The Impact of Named Entity Normalization on Information Retrieval for Question Answering

Mahboob Alam Khalid, Valentin Jijkoun, and Maarten de Rijke

ISLA, University of Amsterdam
mahboob,jijkoun,mdr@science.uva.nl

Abstract. In the named entity normalization task, a system identifies a canonical unambiguous referent for names like *Bush* or *Alabama*. Resolving synonymy and ambiguity of such names can benefit end-to-end information access tasks. We evaluate two entity normalization methods based on Wikipedia in the context of both passage and document retrieval for question anwering. We find that even a simple normalization method leads to improvements of early precision, both for document and passage retrieval. Moreover, better normalization results in better retrieval performance.

1 Introduction

The task of recognizing named entities in text, i.e., identifying character sequences that refer to items like persons, locations, organizations, dates, etc., has been studied extensively. The Named Entity Recognition (NER) task has been thouroughly evaluated within the Conference on Computational Natural Language Learning (CoNNL) framework in a language-independent setting; techniques applied to NER range from rule-based [9] to machine learning-based [12, 4]. Though significant progress has been achieved, the task remains challenging due to a lack of uniformity in writing styles and domain-dependency. Moreover, NER results are often difficult to use directly, due to high synonymy and ambiguity of names across documents [12]. E.g., the strings *U.S.*, *USA*, *America* can all be used to refer to the concept *United States of America*. Similarly, the string *Washington* can be used to refer to different entities (e.g., *Washington, DC*, or *USA*, or *George Washington*). For information access tasks, such as document retrieval or question answering, these phenomena may harm the performance.

One approach to addressing these problems is Named Entity Normalization (NEN), which goes beyond the NER task: names are not only identified, but also normalized to the concepts they refer to. NEN addresses two phenomena. First, *ambiguity* arises when distinct concepts share the same name; e.g., *Alabama* may refer to the *University of Alabama*, the *Alabama river*, or the *State of Alabama*. This calls for the named entity disambiguation. Second, *synonymy* arises when different names refer to the same entity; e.g., *America* and *U.S.* refering to the *United States of America*.

The multi-referent ambiguity problem was considered at the SemEval Web People Search task [3] and in the Spock Entity Resolution Challenge.[1] Both efforts focus on a web search task where the goal is to organize web pages found using a person name as a search engine query, into clusters where pages within a cluster refer to the same person. Cucerzan [8] describes a method for addressing both ambiguity and synonymy; the method uses Wikipedia data and is applied to news texts as well as to Wikipedia itself.

We investigate the impact of NEN on two specific information access tasks: document and passage retrieval for question answering (QA). The tasks consist in finding items in a collection of documents, which contain an answer to a natural language question. E.g., for the question *Who is the queen of Holland?*, an item containing *Beatrix, the Queen of the Kingdom of the Netherlands...* is a relevant response, given that *Holland* is used as a synonym of the *Kingdom of the Netherlands*. Here, NEN may allow a retrieval system to find the answer passage which may have been missed with a standard term-based retrieval.

Specifically, we answer the following research questions: (1) Does NEN improve performance of passage or document retrieval for QA? and (2) To what extent does better entity normalization result in better retrieval for QA? We describe and compare two Wikipedia-based entity normalization methods and evaluate their effectiveness in the setting of passage and document retrieval for QA, using the test collection of the TREC QA track [14].

In Section 2 we review related work. Then, in Section 3, we present two entity normalization methods. Section 4 provides the details of the experimental setup, and shows the results. We conclude in Section 5.

2 Related Work

NEN has been studied both in restricted and in open domains. In the domain of genomics, where gene and protein names can be both synonymous and ambiguious, Cohen [7] normalizes entities using dictionaries automatically extracted from gene databases. Zhou et al. [15] show that appropriate use of domain-specific knowledge base (i.e., synonyms, hypernyms, etc., in a certain domain) yields significant improvement in passage retrieval. For the news domain, Magdy et al. [12] address cross-document Arabic person name normalization using a machine learning approach, a dictionary of person names and frequency information for names in a collection. They apply their method for normalizing Arabic names on the documents related to the situation of Gaza and Lebanon taken from news.google.com. Cucerzan [8] addresses an open domain normalization task, normalizing named entities with information extracted from Wikipedia and machine learning for context-aware disambiguation.

3 Named Entity Normalization

We experimented with two versions of an NEN method based on Wikipedia. Wikipedia is widely used as a rich semantic resource, with natural language processing applications ranging from question answering [2] to text classification [11]

[1] http://challenge.spock.com

to named entity disambiguation [6, 8]. Wikipedia is especially attractive for the task of entity normalization. It covers a huge number of entities (over 2M article titles as of October 2007), most of them named entities. The anchor text of inter-article links allows one to identify different text strings that can be used to refer to the same entity or concept. So-called "redirects" provide information about synonyms or near synonyms (e.g., the article *King of pop* is empty and redirects to the article *Michael Jackson*). Special "disambiguation" pages list possible referents of ambiguous names (such as *George Bush* that lists five persons with that name). Moreover, each Wikipedia entity page has a unique identifier (URL)—a unique and unambiguous way of refering to the entity.

The baseline NEN method in [8] uses this information in the following manner, for each surface form recognized as an NE by an NE recognizer. If there is an entity page or redirect page whose title matches exactly with the surface form, then the corresponding entity is chosen as the normalization result; otherwise the entity most frequently mentioned in Wikipedia using that form as anchor text is selected as the baseline disambiguation. We re-implemented this baseline using the named entity tagger of [10], and refer to it as *MS*.

We also implemented a simple extension of the method by adding a link frequency-based disambiguation algorithm. Whenever a surface form can be resolved to more than one entity using the algorithm above, we select the entity with the highest number of incoming hyperlinks. Our hypothesis of disambiguation is based on the assumption that a more useful and/or popular Wikipedia entity will have many links pointing to it [5]. In other words, we assume that a name found (e.g., "Bush") mostly refers to the most popular compatible Wikipedia entity ("George W. Bush"). We refer to this method as *NN*.

Cucerzan [8] also describes a more sophisticated, context-aware normalization algorithm. We did not use this version of the algorithm in our experiments below because it would have involved classifying each name in the collection—a very computationally expensive step.

We compared the *MS* and *NN* normalization methods, using the evaluation data as described in [8] for intrinsic, stand-alone evaluation of the two methods. The accuracy of *NN* on Wikipedia articles and news articles was 86.5% and 73% respectively, outperforming the accuracy of *MS* (86.1% and 51.7% on Wikipedia and news articles, respectively).

4 Experiments and Results

We performed a number of experiments in a setting similar to [13]. We used a standard set of question/answer pairs from TREC QA tasks of 2001–2003. In addition to using full documents, we split the AQUAINT corpus into 400-character passages (aligned on paragraph boundary). We ran the NER tool of [10] to detect named entities and normalized them using *NN* and *MS*, separately. We used the dump of English Wikipedia from November 2006. Documents and passages were separately indexed using Lucene [1]. Out of 2,136 question/answer pairs in the TREC QA data, we used only 1,215 whose questions contained a named entity. We normalized named entities in questions in the same way as in the collection. We compared the retrieval performance of the baseline (no normalization,

	MRR	s@1	s@5	s@10	p@5	p@10
NONORM	0.532	44.8%	64.6%	72.8%	0.37	0.34
MS	0.511	42.2%	63.4%	71.6%	0.36	0.32
NN	0.523	43.4%	64.4%	72.7%	0.37	0.33
MS+NONORM	0.55	46.4%	67.57%*	74.8%*	0.39**	0.35*
NN+NONORM	**0.56**	**47**%*	**68.2**%**	**75.3**%**	**0.4****	**0.36***

Table 1. Impact of named entity normalization on document retrieval for QA; * and ** indicate significant improvements over the baseline at p=0.05 and p=0.01.

	MRR	s@1	s@5	s@10	p@5	p@10
NONORM	0.411	30.9%	53.6%	**63.3**%	0.26	**0.23**
MS	0.387	29.2%	50%	58.9%	0.23	0.2
NN	0.405	30.7%	51.7%	60.5%	0.24	0.21
MS+NONORM	0.407	30.7%	53%	61.2%	0.24	**0.23**
NN+NONORM	**0.424**	**32.6**%	**54.4**%	62.3%	**0.27**	**0.23**

Table 2. Impact of named entity normalization on passage retrieval for QA. *NN+NONORM* outperforms *MS+NONORM* (at p=0.01).

standard vector space retrieval), for both normalization methods and for equally weighted mixture models of the baseline with both methods. Following [13], we measured performance using the Mean Reciprocal Rank (MRR), success at rank n ($s@n$), and average precision at n ($p@n$). For significance testing we applied the McNemar significance test on success evaluations, and Student's t-test on precision evaluations. Tables 1 and 2 show the evaluation results for passage and document retrieval, respectively.

The results show that the combination of NEN with the baseline improves the MRR value, precision and early success of the retrieval system for QA. They also show that *NN* helps more than *MS*, for document and passage retrieval.

An analysis of the effect of NEN on text retrieval shows that for questions where normalization did not improve the retrieval, this was mostly due to NER errors. E.g., for *What river is under New York's George Washington bridge?*, the entity *George Washington* was detected as a person name, while the answer passage contains the entity *George Washington Bridge* correctly detected as LOCATION. Where normalization helped to find relevant passages, this was often due to the correct "gluing" of multiword units: *Buffalo Bill*, *Crater Lake*, *Joe Andrew*, *Andrew Jackson*. Here, without normalization, retrieval failed.

Finally, for the passage retrieval experiments, the difference between *NN* and *MS* is statistically significant (at $p = 0.01$). This indicates that better normalization does indeed lead to better retrieval performance.

5 Conclusion

We described experiments evaluating the impact of name entity normalization on document and passage retrieval for QA. We implemented the normalization method of [8] and a simple refinement. Although our disambiguation methods are not context-aware, we observed improved retrieval performance with entity normalization. Moreover, better normalization has led to better QA performance.

The error analysis shows that entity recognition errors are a main source of retrieval errors due to normalization. This indicates an obvious direction for improving the system. Another item for future work is to include surface form context into the disambiguation model in such a way that normalizing a large text collection remains computationally tractable.

Acknowledgements. Mahboob Alam Khalid was supported by the Netherlands Organization for Scientific Research (NWO) under project number 612.-066.512, Valentin Jijkoun—by NWO projects 220.80.001, 600.065.120 and 612.-000.106. Maarten de Rijke was supported under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.065.120, 612-13-001, 612.000.106, 612.-066.302, 612.069.006, 640.001.501, 640.002.501, and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

Bibliography

[1] Jakarta Lucene ext search engine. http://lucene.apache.org, 2002.

[2] D. Ahn, V. Jijkoun, G. Mishne, K. Müller, M. de Rijke, and S. Schlobach. Using Wikipedia at the TREC QA Track. In *TREC '04*, 2005.

[3] J. Artiles, J. Gonzalo, and S. Sekine. The SemEval-2007 WePS Evaluation: Establishing a benchmark for Web People Search Task. In *Semeval '07*, 2007.

[4] A. Borthwick. *A Maximum Entropy Approach to Named Entity Recognition*. PhD thesis, New York University, 1999.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[6] R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL 2006*, 2006.

[7] A. M. Cohen. Unsupervised gene/protein named entity normalization using automatically extracted dictionaries. In *ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases*, pages 17–24, 2005.

[8] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL '07*, pages 708–716, 2007.

[9] D. Farmakiotou, V. Karkaletsis, J. Koutsias, G. Sigletos, C. Spyropoulos, and P. Stamatopoulos. Rule-based named entity recognition for Greek financial texts. In *Proceedings COMLEX 2000*, 2000.

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

[11] E. Gabrilovich and S. Markovitch. Overcoming the Brittleness Bottleneck using Wikipedia. In *AAAI 2006*, 2006.

[12] W. Magdy, K. Darwish, O. Emam, and H. Hassan. Arabic cross-document person name normalization. In *CASL Workshop '07*, pages 25–32, 2007.

[13] C. Monz. Minimal span weighting retrieval for question answering. In *Proceedings of SIGIR 2004 Workshop on Information Retrieval for Question Answering*, 2004.

[14] E. M. Voorhees. Overview of the TREC 2003 Question Answering Track. In *TREC*, pages 54–68, 2003.

[15] W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *SIGIR '07*, pages 655–662, 2007.