Towards Entity Correctness, Completeness and Emergence for Entity Recognition

Lei Zhang, Yunpeng Dong, Achim Rettinger
Karlsruhe Institute of Technology (KIT), 76128 Karlsruhe, Germany
{l.zhang,rettinger}@kit.edu, yunpeng.dong@student.kit.edu

ABSTRACT

Linking unstructured text to knowledge bases (KBs) by mapping words or phrases to the corresponding entities in KBs, is the problem of entity recognition and disambiguation. In this paper, we focus on the task of entity recognition in Web text to address the challenges of *entity correctness*, *completeness* and *emergence* that exiting approaches mainly suffer from. Experimental results show that the proposed approach significantly outperforms the state-of-the-art approaches in terms of precision, F-measure, micro-accuracy and macro-accuracy, while still preserving high recall.

1. INTRODUCTION

In recent years, large repositories of structured knowledge, such as Wikipedia, have become valuable resources for knowledge extraction, especially for the automatic aggregation of knowledge from Web text. In this regard, entity linking (EL), which leverages such knowledge bases (KBs) to link words or phrases in Web text with entities, has emerged as a topic of major interest. The challenges of entity linking lie in entity recognition (ER) and disambiguation (ED). The first stage, ER, serves to detect words or phrases in text, also called *mentions*, that are likely to denote entities; the second stage, ED, performs the disambiguation of the recognized mentions into entities. This paper focus on ER in Web text for linking entities with Wikipedia, where entities can be either *named entities (NEs)* or *nominal entities (NOEs)*. For instance, in the sentence "US President Barack Obama will land in India for a three-day visit.", two mentions *Barack Obama* and *India* refer to the NEs Barack_Obama and India, the other two mentions *US President* and *visit* refer to the NOEs President_of_the_United_States and State_visit.

For ER, some existing approaches [4, 5] first gather all n-grams from text and the ones matching surface forms that can refer to any entities in KB are retained for ED. These approaches can detect both NEs and NOEs but could generate a lot of noise, i.e., mentions without actual referent entities in KB, which results in the challenge of *ER correctness*. Recently, part-of-speech (POS) tags have been exploited to find mentions that are primarily noun phrases [6]. However, such approaches do not solve the problem, since either the noise still remains or some expected mentions are missing.

Another major challenge regarding *ER correctness* is *overlapping mentions*. EL systems usually make the assumption that only the complete mentions are linked with entities. For example, the complete mention *US President* should be detected and linked but its partial ones *US* and *President* should not. However, both n-gram and POS based approaches suffer from the problem of overlapping mentions.

In some other work [3, 2], named entity recognition (NER) has been performed on the input text to detect NEs, which are then used for ED, where NOEs are ignored. And NER cannot even detect NEs when they are mentioned as abbreviations. For instance, in the sentence "Edward Snowden revealed PRISM.", most of the NER systems, e.g. the Stanford NER Tagger [1], normally only detect *Edward Snowden* as NE but not *PRISM*, which actually refers to the NE PRISM_(surveillance_program). Therefore, NER based approaches result in the challenge of *ER completeness*.

Due to the highly dynamic Web contents, *emerging entities (EEs)* has become an additional challenge of ER in Web text[1]. Consider the Web news about the disclosure of the PRISM program by Edward Snowden containing two EEs PRISM_(surveillance_program) and Edward_Snowden, which are assumed not to be covered by the indexed KB. NER based approaches usually can only detect *Edward Snowden* but not *PRISM* as discussed before. Regarding n-gram and POS based approaches, EEs can not be detected due to fact that there might be no corresponding surface forms in KB.

2. APPROACH

In order to address the challenges of *ER correctness* and *completeness*, we combine NER with POS analysis. Given a Web text t published on date d, we first feed it into a NER system and collect the output $\mathsf{M}_{\mathsf{NER}} = \{m_e | \forall e \in \text{NER}(t)\}$, which is the set of mentions m_e of the NEs e detected by NER. Then we perform the POS analysis on t and extract all sequences conforming to a set of predefined POS patterns[2], which extract all proper nouns and other possible combinations matching entities, serving as candidate mentions.

For the challenge of *emerging entities (EEs)*, we exploit the Wikipedia page view statistics, which capture the num-

[1]Since most of the EL systems have to index the KB before online processing, EEs denote entities that are not covered by the indexed KB instead of the available latest version of the KB, which makes ER in Web text more challenging.

[2]The POS patterns used in this work are available online at `http://people.aifb.kit.edu/lzh/er/pos_patterns.pdf`.

ber of times Wikipedia pages, including non-existent pages, have been requested, and can be treated as a query log of entities, including EEs. Since it is very likely that EEs will be requested in real-time (such as due to a current event), the page view statistics are valuable source for detecting EEs.

In this regard, the mentions detected using POS patterns have to satisfy one of the following conditions: (1) they have been used to refer to entities in Wikipedia; (2) they have been requested in Wikipedia page view in the vicinity of the publishing date d of t. Then we obtain the set of mentions using POS patterns as $\mathsf{M}_{\mathsf{POS}} = \{m | \forall m \in \mathrm{POS}(t) : m \in \mathrm{POS\ Patterns} \wedge (freq_{link}(m) > 0 \vee freq_{view}(m,d) > 0)\}$, where $freq_{link}(m)$ is the number of links using m as anchor text pointing to entities and $freq_{view}(m,d)$ is the maximum frequency of page view requests in the vicinity of the publishing date d of the input Web text. More specifically, we track the page view requests of m on d and the preceding $n-1$ days and calculate $freq_{view}(m,d)$ as

$$freq_{view}(m,d) = \max_{d_i \in [d-n+1,d]} freq^{d_i}_{view}(m) \qquad (1)$$

where $freq^{d_i}_{view}(m)$ is the number of page view requests of m on date d_i. By taking into account both $\mathsf{M}_{\mathsf{NER}}$ and $\mathsf{M}_{\mathsf{POS}}$, we obtain the set of candidate mentions as $\mathsf{M} = \mathsf{M}_{\mathsf{NER}} \cup \mathsf{M}_{\mathsf{POS}}$.

In order to overcome the challenge of *overlapping mentions*, we then calculate the score of each mention as follows

$$\mathrm{Score}(m) = \mathrm{Score}_{freq}(m) \cdot \mathrm{Score}_{idf}(m) \cdot \mathrm{Boost}(|m|) \qquad (2)$$

The conflicting mentions with smaller score can be filtered out. In the following, we discuss the components in Eq. 2.

First of all, we calculate the frequency of m by leveraging both Wikipedia links and page view requests of m as

$$freq(m) = freq_{link}(m) + \beta \cdot freq_{view}(m,d) \qquad (3)$$

While $freq_{link}(m)$ represents the general popularity of m based on KB, $freq_{view}(m,d)$ captures the temporal importance of m w.r.t. the date d based on the page view requests, which thus can also help with the problem of *overlapping mentions*. Due to the different scales of Wikipedia link and page view request frequencies, $freq_{view}(m,d)$ is adjusted by a balance parameter $\beta = \frac{\text{total number of links in Wikipedia}}{\text{average number of page views per day}}$, where the fraction accounts for the difference in frequency of Wikipeida links and per-day page view requests of m.

In general, we can use $freq(m)$ to calculate $\mathrm{Score}_{freq}(m)$. However, for mention m of EE detected by NER that appears neither in KB nor in page view requests, i.e., $freq(m) = 0$, we make use of the maximal frequency among its term subsequences $m' \sqsubseteq m$, given by the following score

$$\mathrm{Score}_{freq}(m) = \begin{cases} \max_{m' \sqsubseteq m} freq(m') & \text{if } m \in M_{NER}, \\ & freq(m) = 0 \\ freq(m) & \text{otherwise} \end{cases} \qquad (4)$$

Furthermore, we consider the inverse document frequency (idf) of m, which captures a notion of how important the terms in m are, to penalize common terms. Thus we have

$$\mathrm{Score}_{idf}(m) = \frac{1}{freq_{doc}(m)} \qquad (5)$$

where $freq_{doc}(m)$ is the number of articles containing m.

The function Boost() is used to boost the score of a long mention by its length $|m|$, i.e. the number of terms in m, as

$$\mathrm{Boost}(|m|) = \exp(\gamma \cdot |m|) \qquad (6)$$

where the tunable parameter γ reflects the sensitivity to long mentions in the entity recognition process.

Methods	Prec.	Rec.	F1	Mic. Acc.	Mac. Acc.
n-gram [4, 5]	0.22	0.93	0.35	0.21	0.21
NER [1]	0.80	0.24	0.36	0.22	0.22
POS [6]	0.61	0.90	0.73	0.56	0.58
NER+n-gram	0.22	**0.96**	0.36	0.21	0.22
NER+POS	0.61	0.94	0.74	0.58	0.59
Our Approach	**0.86**	0.90	**0.88**	**0.78**	**0.79**

Table 1: The Experimental Results.

3. EVALUATION

We now discuss the experiments performed to assess the performance of our ER approach. We use the English version of Wikipedia from July 2013 as the KB. Since there are no existing datasets of recent Web text containing EEs, we have created a new dataset[3] by collecting 100 news articles on the Web in 2014, where 26 of them contain EEs. We first automatically recognized mentions using the Stanford POS and NER taggers and then corrected them manually.

We conducted the experiments with our approach and several baselines: the *n-gram* based approach introduced by [4, 5]; the *NER* based approach using the Stanford NER Tagger [1]; the *POS* based approach proposed by [6]; the other two baselines combining *NER* with *n-gram* and *POS* respectively, i.e., using the union of their outputs.

We experimented with different values of γ and observed that the performance of our approach improves from $\gamma = 0.1$ to $\gamma = 0.9$, then reaches a plateau. The experimental results of the baselines and our approach with $\gamma = 0.9$ are shown in Table 1. Our approach clearly outperforms the baselines in terms of precision, F-measure, micro-accuracy and macro-accuracy, while still preserving high recall.

4. CONCLUSIONS

In this paper, we propose a new approach to entity recognition in Web text for addressing the challenges of entity emergence, correctness and completeness that existing approaches mainly suffer from. We have experimentally shown that our approach achieves a significant improvement over the baselines. Our future work will integrate the proposed approach to entity recognition into an entity linking system to show that the improvement of entity recognition can also carry over to entity linking.

5. REFERENCES

[1] J. R. Finkel, T. Grenager, and C. D. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, 2005.

[2] J. Hoffart, Y. Altun, and G. Weikum. Discovering emerging entities with ambiguous names. In *WWW*, pages 385–396, 2014.

[3] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *EMNLP*, pages 782–792, 2011.

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

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

[6] S. Zhao, C. Li, S. Ma, T. Ma, and D. Ma. Combining pos tagging, lucene search and similarity metrics for entity linking. In *WISE*, pages 503–509, 2013.

[3] The dataset used in our experiments is online available at `http://people.aifb.kit.edu/lzh/er/er_dataset.tgz`.