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 words or phrases in unstructured text to entities in knowledge bases 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 existing approaches mainly suffer from. Experimental results show that our 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, which leverages such knowledge bases (KBs) to link words or phrases in unstructured text with entities in KBs, has emerged as a topic of major interest. The challenges of entity linking lie in entity recognition (ER) and entity 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 focuses on the task of ER in Web text for later ED with Wikipedia, where entities can be either named entities (NEs) or nominal entities (NOEs) in Wikipedia. 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 are used to refer to any entities in KBs 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, 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. Entity linking 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 missing. Furthermore, NER systems sometimes cannot detect all NEs, due to the limitation of selected algorithms and training data. For instance, in the sentence “Edward Snowden revealed Prism”, the Stanford NER Tagger [1] only detects 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) have become an additional challenge of ER in Web text1. 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 cannot be detected because there might be no corresponding surface forms found in KBs.

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 \( M_{\text{NER}} = \{m_e | e \in \text{NER}() \} \), which is the set of mentions \( m_e \) of the NEs detected by NER. Then we perform the POS analysis on \( t \) and extract all sequences conforming to a set of predefined POS patterns2, 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-

---

1Since entity linking systems usually 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.

2The POS patterns used in this work are available online at http://people.aifb.kit.edu/lzh/er/pos_patterns.pdf.
Table 1: The Experimental Results.

3. EXPERIMENTS

For the experiments, we used the English Wikipedia snapshot from July 2013 as the KB. Since there are no existing datasets of recent Web text containing EEs, we created a new dataset\(^3\) of 100 Web documents in 2014, where 26 of them contain EEs. Every mention was manually annotated by two volunteers, conflicts were reconciled by the authors.

We conducted the experiments with our approach and several baselines: the n-gram based approach used in [4,5]; the NER based approach used in [3,2] based on the Stanford NER Tagger [1]; the POS based approach proposed by [6]; the two other baselines combining NER with n-gram and POS respectively, i.e., using the union of their outputs.

We experimented with different values of \(\gamma\) 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 ER for addressing the main challenges of entity correctness, completeness and emergence. We have experimentally shown that our approach achieves a significant improvement over the baselines. Our future work will integrate our ER approach into an entity linking system to show that the improvement of ER can also carry over to the entire entity linking process.

5. ACKNOWLEDGMENTS

This work was partially supported by the EU FP7 project XLI-Me (Grant 611346).

6. REFERENCES


\(^{\text{3}}\)The dataset used in our experiments is online available at http://people.salford.ac.uk/lzh/er/er_dataset.1tg.