Generating Pseudo-ground Truth for Predicting New Concepts in Social Streams

David Graus, Manos Tsagkias, Lars Buitinck, and Maarten de Rijke

ISLA, University of Amsterdam, Amsterdam, The Netherlands
{d.p.graus, e.tsagkias, l.j.buitinck, derijke}@uva.nl

Abstract. The manual curation of knowledge bases is a bottleneck in fast paced domains where new concepts constantly emerge. Identification of nascent concepts is important for improving early entity linking, content interpretation, and recommendation of new content in real-time applications. We present an unsupervised method for generating pseudo-ground truth for training a named entity recognizer to specifically identify entities that will become concepts in a knowledge base in the setting of social streams. We show that our method is able to deal with missing labels, justifying the use of pseudo-ground truth generation in this task. Finally, we show how our method significantly outperforms a lexical-matching baseline, by leveraging strategies for sampling pseudo-ground truth based on entity confidence scores and textual quality of input documents.

1 Introduction

We increasingly harvest the power of knowledge bases to interpret the content generated around us. This is achieved via semantic linking, a process that identifies mentions of real-life entities or concepts in text, and links them to concepts in a knowledge base (KB) [20]. Core to its success is the extensive coverage of today's KBs; they span the majority of popular and well established concepts. For most content domains this coverage is enough; however it does not provide a solid basis for domains that refer to "long-tail" entities or where new entities are constantly born: e.g. in news and social streams. Here, entities emerge (and sometimes disappear) before editors of a KB reach consensus on whether an entity should be included. Identifying newly emerging entities that will become concepts in a knowledge base is important in knowledge base population and complex filtering tasks, where users are not just interested in any entity, but also in attributes like impact or importance.

The target users we have in mind are media analysts in an online reputation management setting who track entities that can impact the reputation of their customer in social streams, e.g., Twitter. Our problem is related to named entity detection, classification and disambiguation, with the additional constraint that an entity should have "impact" or be important. Although impact or importance are hard to model because they depend on the context of a task, we argue that entities that are included in a knowledge base are more important than those that are not, and use this signal for modeling the importance of an entity.

Named entity recognition is a natural approach for identifying newly emerging entities, that are not in the KB. However, current models fall short as they do not account for the importance, or impact of the entity. In this paper, we present an unsupervised method for generating pseudo-ground truth for training a named entity recognizer to predict new concepts. Our method is applicable to any trainable model for named entity recognition. In addition, our method is not restricted to a particular class of entities, but can be trained to predict any type of concept that is in the KB.

The challenge here is two-fold: (a) how to model the attribute of importance, and (b) the system needs to adapt to its input (social streams) and the updates in the knowledge base; content that is eligible for addition in Wikipedia today, may no longer be in the future, which renders static training annotations unusable.

Our approach answers both challenges. For the first challenge, we carefully craft the training set of a named entity recognizer to steer it towards identifying new concepts. That is, we leverage prior knowledge of important concepts, to identify new concepts that are likely to share the same attributes. Just as a named entity recognizer trained solely on English person-type entities will recognize only such entities, a named entity recognizer trained on entities with referent KB concepts is likely to recognize only this type of entity. For the second challenge, we provide an unsupervised method for generating pseudo-ground truth from the input stream. This way, we are not dependent on human annotations that are necessarily limited and domain and language specific, and newly added knowledge will be automatically included.

We focus on social streams because of the fast paced evolution of content and its unedited nature, which make it a challenging setting for predicting which entities will feature in a knowledge base. The main research question we seek to answer is: *What is the utility of our sampling methods for generating pseudo-ground truth for a named entity recognizer?* We measure utility within the task of predicting new concepts from social streams as the prediction effectiveness of a named entity recognizer trained using our method. We also study the impact of prior knowledge, in our second research question: *What is the impact of the size of prior knowledge on predicting new concepts?* Our main contribution is a method that uses entity linking for generating training material for a named entity recognizer.

2 Approach

We view the task of identifying new concepts in social streams as a combination of an *entity linking* (EL) problem and a *named-entity recognition and classification* (NERC) problem. We visualize our method in Fig. 1. Starting from a document in a document stream, we extract sentences, and use an EL system to identify referent concepts in each sentence. If any is identified, the sentence is pooled as a candidate training example for NERC (we refer to this type of sentence as a *linkable sentence*), otherwise it is routed to NERC for identifying new concepts (*unlinkable sentences*): an underlying assumption behind our method is that the first place to look for new entities is the set of unlinkable sentences. Most of our attention in this paper is devoted to training NERC. Two ideas are important here. First, we extend the distributional hypothesis [11] (i.e., words that occur in the same contexts tend to have similar meaning) from words to entities and concepts; we hypothesize that new entities that should be included in the knowledge base occur in similar contexts as current knowledge base concepts. Second, we apply EL on the input stream and transform its output into pseudo-ground truth for NERC; this

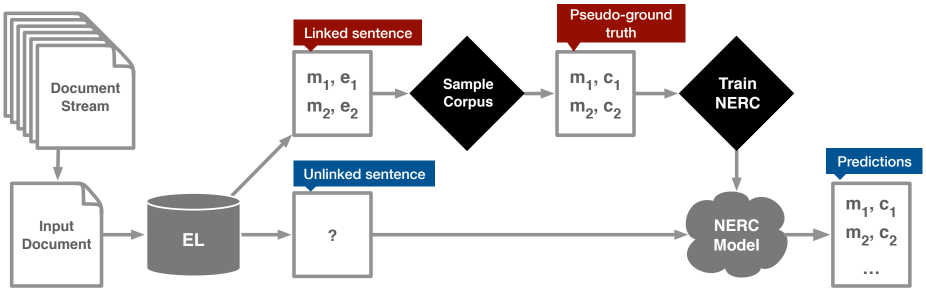

Fig. 1. Our approach for *generating pseudo-ground truth* for training a NERC method, and for *predicting new concepts*.

results in an unsupervised way of generating pseudo-ground truth, with the flexibility of choosing any type of entity or concept described in the KB.

3 Unsupervised generation of pseudo-ground truth

We start with the output of an entity linking method[1] [19, 23], on a sentence of a document in the stream. This output is our source for generating training material. The output consists of tuples of *entity mentions* and *referent entities* (m, e pairs in Fig. 1).

Since we are allowed to use generic corpora from any domain, e.g., news, microblog posts, we may expect to have noise in our pseudo-ground truth. We apply various sampling methods to select sentences that make up a high quality training corpus. These sampling methods are described in Section 4.

After sampling, we convert the remaining sentences in a format suitable for input for NERC. This format consists of the *entity span*; the sequence of tokens that refers to an entity, i.e. the entity mention, and *entity class* for each linked entity (m, c pairs in Fig. 1). To denote the entity span, we apply the BIO-tagging scheme [24], where each token is tagged with whether it is the **B**eginning of an entity, **I**nside an entity, or **O**utside one, so that a document like *Kendrick Lamar and A$AP Rocky. That's when I started listening again. Thanks to Brendan.* becomes: *Kendrick*$_B$ *Lamar*$_I$ *and*$_O$ *A$AP*$_B$ *Rocky*$_I$. *That's*$_O$ *when*$_O$ *I*$_O$ *started*$_O$ *listening*$_O$ *again*$_O$. *Thanks*$_O$ *to*$_O$ *Brendan*$_B$. The final step is to assign a class label to an entity. For this, we need to move from a concept to a concept class. As not all knowledge bases associate a concept class to their concepts, we use DBPedia for looking up the concept and extracting the concept's DBPedia ontology class, if any; see Section 5 for details. Our example then becomes: *Kendrick*$_{B\text{-}PER}$ *Lamar*$_{I\text{-}PER}$ *and*$_O$ *A$AP*$_{B\text{-}PER}$ *Rocky*$_{I\text{-}PER}$. *That's*$_O$ *when*$_O$ *I*$_O$ *started*$_O$ *listening*$_O$ *again*$_O$. *Thanks*$_O$ *to*$_O$ *Brendan*$_{B\text{-}PER}$. Now we can proceed and train NERC with our generated pseudo-ground truth. We do so using a two-stage approach [4] where the recognition stage is implemented using the fast structured perceptron algorithm [7].[2]

[1] http://semanticize.uva.nl

[2] https://github.com/larsmans/seqlearn

Table 1. Nine features used for sampling documents from which we train a NERC system.

Feature	Description	Feature	Description
n_mentions	Number of usernames (@)	avg_token_len	Average token length
n_hashtags	Number of hashtags (#)	tweet_len	Length of tweet
n_urls	Number of URLs	density	Density as in [13]
ratio_upper	Percentage of uppercased chars	personal	Contains personal pronouns (I, me, we, etc.)
ratio_nonalpha	Percentage of non-alphanumeric chars		

4 Sampling pseudo-ground truth

In this section, we present two methods for sampling pseudo-ground truth: (a) sampling based on the EL system's confidence score for a detected entity, and (b) sampling based on the textual quality of an input document.

Sampling based on entity linker's confidence score. Typically, entity linkers return a confidence score with each entity mention (n-gram) they are able to link to a knowledge base concept. These confidence scores can be used to rank possible concepts for an n-gram, but also for pruning out entity mention-concept pairs about which the linker is not confident. Although the scale of the confidence score is dependent on the model behind the entity linker, the scores can be normalized over the candidate concepts for an entity, e.g., using linear or z-score normalization. We use the *SENSEPROB* [23] metric as confidence score. This score is calculated by combining the probability of an n-gram being used as an anchor text on Wikipedia, with the *commonness* [21] score.

Sampling based on textual quality of an input document. Taking the textual quality of content into account has proved helpful in a range of tasks. Based on [18, 26], we consider nine features indicative of textual quality; see Table 1. While not exhaustive, our feature set is primarily aimed at social streams as our target document stream (see §5) and suffices for providing evidence on whether this type of sampling is helpful for our purposes. Based on these features, we compute a final score for each document d as $\text{score}(d) = \frac{1}{|F|}\sum_{f \in F}\frac{f(d)}{\max_f}$, where F is our set of feature functions (Table 1) and $\max_f$ is the maximum value of f we have seen so far in the stream of documents. Since all features are normalized in $[0,1]$, $\text{score}(d)$ has this same range. As a sanity check, we rank documents from the MSM2013 [1] dataset using our quality sampling method and list the top-3 and bottom-3 scoring documents in Table 2. Top scoring documents are longer and denser in information than low scoring documents. We assume that these documents are better examples for training a NERC system.

In the next section, we follow a linear search approach to sampling training examples as input for NERC. First, we find an optimal threshold for confidence scores, and fix it. For sampling based on textual quality, we turn to the MSM2013 dataset to determine sampling thresholds. We calculate the scores for each tweet, and scale them to fall between [0,1]. We then plot the distribution of scores, and bin this distribution in three parts: tweets that fall within a single standard deviation of the mean are considered *normal*, tweets to the left of this bin are considered *noisy*, whilst the remaining tweets to the right of the distribution are considered *nice*. We repeat this process for our tweet corpus, using the bin thresholds gleaned from the MSM2013 set.

Table 2. Ranking of documents in the MSM2013 dataset based on our quality sampling method. Top ranking documents appear longer and denser in information than low ranking documents.

Top-3 quality documents
“Watching the History channel, Hitler’s Family. Hitler hid his true family heritage, while others had to measure up to Aryan purity.”
“When you sense yourself becoming negative, stop and consider what it would mean to apply that negative energy in the opposite direction.”
“So. After school tomorrow, french revision class. Tuesday, Drama rehearsal and then at 8, cricket training. Wednesday, Drama. Thursday ... (c)”
Bottom-3 quality documents
*Toni Braxton ˜ He Wasnt Man Enough for Me _HASHTAG_ _HASHTAG_? _URL_ RT _Mention_*
“tell me what u think The GetMore Girls, Part One _URL_”
this girl better not go off on me rt

5 Experimental setup

In addressing the new concept prediction in document streams problem, we concentrate on developing an unsupervised method for generating pseudo-ground truth for NERC and predicting new concepts. In particular, we want to know the effectiveness of our unsupervised pseudo-ground truth (UPGT) method over a random baseline and a lexical matching baseline, and the impact on effectiveness of our two sampling methods. To answer these questions, we conduct both optimization and prediction experiments.

Dataset. As a document stream, we use tweets from the TREC 2011 Microblog dataset [16], a collection of 4,832,838 unique English Twitter posts. This choice is motivated by the unedited and noisy nature of tweets, which can be challenging for prediction. Our knowledge base (KB) is a subset of Wikipedia from January 4, 2012, restricted to concepts that correspond to the NERC classes person (PER), location (LOC), or organization (ORG). We use DBPedia to perform selection, mapping the DBPedia classes *Organisation*, *Company*, and *Non-ProfitOrganisation* to ORG, *Place*, *PopulatedPlace*, *City*, and *Country* to LOC, and *Person* to PER.[3] Our final KB consists of 1,530,501 concepts.

Experiment I: Sampling pseudo-ground truth. To study the utility of our sampling methods, we turn to the impact of setting a threshold on the entity linker’s confidence score (Experiment Ia.) and the effectiveness of our textual quality sampling (Experiment Ib.). In Experiment Ia., we do a sweep over thresholds from 0.1 up to 0.9, using the same threshold for both the generation of pseudo-ground truth and evaluating the prediction effectiveness of new concepts. Lower thresholds allow low confidence entities in the pseudo-ground truth, and likely generates more data at the expense of noisy output. We emphasize that we are not interested in the correlation between noise and confidence score, but rather in the performance of finding new concepts given the EL

[3] http://mappings.dbpedia.org/server/ontology/classes/

system's configuration. In Experiment Ib., we compare our methods performance with differently sampled pseudo-ground truths, containing nice, normal, or noisy tweets.

Experiment II: Prediction experiments. To answer our second research question, and study the impact of prior knowledge on detecting new concepts, we compare the performance of our method (UPGT) to two baselines: a *random baseline* (RB) that extracts all *n*-grams from test tweets and considers them new concepts, and a *lexical-matching baseline* (NB) that follows our approach, but generates pseudo-ground truth by applying lexical matching of KB entity titles, instead of an EL system, and refrains from sampling based on textual quality. For this experiment, we use the optimal sampling parameters for generating pseudo-ground truth from our previous experiment, i.e., include linked entities with a confidence score higher than 0.7, and use only normal tweets. As we will show in Section 6, the threshold of 0.7 balances performance with high recall of entities in the pseudo-ground truth.

Evaluation. We evaluate the quality of the generated pseudo-ground truth on the effectiveness of a NERC system trained to predict new concepts. As measuring the addition of new concepts to the knowledge base is non-trivial, we consider a retrospective scenario: Given our KB, we random sample concepts to yield a smaller KB ($\mathrm{KB_s}$). This $\mathrm{KB_s}$ simulates the available knowledge at the present point in time, whilst KB represents the future state. By measuring how many concepts we are able to detect in our corpus that feature in KB, but not $\mathrm{KB_s}$, we can measure new concept prediction. We create $\mathrm{KB_s}$ by taking random samples of 20–90% the size of KB (measured in concepts), in steps of 10%. We repeat each sampling step ten times to avoid bias.

We generate test sets and pseudo-ground per $\mathrm{KB_s}$. We link the corpus of tweets using $\mathrm{KB_s}$, and yield two sets of tweets: (a) tweets that contain new concepts, and (b) tweets with linked concepts, analog to the *unlinked* and *linked sentences* in Fig. 1. The size of these two sets depend on the size of $\mathrm{KB_s}$ and makes the comparison of results difficult across different $\mathrm{KB_s}$. We cater for this bias by randomly sampling 10,000 tweets from both the test set and the pseudo-ground truth and repeating our experiments ten times.[4] Ground truth is then assembled by linking the corpus of tweets using KB. This ground truth consists of 82,305 tweets, with 12,488 unique concepts.

We evaluate the effectiveness of our method in two ways: (a) the ability of NERC to generalize from our pseudo-ground truth, and (b) the accuracy of our predictions. For the first, we compare the predicted concept *mentions* to those in our ground truth, akin to traditional NERC evaluation. For the second we take the set of correct predictions (true positives), and link each mention to the referent concept in the ground truth. This allows us to measure what we're actually interested in: the fraction of newly discovered concepts. For both types of evaluation, we report on average precision and recall over 100 runs per $\mathrm{KB_s}$. Statistical significance is tested using a two-tailed paired t-test and is marked as $^\blacktriangle$ for significant differences for $\alpha = .01$.

6 Results

Experiment I: Sampling pseudo-ground truth. Our first experiment aims to answer RQ1: *What is the utility of our sampling methods for generating pseudo-ground truth*

[4] Using the smallest $\mathrm{KB_s}$ (20%) results in about 15,000 tweets in the pseudo-ground truth.

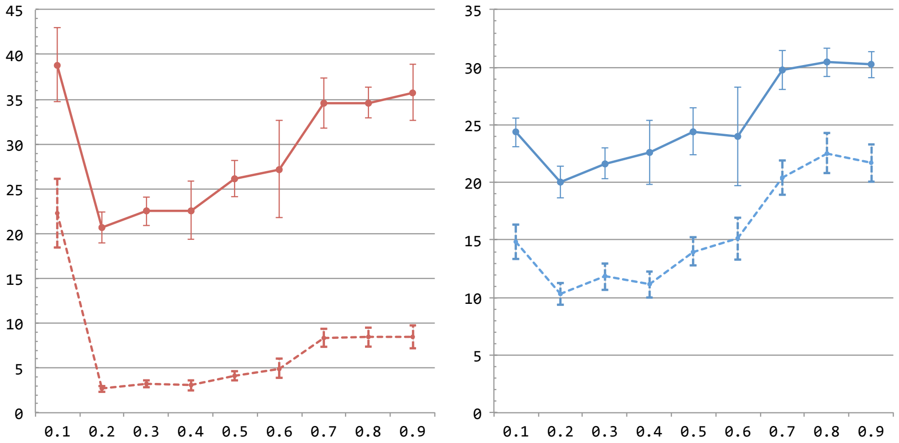

Fig. 2. Experiment Ia. Impact of confidence score on UPGT. Effectiveness of identifying mentions of new concepts (left). New concept prediction effectiveness (right). Threshold set on the confidence score in the x-axis. Precision (solid line) and recall (dotted) are shown in the y-axis.

for a named entity recognizer? We fix the $\mathrm{KB_s}$ at 50%. We start by looking at the ability of NERC to generalize from our pseudo-ground truth, measured on two aspects: (i) effectiveness for identifying mentions of new concepts, and (ii) predicting new concepts.

In Experiment Ia., we look at our confidence-based sampling method. For identifying mentions of new concepts, we find that effectiveness peaks at 0.1 confidence threshold with a precision of 38.84%, dips at 0.2 and slowly picks up to 35.75% as threshold increases (Fig. 2, left). For new concept prediction, effectiveness positively correlates with the threshold. Effectiveness peaks at the 0.8 confidence threshold, statistically significantly different from 0.7 but not from 0.9 (Fig. 2, right).

Interestingly, besides precision, recall also shows a positive correlation with thresholds. This suggests that in new concept prediction, missing training labels are likely to have less impact on performance than generating incorrect, noisy labels. This is an interesting finding as it sets the concept prediction task apart from traditional NERC, where low recall due to incomplete labeling is a well-understood challenge.

Next, we turn to the characteristics of the pseudo-ground truth that results for each of these thresholds, and provide an analysis of their potential impact on effectiveness. We find that more data through a larger pseudo-ground truth allows NERC to better generalize and predict a larger number of new concepts.

This claim is supported by the number of predicted concepts per threshold in Table 3. We find a similar trend as in the precision and recall graph above: the number of predicted concepts peaks for the threshold at 0.1 (6,653 concept mentions), and drops between 0.2 and 0.4, and picks up again from 0.5 reaching another local maximum at 0.8. The increasing number of predicted concept mentions with stricter thresholds indicates that the NERC model is more successful in learning patterns for separating concepts from noisy labels. This may be due to the entity linker linking only those entities it is most confident about, providing a clearer training signal for NERC.

Table 3. Number of predicted concept mentions per threshold on the confidence score.

Threshold	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Predictions	6,653	1,500	1,618	1,512	1,738	2,025	2,662	2,713	2,614
Ground truth	11,429	11,533	11,291	11,078	10,955	10,935	10,799	10,881	10,855

For the rest of our experiments we use a threshold of 0.7 on confidence score because it is deemed optimal in terms of trade-off between performance and quantity of concepts in pseudo-ground truth.

In Experiment Ib., we study three different textual quality-based sampling strategies: we consider only normal tweets (i), nice tweets (ii), or both normal and nice tweets (iii); see Section 4. For reference, we also report on the performance achieved when no textual quality sampling is used. We keep $\mathrm{KB_S}$ fixed at 50%, and use 0.7 for confidence threshold.

Table 4. Experiment Ib. Precision and recall for three sampling strategies based on textual quality of documents: nice, normal, normal+nice. We also report on effectiveness of a system that uses no sampling for reference. Boldface indicates best performance. Statistical significance is tested against the previous sampling method, e.g., nice to normal.

	Mention		**Concept**	
Sampling	**Precision**	**Recall**	**Precision**	**Recall**
No sampling	34.26±2.65	8.21±0.83	29.63±1.67	20.25±1.56
Normal+nice	45.50±4.71$^{\blacktriangle}$	12.97±2.03$^{\blacktriangle}$	36.22±2.07$^{\blacktriangle}$	24.86±1.69$^{\blacktriangle}$
Normal	66.09±3.86$^{\blacktriangle}$	30.94±3.46$^{\blacktriangle}$	44.62±1.51$^{\blacktriangle}$	**32.20±1.67**$^{\blacktriangle}$
Nice	**70.36±3.07**$^{\blacktriangle}$	**30.98±3.25**	**45.99±1.34**$^{\blacktriangle}$	29.69±1.79

Textual quality-based sampling turns out to be twice as effective as no sampling on both identifying mentions of new concepts and new concept prediction. Among our sampling strategies, nice proves to be the most effective with a precision of 70.36% for concept mention identification.

In new concept prediction, the performance of nice and normal strategies hovers around the same levels. In terms of recall, nice and normal methods are on par, outperforming both other strategies. The success of nice and normal sampling methods can be attributed to the fact that a more coherent and homogeneous training corpus allows the NERC model to more easily learn patterns.

Experiment II: Impact of prior knowledge Next, we seek to answer RQ2: *What is the impact of the size of prior knowledge on predicting new concepts?* We use the optimal combination of our sampling methods from the previous experiments, i.e., a confidence threshold of 0.7, and the normal textual quality sampling. We again look at the effectiveness of our methods in identifying new concept mentions and predicting new concepts.

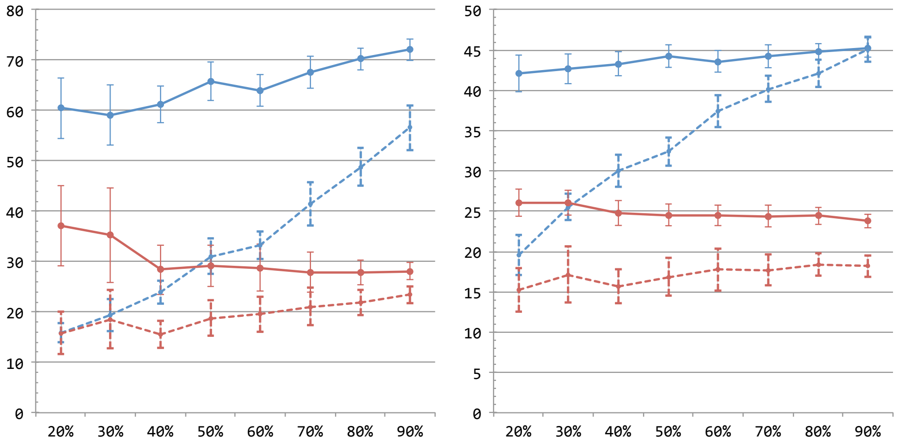

Fig. 3. Our method (UPGT, blue line) versus the lexical baseline (red) for both identifying mentions of new concepts (left) and new concept detection (right). Knowledge base size is on the *x*-axis, and precision (solid lines) and recall (dotted lines) are marked on the *y*-axis.

Fig. 3 shows the effectiveness of our methods as a function of the size of the knowledge base. For identifying mentions of new concepts, our method (UPGT, blue line) constantly and statistically significantly outperforms both lexical (red line) and random baselines (not shown). In terms of recall, the lexical baseline is on par with UPGT for KB sizes up to 30%. The random baseline shows very low precision for both identifying mentions of new concepts, and new concept prediction over all KB_s, but almost perfect recall—which is expected given that it assigns all possible *n*-grams as new concept mentions and new concepts (0.69% precision and 65% recall for entity mention identification, 1.82% precision and 94.95% recall for new concept prediction).

Next, we take a closer look at the results. The lexical baseline's recall increases slightly when more prior knowledge is added to the knowledge base. This is expected behavior because, as we saw in our previous experiment, the pseudo-ground gets more diverse labels, and helps the NERC to generalize. Looking at the number of unique concepts in the pseudo-ground truth, we find that the number increases with the size of KB_s. UPGT assigns labels to 2,500 unique concepts at 20% KB_s, which tops at 11,000 unique concepts at 90%. These numbers are lower for the lexical baseline (1,800 at 20% KB_s, and 7,000 at 90% KB_s). However, for both methods, the number of concepts in the ground truth stays around the same. The gradual improvement in precision and recall of UPGT for increasing KB_s can be attributed to a broader coverage for labeling (observed through looking at the prior entities), and the main distinction between the lexical baseline and UPGT: more strict labeling through leveraging the entity linker's confidence score.

Finally, to better understand the performance of our method, we have looked at the set of correct predictions (new concepts), and false positives, or incorrectly identified new concepts. Our analysis revealed examples of out of knowledge base entities, that were not included in the initial KB, highlighting the challenging setting of evaluating the task.

On the whole, however, our method is able to deal with missing labels and incomplete data, as observed through its consistent and stable precision, justifying our assumption that data is incomplete by design.

7 Related work

The problem we concentrate on, and the method we propose for approaching it relate to literature in (a) entity linking in document streams, (b) training methods on automatically annotated data, and (c) predicting new concepts.

Document streams. Lexical matching-based entity linking approaches have shown to be successful in the challenging genre of social streams [19], and have shown to be suitable for adaptation to new genres and languages [23]. They provide a strong baseline, and as an added advantage are independent of intricate NLP pipelines, linguistic features, etc. Cassidy et al. [6] expand on this approach by considering "coherence" between entities to aid disambiguation. Guo et al. [12] propose a weakly supervised method for detecting so-called NIL entities, but this cannot handle or recognize out-of-KB entities. In addition, the noisy character of social streams degrades the effectiveness of NER methods [9, 10], and current approaches largely tailor the NLP pipeline to Twitter and heavily rely on large amounts of labeled data [3, 25]. We generate large amounts of training data for these types of system to improve NER effectiveness on social streams.

Automatically generated pseudo-ground truth. Several attempts have been made to either simulate or generate human annotations. Kozareva [15] uses regular expressions for generating training data for NERC. Zhou et al. [28] generate training data by considering Wikipedia links as positive examples, and consider each other entity that may be referred to by the same anchor as negative examples Nothman et al. [22] leverage the anchor text of links inbetween two same articles in different Wikipedia translations for training a multilingual NERC system. Wu et al. [27] investigate generating training material for NERC from one domain and testing on another. Becker et al. [2] study the effects of sampling automatically generated data for training NER. Our setting differs from settings considered before, and our approach to automatically generating ground truth by using entity linking is new too.

Predicting new concepts. Viewed abstractly, our task is similar to named entity normalization in user generated content [14], and named entity disambiguation in streams [8], but the conditions are different because of the lack of "context" and discussion structure (e.g., comments on an article). Our task is also different because of our focus on knowledge bases and the emergence of unknown entities. Bunescu [5] study out-of-Wikipedia entity detection by setting a threshold on their candidate ranker. Lin et al. [17] leverage n-gram statistics from Google Books for predicting new concepts. Our method generates training data solely from the input stream and a knowledge base and does not depend on third sources which may have different evolution rate.

8 Conclusions

We tackled the problem of predicting new concepts in social streams. We presented an unsupervised method for generating pseudo-ground truth to train NERC for detecting entities that are likely to become concepts in a knowledge base. Our method uses the output of an entity linker to generate training material for NERC. We introduced two sampling methods, based on the entity linker confidence, and the textual quality of an input document. We found that sampling by textual quality improves performance of NERC and consequently our method's performance in new concept prediction. As setting a higher threshold on the entity linker's confidence score for generating pseudo-ground truth results in fewer labels but better performance, we show that the NERC is better able to separate noise from entities that are worth including in a knowledge base. The entity linker's confidence score is an effective signal for this separation. Our sampling methods significantly improve detection of knowledge base worthy entities.

In the case of a small amount of prior knowledge , i.e., size of the available KB, our method is able to cope with missing labels and incomplete data, as observed through its consistent and stable precision, justifying our proposed method that assumes incomplete data by design. This finding furthermore suggests the scenario of an increasing rate of new concept prediction, as more data is fed back to the KB. Additionally, we found that a larger number of entities in the KB allows for setting a desirable stricter threshold on the confidence scores, and leads to improvements in both precision and recall. This finding suggests an adaptive threshold that takes prior knowledge into account could prove effective.

Our proposed method can be applied with any trainable NERC model and entity linker that is able to return a confidence score for a linked entity. In addition, our method is suitable for domain and/or language adaptation as it does not rely on language specific features or sources.

Acknowledgments. This research was partially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 288024 (LiMoSINe project), the Netherlands Organisation for Scientific Research (NWO) under project nrs 640.-004.802, 727.011.005, 612.001.116, HOR-11-10, the Center for Creation, Content and Technology (CCCT), the QuaMerdes project funded by the CLARIN-nl program, the TROVe project funded by the CLARIAH program, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project number 027.012.105 and the Yahoo! Faculty Research and Engagement Program.

Bibliography

[1] A. E. C. Basave, A. Varga, M. Rowe, M. Stankovic, and A.-S. Dadzie. Making sense of microposts (#msm2013) concept extraction challenge. In *MSM '13*, pages 1–15, 2013.

[2] M. Becker, B. Hachey, B. Alex, and C. Grover. Optimising selective sampling for bootstrapping named entity recognition. In *ICML-LMV '05*, pages 5–11, 2005.

[3] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, and N. Aswani. TwitIE: An open-source information extraction pipeline for microblog text. In *RANLP '13*. ACL, 2013.

[4] L. Buitinck and M. Marx. Two-stage named-entity recognition using averaged perceptrons. In *NLDB'12*, pages 171–176. Springer-Verlag, 2012.

[5] R. Bunescu. Using encyclopedic knowledge for named entity disambiguation. In *EACL '06*, pages 9–16, 2006.

[6] T. Cassidy, H. Ji, L.-A. Ratinov, A. Zubiaga, and H. Huang. Analysis and enhancement of wikification for microblogs with context expansion. In *COLING '12*, pages 441–456, 2012.

[7] M. Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *ACL '02*, pages 1–8, 2002.

[8] A. Davis, A. Veloso, A. S. da Silva, W. Meira, Jr., and A. H. F. Laender. Named entity disambiguation in streaming data. In *ACL '12*, pages 815–824, 2012.

[9] L. Derczynski, D. Maynard, N. Aswani, and K. Bontcheva. Microblog-genre noise and impact on semantic annotation accuracy. In *HT '13*, pages 21–30. ACM, 2013.

[10] J. R. Finkel, C. D. Manning, and A. Y. Ng. Solving the problem of cascading errors: approximate bayesian inference for linguistic annotation pipelines. In *EMNLP '06*, pages 618–626, 2006.

[11] J. R. Firth. A Synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis*, pages 1–32, 1957.

[12] S. Guo, M.-W. Chang, and E. Kiciman. To link or not to link? a study on end-to-end tweet entity linking. In *NAACL '13*, pages 1020–1030, 2013.

[13] M. Hu, A. Sun, and E.-P. Lim. Comments-oriented document summarization: understanding documents with readers' feedback. In *SIGIR '08*, pages 291–298. ACM, 2008.

[14] V. Jijkoun, M. Khalid, M. Marx, and M. de Rijke. Named entity normalization in user generated content. In *AND '08*, 2008.

[15] Z. Kozareva. Bootstrapping named entity recognition with automatically generated gazetteer lists. In *EACL-SRW '06*, pages 15–21. ACL, 2006.

[16] J. Lin, C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2011 Microblog track. In *TREC 2011*, 2012.

[17] T. Lin, Mausam, and O. Etzioni. No noun phrase left behind: detecting and typing unlinkable entities. In *EMNLP-CoNLL '12*, pages 893–903, 2012.

[18] K. Massoudi, M. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In *ECIR '11*, pages 362–367, 2011.

[19] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *WSDM '12*, pages 563–572, 2012.

[20] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07*, pages 233–242, 2007.

[21] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM '08*, pages 509–518, 2008.

[22] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151 – 175, 2013.

[23] D. Odijk, E. Meij, and M. de Rijke. Feeding the second screen: Semantic linking based on subtitles. In *OAIR '13*, 2013.

[24] L. Ramshaw and M. Marcus. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*. ACL, 1995.

[25] A. Ritter, Mausam, O. Etzioni, and S. Clark. Open domain event extraction from twitter. In *KDD '12*, 2012.

[26] W. Weerkamp and M. de Rijke. Credibility-inspired ranking for blog post retrieval. *Information Retrieval*, 15(3–4):243–277, 2012.

[27] D. Wu, W. S. Lee, N. Ye, and H. L. Chieu. Domain adaptive bootstrapping for named entity recognition. In *EMNLP '09*, pages 1523–1532. ACL, 2009.

[28] Y. Zhou, L. Nie, O. Rouhani-Kalleh, F. Vasile, and S. Gaffney. Resolving surface forms to wikipedia topics. In *COLING '10*, pages 1335–1343, 2010.