Named Entity Extraction and Disambiguation: The Missing Link

Mena B. Habib and Maurice van Keulen
Faculty of EEMCS, University of Twente, Enschede, The Netherlands
{m.b.habib,m.vankeulen}@ewi.utwente.nl

ABSTRACT

Named entity extraction (NEE) and disambiguation (NED) are two areas of research that are well covered in literature. Typical fields addressing these topics are information retrieval, natural language processing, and semantic web. Although these topics are highly dependent, almost no existing works examine this dependency. It is the aim of this position paper to explore that dependency and show how one affects the other, and vice versa. We show the benefit of using this reinforcement effect on two domains: NEE and NED for toponyms in formal text; and for arbitrary entity types in informal short text in tweets. Finally we give an insight about the potential of this approach for future research.

Categories and Subject Descriptors

I.7 [**Document and Text Processing**]: Miscellaneous; H.3.1 [**Information Systems**]: Content Analysis and Indexing-Linguistic processing

General Terms

Algorithms

Keywords

Named Entity Extraction; Named Entity Disambiguation; Uncertain Annotations

1. INTRODUCTION

Named entities (NEs) are atomic elements in text belonging to predefined entity types such as persons, organizations, locations, etc. NEE is a sub task of information extraction that seeks to locate those elements in text. NED is the task of determining which real entity is referred to by a certain mention of a name. In this position paper we answer the following research questions regarding the relation between NEE and NED: *a*) How the imperfection of the extraction process affects the effectiveness of disambiguation process. *b*) Whether the extraction confidence can be used to improve the effectiveness of disambiguation. *c*) How disambiguation results can be used to improve the quality of extraction. *d*) How NEE and NED can be domain and language independent. We investigate the answers for the aforementioned questions on two domains: NEE and NED for toponyms in formal text; and for arbitrary entity types in informal short text in tweets.

The general principal we claim is that NED could be very helpful in improving the NEE process. For example, consider the tweet '– ***Lady Gaga*** - ***Speechless*** *live @* ***Helsinki*** *10/13/2010 http://www.youtube.com/watch?v=yREociHyijk @ladygaga also talks about her Grampa who died recently*' where named entities are marked in bold. It is uncertain, even for humans, to recognize **Speechless** as a song name without having a prior information about **Lady Gaga**'s songs.

Although the logical order for an Information Extraction (IE) system is to do extraction first then the disambiguation, we always start with a phase of extraction which aims to achieve high recall (find as much NE candidates as possible) then we apply the disambiguation for all the extracted NE. Finally we filter those extracted NE candidates into true positives and false positives using features derived from the disambiguation phase in addition to other shape and Knowledge-Base (KB) features. The potential of this order is that disambiguation step would give extra information (such as entity-context similarity and entity-entity coherency) about each NE candidate that might help in the decision if this candidate is a true NE or not.

2. TOPONYM EXTRACTION AND DISAMBIGUATION

Toponyms are names referring to locations such as '*Lake Como*' or '*Museum of Modern Arts*'. To answer the research questions, we conducted experiments with a set of holiday home descriptions with the aim to extract and disambiguate toponyms [1]. The task we focus on is to infer from the extracted toponyms the country where the holiday property is located. The context of country inference aids in disambiguating the extracted toponyms. A rule based approach is used for extraction. We investigated how the effectiveness of disambiguation is affected by the effectiveness of extraction by comparing with results based on manually extracted toponyms. We also investigated the reverse measuring the effectiveness of extraction when filtering out those toponyms found to be highly ambiguous, and in turn, measure the ef-

ESAIR'13, October 28, 2013, San Francisco, CA, USA.

Enter the DOI string/url from the ACM e-form confirmation ...$15.00.

fectiveness of disambiguation based on this filtered set of toponyms. Results showed that the effectiveness of extraction and, in turn, disambiguation improved, thereby showing that both can reinforce each other. We called this potential for mutual improvement, the reinforcement effect (see Figure 1).

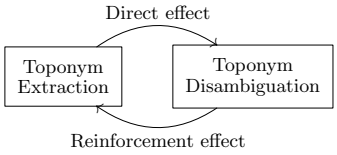

Figure 1: The reinforcement effect between the toponym extraction and disambiguation processes.

In [2], we examined statistical approaches for toponym extraction (Hidden Markov Models (HMM) and Conditional Random Fields (CRF)). The advantage of statistical techniques for extraction is that they provide alternatives for annotations along with confidence probabilities. Instead of discarding these, as is commonly done by selecting the topmost likely candidate, we use them to enrich the knowledge for disambiguation (i.e. annotations are truly probabilistic). The probabilities proved to be useful in enhancing the disambiguation process. We believe that there is much potential in making the inherent uncertainty in information extraction explicit in this way. Furthermore, extraction models are inherently imperfect and generate imprecise confidence. We were able to use the disambiguation result (toponym-country co-occurrence) to enhance the confidence of true positives and reduce the confidence of false positives. This enhancement of extraction improves as a consequence the disambiguation (the aforementioned reinforcement effect). This process can be repeated iteratively, without any human interference, until there is no more improvement in the extraction and disambiguation. In this way, the context in which a certain name occurs can be used to automatically enhance training these by improving the extraction and disambiguation of that name.

To investigate the language independence of our concepts, we proposed a hybrid toponym extraction approach based on HMM and Support Vector Machines (SVM) [3]. HMM is used for extraction with high recall and low precision. Then SVM is used to find false positives based on informativeness features and coherence features derived from the disambiguation results. Experimental results showed that the proposed approach outperform the state of the art methods of extraction and also proved to be robust. Robustness is proved on three aspects: language independence, high and low HMM threshold settings, and limited training data.

3. NAMED ENTITY EXTRACTION AND DISAMBIGUATION IN TWEETS

Short context messages (like tweets and SMS's) are a potentially rich source of continuously and instantly updated information. Shortness and informality of such messages are challenges for Natural Language Processing tasks.

To verify our concepts in the domain of informal text we presented two systems for NEE from tweets. The first is an unsupervised system to improve the extraction process by using clues from the disambiguation process [4]. For extraction we used a simple Knowledge-Base matching technique. This method of extraction achieves high recall and low precision. For disambiguation, we developed a simple algorithm which assumes that the correct entities for mentions appearing in the same message should be related to each other in YAGO KB graph. Based on this coherency feature, we were able to discover false positives and thereby improve the precision and F1 measure.

The second system is a supervised one which represents a hybrid approach for Named Entity Extraction (NEE) and Classification (NEC) for tweets [5]. The system uses the power of the Conditional Random Fields (CRF) and the Support Vector Machines (SVM) in a hybrid way to achieve better results. For named entity type classification we used AIDA [6] disambiguation system to disambiguate the extracted named entities and hence find their type.

NED in tweets is challenging in two ways. First, the limited length of Tweet makes it hard to have enough context while many disambiguation techniques depend on it. The second is that many named entities in tweets do not exist in a knowledge base (KB). We combine ideas from information retrieval (IR) and NED to propose solutions for both challenges [7]. For the first problem we make use of the gregarious nature of tweets to get enough context needed for disambiguation. For the second problem we look for an alternative home page if there is no Wikipedia page represents the entity. Given a mention, we obtain a list of Wikipedia candidates from YAGO KB in addition to top ranked pages from Google search engine. We use Support Vector Machine (SVM) to rank the candidate pages to find the best representative entities. Experiments conducted on two data sets show better disambiguation results compared with the baselines and a competitor.

4. CONCLUSIONS AND FUTURE WORK

Named entity extraction and disambiguation are highly dependent processes. We examined how handling the uncertainty of extraction influences the effectiveness of disambiguation, and reciprocally, how the result of disambiguation can be used to improve the effectiveness of extraction. This concept is proved experimentally to be language independent. Furthermore, we introduced a supervised and an unsupervised approaches for NEE in short context using clues from NED. Finally, we presented a solution to overcome challenges in NED in tweets.

Our general approach is beneficial in many future research directions. The approach can potentially adapt itself to any domain. Moreover, it can be used to enrich existing knowledge bases by new entries. For example, we could find an estimation for a location of a toponym that has no entry in knowledge base given other disambiguated toponyms on the same context. It can also be used to build a knowledge base for closed domains from user generated contents. For example we could draw a rough map for a city center area using tweets sent about some event held there.

5. REFERENCES

[1] Mena B. Habib and M. van Keulen. Named entity extraction and disambiguation: The reinforcement effect. In *Proceedings of the 5th International Workshop on Management of Uncertain Data, MUD 2011, Seatle, USA*, pages 9–16, 2011.

[2] Mena B. Habib and M. van Keulen. Improving toponym disambiguation by iteratively enhancing certainty of extraction. In *Proceedings of the 4th*

International Conference on Knowledge Discovery and Information Retrieval, KDIR 2012, Barcelona, Spain, pages 399–410, 2012.

[3] Mena B. Habib and M. van Keulen. A hybrid approach for robust multilingual toponym extraction and disambiguation. In *Proceedings of the International Conference on Language Processing and Intelligent Information Systems (LP&IIS 2013), Warsaw, Poland*, 2013.

[4] Mena B. Habib and M. van Keulen. Unsupervised improvement of named entity extraction in short informal context using disambiguation clues. In *Workshop on Semantic Web and Information Extraction, SWAIE 2012, Galway, Ireland*, pages 1–10, 2012.

[5] Mena B. Habib, M. van Keulen, and Z. Zhu. Concept extraction challenge: University of twente at #msm2013. In *Proceedings of the 3rd workshop on 'Making Sense of Microposts' (#MSM2013), Rio de Janeiro, Brazil*, 2013.

[6] Mohamed Amir Yosef, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum. Aida: An online tool for accurate disambiguation of named entities in text and tables. *PVLDB*, 4(12):1450–1453, 2011.

[7] Mena B. Habib and M. van Keulen. A generic openworld named entity disambiguation approach for tweets. In *Proceedings of the 5th International Conference on Knowledge Discovery and Information Retrieval, KDIR 2013, Vilamoura, Portugal*, 2013.