiments.

We use precision, recall, and f1 on salient entities as our evaluation metrics. In all experiments, a classification threshold of 0.5 is used by default, since in each case it is close to threshold that maxmized F1.

Table 6 shows experimental results on two datasets on the task of entity salience detection. As we can see in the results, positional features achieve reasonable performance, which conforms with the results in (Dalton et al., 2014). Adding the first location of an entity does not help much. The reason is that they are both positional features and thus indicate similar information.

To our surprise, features that are used to approximate entity frequency, i.e., head counts and mentions, have a negative impact on the performance. As also observed by Dalton et al. (2014), the precision decreases on both datasets compared to the positional baseline. However, the recall shows different trends (increasing on NYT Salience, decreasing on WN-Salience). This might come from the fact that documents in WN-Salience are not very long and entities might not appear in documents many times, which makes entity frequency less meaningful as a feature.

The effectiveness of using entity centrality feature is not as good as expected. Comparing the performance in two datasets, it works better in NYT Salience. The recall decreases a lot after using the centrality feature in WN-Salience, which means that the salience of entities is less sensitive to centrality rank, compared to NYT Salience.

7. Conclusions and Future Work

In this work, we uncover entity salience information in Wikinews website. Based on our observations, we propose an automated method to extract datasets with entity salience annotations, which leverages the category annotations in Wikinews news articles. Our extracted dataset, WN-Salience is presented. Experiments are performed to validate our proposed assumptions, measure the consistency between our dataset and an existing dataset and set a benchmark for evaluating on the task of entity salience

Table 6: The results of entity salience detection over two datasets.

	NYT-Salience			WN-Salience		
Features	P	R	F1	P	R	F1
positional baseline	0.5598	0.4095	0.4730	0.4794	0.5322	0.5044
head count	0.3346	0.5221	0.4078	0.2422	0.2138	0.2271
mentions	0.4198	0.4167	0.4182	0.2422	0.2138	0.2271
1st-loc	0.1901	0.4133	0.2604	0.2908	0.7890	0.4250
+ head count	0.3206	0.7079	0.4413	0.2643	0.8124	0.3988
+ mentions	0.3919	0.5970	0.4732	0.2920	0.4806	0.3633
+ centrality	0.3506	0.6554	0.4568	0.2921	0.4850	0.3646

detection. We believe that WN-Salience will stimulate the development of more advanced method for entity salience detection and salient entity linking. Here we focus on English language only. Our method for extracting a similar dataset in other languages is possible.

8. Acknowledgements

This research was supported by Ahold Delhaize, the Association of Universities in the Netherlands (VSNU), the China Scholarship Council, the European Union, under project number H2020-EU.3.4. - SOCIETAL CHALLENGES - Smart, Green And Integrated Transport (814961), the Google Faculty Research Awards program, the Innovation Center for Artificial Intelligence (ICAI), the National Natural Science Foundation of China, under project numbers 71473183 and 61876003, and the Netherlands Organisation for Scientific Research (NWO), under project numbers 016.Vidi.189.039 and 314-99-301.
All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

9. Bibliographical References

Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. (2008). An experimental comparison of click position-bias models. In *WSDM*, pages 87–94. ACM.

Dalton, J., Dietz, L., and Allan, J. (2014). Entity query feature expansion using knowledge base links. In *SIGIR*, pages 365–374. ACM.

Dojchinovski, M., Reddy, D., Kliegr, T., Vitvar, T., and Sack, H. (2016). Crowdsourced corpus with entity salience annotations. In *LREC*, Paris, France. European Language Resources Association (ELRA).

Dunietz, J. and Gillick, D. (2014). A new entity salience task with millions of training examples. In *EACL*, volume 14, page 205.

Fetahu, B., Markert, K., and Anand, A. (2015). Automated news suggestions for populating Wikipedia entity pages. In *CIKM*, pages 323–332. ACM.

Gabrilovich, E., Ringgaard, M., and Subramanya, A. (2013). Facc1: Freebase annotation of clueweb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0). http://lemurproject.org/clueweb09/FACC1/.

Gamon, M., Yano, T., Song, X., Apacible, J., and Pantel, P. (2013). Identifying salient entities in web pages. In *CIKM*, pages 2375–2380. ACM.

Ieong, S., Mishra, N., Sadikov, E., and Zhang, L. (2012). Domain bias in web search. In *WSDM*, pages 413–422. ACM.

Ponza, M., Ferragina, P., and Piccinno, F. (2019). Swat: A system for detecting salient Wikipedia entities in texts. *Computational Intelligence*, 35(4):858–890.

Raviv, H., Kurland, O., and Carmel, D. (2016). Document retrieval using entity-based language models. In *SIGIR*, pages 65–74. ACM.

Röder, M., Usbeck, R., Hellmann, S., Gerber, D., and Both, A. (2014). N^3-a collection of datasets for named entity recognition and disambiguation in the NLP interchange format. In *LREC*, pages 3529–3533.

Sandhaus, E. (2008). The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Shen, W., Wang, J., and Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *TKDE*, 27(2):443–460.

Tran, T. A., Niederée, C., Kanhabua, N., Gadiraju, U., and Anand, A. (2015). Balancing novelty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. In *CIKM*, pages 1201–1210. ACM.

Trani, S., Lucchese, C., Perego, R., Losada, D. E., Ceccarelli, D., and Orlando, S. (2018). SEL: A unified algorithm for salient entity linking. *Computational Intelligence*.

Wu, C., Kanoulas, E., and de Rijke, M. (2019). It all starts with entities: A salient entity topic model. *Natural Language Engineering*, pages 1–19.

Xiong, C., Liu, Z., Callan, J., and Liu, T.-Y. (2018). Towards better text understanding and retrieval through kernel entity salience modeling. In *SIGIR*, pages 575–584. ACM.