Exploring Linguistic Features for Cross-document Named Entity Disambiguation

Shuangshuang Zhou Canasai Kruengkrai Kentaro Inui
Graduate School of Information Sciences, Tohoku University
{shuang,canasai,inui}@ecei.tohoku.ac.jp

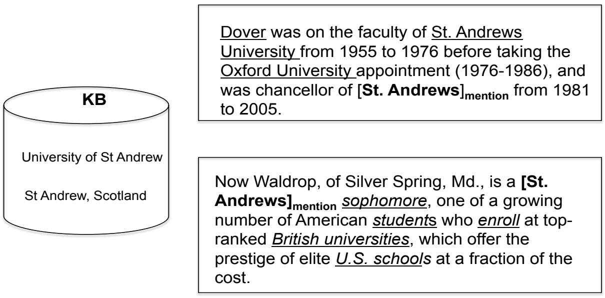
1 Introduction

In natural language processing, named entities are important components. However, due to various ways of writing, named entities have multiple surfaces in texts, e.g., 'Big Blue' and 'IBM'. Moreover, different entities can share the same surface. For example, 'New York' as a place name has dozens of different referents in Wikipedia.[1] Thus, cross-document named entity disambiguation is the task of identifying whether a mention refers to a certain entity and linking mentions in different documents to their corresponding entires in a large-scale knowledge base. Disambiguating named entities relies on context information obtained from source documents and knowledge base texts.

State-of-the-art systems [14, 13, 8] simultaneously resolve multiple entities and mostly adopt link-based methods which leverage relationships of co-occurring entities in the knowledge base while linguistic-based context information can also significantly affect disambiguation [1, 3].

For example, in Figure 1, there are two documents containing the mention 'St.Andrew'. This mention may refer to [University of St.Andrew] or [St.Andrew, Scotland]. In the first document, co-occurring entities like 'Oxford University' and 'St.Andrew University' strongly support the 'University of St.Andrew' candidate. In the second document, there is seldom co-occurring entities in texts, but linguistic information can be used. Words like 'sophomore', 'British universities', and 'U.S. schools' strongly suggest 'University of St.Andrew' as the correct entity for 'St.Andrew'.

An ideal solution is to combine global resolution method with sophisticated linguistic features. We desire to explore important linguistic features from context as the first step, which could be as the fundamental part of the future combination. Therefore, we study and compare the effects of linguistic features in a comprehensive way. Moreover, we provide salient findings according to the experiment results.



Dover was on the faculty of St. Andrews University from 1955 to 1976 before taking the Oxford University appointment (1976-1986), and was chancellor of **[St. Andrews]**$_{\textbf{mention}}$ from 1981 to 2005.

Now Waldrop, of Silver Spring, Md., is a **[St. Andrews]**$_{\textbf{mention}}$ *sophomore*, one of a growing number of American *students* who *enroll* at top-ranked *British universities*, which offer the prestige of elite *U.S. schools* at a fraction of the cost.

Figure 1: Example of documents containing mentions that refer to the same entity.

2 Related Work

According to Erbs et al. [4], features for candidate ranking could be grouped into: linguistic-based (text in source document and text extracted from the KB titles) and link-based.

Multiple linguistic-based features showed promising results in some previous studies [1, 3, 17, 6], such as document similarity, word overlapping, entity-level word overlapping, document topics, and so on. Recently, several systems [16, 11] tried to generate context (co-occurring words or entities in the external sources) for mentions, but their methods are on the word level and lack semantic information.

On the other hand, link-based features are explored by many previous work [12, 14, 7, 8, 13]. Especially, Garcia et al. [5] systemically reviewed and evaluated several state-of-the-art link-based approaches, but they did not mention linguistic-based context features.

The comparative study of linguistic-based features has been little examined. Therefore, we try to explore context information on the linguistic level in a comprehensive way, and aim at decreasing the dependence on the knowledge base.

3 Platform System

Since we want to simplify our evaluation process and only to focus on linguistic features, we build a

[1] http://en.wikipedia.org/

pipeline platform system for this task.[2] The system consists of basic components: mention detection, candidate generation, candidate ranking, NIL classification and NIL clustering. These five components are commonly required for performing Entity Discovery and Linking (EDL) [9]. We add the candidate pruning process after candidate generation to eliminate noisy candidates. In this work, since we aim at exploring the effects of different evidences in context information, we train and test on gold mention data set and start from the candidate generation phrase.

Candidate generation In the candidate generation phrase, we need a high-recall candidate list for each mention. Therefore, we first group mentions in the source document to handle misspelling, abbreviation, and partial names. For example, the candidate mentions “Gretzy” and “Wayne Gretzky” occur in the same source document, and they likely refer to the same entity. Moreover, we construct a name variation database, SurfaceSet, by extracting entity title-surface pairs from various Wikipedia sources, such as disambiguation pages, redirection pages, and anchor texts. For example, we extracted name variations like ‘Barcodes’, ‘Toon’, ‘mags’, ‘magpies’, and ‘Newcastle’ for ‘Newcastle United F.C.’, a famous England football club. SurfaceSet contains 548,084 entities and 2,080,491 surfaces. We achieved 98.43% recall on the training set.

Candidate pruning Note that the initial candidate lists are too noisy. Ranking document similarities between source documents and wiki texts is a simple and efficient way to eliminate noisy candidates. Therefore we apply Latent Semantic Index (LSI) to rank each candidate list and retain the top 50 candidates as the final candidate list. We use an off-the-shelf tool, gensim [15]. We achieved 97.28% recall on the training set. The average number of candidates per list is 41.

Candidate ranking In the candidate ranking phrase, we formulate the ranking problem similar to [1]. We represent each candidate as a feature vector, and learn to rank each candidate list. Then we select the top 1 candidate as a temporary entity label for each mention. We use SVM$^{\text{rank}}$ with the linear kernel [10].

NIL Classification and NIL Clustering NIL means mentions that do not have entries in the KB. Mentions are labeled as NIL if there is no candidate in the candidate list or the ranking score of the top 1 candidate is below a threshold. After determining NIL mentions, we group them into clusters.

4 Feature Study

We extract multiple features for candidate ranking. First, we extract basic features from mention surfaces. In order to explore linguistic information in context, we categorize those linguistic-based features into several groups.

4.1 Basic Features

We focus on the surface properties of the KB title and the mention surface. Acronym features try to capture characteristics of acronyms. For example, given a mention ‘WTO’, acronym features can detect ‘**W**orld **T**rade **O**rganization’. We also incorporate other similarity features used in previous work [6, 2], such as dice coefficient scores and jaccard index scores.

4.2 Linguistic Context Features

We extract linguistic information from both mention source documents and texts of knowledge base entries (candidates) for disambiguation.

String Appearance String appearance features are related with the appearance of a candidate title in the source document, or the appearance of mentions in candidate texts. For example, if a given mention is the family name of a person like ‘Daughtry’, the title of a candidate like ‘Chris Daughtry’ may appear in the source document. Similarly, this given mention ‘Daughtry’ may occur in the text of KB entry ‘Chris Daughtry’. Among them, a salient feature detects disambiguators in candidate titles, e.g., ‘magazine’ in People (magazine) and ‘basketball’ in Maurice Williams (basketball).

Document Similarity We use two measures to compare the text similarity between source documents and KB texts: cosine similarity with TF/IDF and dice coefficient on tokens. Since the first paragraphs of KB and text surrounding mention are supposed to be more informative, we consider to use different ranges of source documents and KB texts. We divide text in a source document into local text (window size = 50 tokens) and global text (the whole source document), and use the first paragraph and the whole KB text receptively.

Entity Mention Occurrence Named entities in mention context are more salient than common words. We capture co-occurring named entities between source documents and KB texts. For example, for a given mention ‘Obama’, the named entities ‘White House’ and ‘United States’ may appear in both the source document and the KB text if it refers to the American president ‘Barack Obama’.

Entity Fact The infobox contains important attributes of entries. For example, for entity ‘Apple

[2] We submitted the system to TAC KBP 2014 entity discovery and linking shared task.

Inc.', we can extract attributes, such as Founder (*'Steve Jobs'*) and CEO (*'Tim Cook'*). Therefore we extract fact texts from KB and check whether fact texts are in source documents. Notice that we use Wikipedia infobox here, but attributes of entities can be extracted from other KBs.

Document Topics Semantic information cannot be detected by simply counting occurrences of tokens, n-grams, and entities. Therefore we use topic models to discover the implicit topics of source documents and KB texts. We train LDA (Latent Dirichlet Allocation) model with gensim [15], which provides a fast on-line LDA model. We treat each KB entry as one document and use two different corpus for training. The first is the KBP knowledge base and the second is the latest wikidump.[3] The KBP knowledge base is a partial KB and contains about one third of Wikipedia entities. We use two similarity measures to check the topic similarity between source documents and KB entries including cosine similarity and Hellinger distance. We also generate topics of partial text surrounding mention as the local topics, to compare with using the whole source document (global topics).

Part-of Speech We hypothesize that nouns and verbs compared to other type of words could contribute more on disambiguating. Therefore we collect this two type of tokens in context and calculate cosine similarity with TF/IDF weighing respectively.

Entity Characters We use entity type matching to detect whether the KB entity type is identical to the mention entity type. For example, the mention 'St.Andrew' is an ORG (Organization) entity in the top document in Figure 1. The candidate 'University of St.Andrew' (ORG) is more correct than 'St.Andrew, Scotland' (GPE) because of entity type matching.

5 Evaluation

5.1 Experiment

We use the training data from the 2014 TAC KBP Entity Discovery and Linking (EDL) track [9]. The TAC data set consists of 5878 mentions over 158 documents. Since we focus on the ranking performance of each group of linguistic-based context features, we compute the accuracy of mentions system resolved (excluding performance on NIL clustering). In order to eliminate the effect of feature combination, we add only one feature group to the basic feature group each time. We performed 5-fold cross-validation on the training set. Table 1 shows micro-averaged accuracies of feature addition experiments.

Feature Group	Non-NIL	NIL	ALL
Basic Features	0.5946	0.6940	0.6384
String Appearance	0.5996	0.7190	0.6440
Entity Facts	0.6094	0.6754	0.6382
Entity Mention Occurrence	0.6228	0.7648	0.6856
Document Similarity	0.6612	0.7565	0.7036
Document Similarity (LOCAL)	0.6410	0.7006	0.6671
Document Similarity (GLOBAL)	0.6444	0.7730	0.7010
Document Topics	0.6412	0.6794	0.6582
Document Topics (WIKI)	0.6223	0.6765	0.6464
Document Topics (KBP)	0.6374	0.6713	0.6554
POS	0.6444	0.7182	0.6768
POS (noun)	0.6394	0.7074	0.6704
POS (verb)	0.6159	0.6943	0.6501
Type	0.5996	0.7048	0.6458
All	0.7303	0.7536	0.7410

Table 1: Feature additive test results.

In order to clarify feature effects, we divide features into more fine-grained groups, such as local topics (DT_WIKI_LOC, DT_KBP_LOC), global topics (DT_WIKI_GLO, DT_KBP_GLO), and document similarity by using the first paragraph of KB texts (DS_CON_FIR) or using the whole KB texts (DS_CON_ALL). Table 2 shows the increment of each fine-grained feature group to basic features on non-NIL mentions before NIL classification processing, and feature group names are capitalized referring to Table 1.

Fine-grained Feature Group	Accuracy Increment
C_DS_LOCAL	0.0614
C_DS_GLOBAL	0.1038
C_DS_CON_FIR	0.0003
C_DS_CON_ALL	0.056
C_DT_WIKI	0.0352
C_DT_KBP	0.0628
C_DT_WIKI_GLO	0.0392
C_DT_WIKI_LOC	0.0216
C_DT_KBP_GLO	0.0664
C_DT_KBP_LOC	0.0356
C_PS_Noun	0.064
C_PS_Verb	0.0258

Table 2: Accuracy increment on non-NIL mentions before NIL classification.

5.2 Findings

Basic features only include features related to surface similarity, which is not effective enough to find correct entities. Features based on document similarity (both words and part-of-speech levels), named entities co-occurrence, and document topics contribute the most gains.

In both document similarity and document topics group, global features are better than local features. Since we leverage measures based on bag-of-words calculation, the larger text of context contains more co-occurring words than window-size context. Although we suggest that the first paragraph in the KB is much informative, using the whole KB text ('DS_CON_ALL') is much better than only using the first paragraph ('DS_CON_FIR'). We found that, in the KBP KB, several first paragraphs of KB texts are

[3] http://dumps.wikimedia.org/enwiki/20140707/enwiki20140707-pages-articles.xml.bz2

very short, sometimes only one sentence. For example, for 'Jeff Perry (American actor)', there is only one sentence like "Jeff Perry (born August 16, 1955 in Highland Park, Illinois) is an American character actor."

Moreover, based on the results in Table 2, the increment of global topics is more than that of local topics by 0.03 (KBP corpus). Since the distribution of partial document topics is inconsistent with document topics, global topics can better represent the semantic context of a mention.

Although the KBP KB contains around one third entities of Wikipedia, the performance on the KBP KB corpus is better because we use the KBP KB as entities database. We found that words of KBP KB topics could represent source document better than using the Wikipedia corpus for some entities. For example, 'Salvador Dali' entity is a painter, who is also known for writing and film. Words of top topics given by the KBP LDA corpus of this entity are 'film', 'book', 'album', 'play', and so on. However, words given by the Wikipedia LDA corpus are 'Louisiana', 'disease', 'species', and so on. The Wikipedia LDA corpus is not well-built, which may also affect the performance, because we follow an off-the-shelf training process.[4]

In addition, we found that nouns and verbs are more informative than other type of words. Nouns contain more information than verb words because named entities are more salient.

6 Conclusion and Future Work

We comprehensive study the effects of different linguistic-based features based on a platform system. According to the evaluation results, we find several useful features, such as document similarity, co-occurring entities overlap, and document topics. Moreover, global topics are more effective than local topics for representing implicit semantic information of mentions. In addition, nouns are quite effective than verbs on disambiguation.

We also compare the performance of current features with systems from the 2014 EDL Diagnostic task [9]. The accuracy on all mentions of our current system could beat the median system by 0.5, but we still have a huge gap with the best system.

In future work, we plan to combine linguistic features with link-based methods to further improve our system.

References

[1] R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, volume 6, pages 9–16, 2006.

[2] L. Dietz and J. Dalton. Acrossdocument neighborhood expansion: Umass at tac kbp 2012 entity linking. In *Proceedings of TAC 2012*, 2012.

[3] M. Dredze, P. McNamee, D. Rao, A. Gerber, and T. Finin. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 277–285, 2010.

[4] N. Erbs, T. Zesch, and I. Gurevych. Link discovery: A comprehensive analysis. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 83–86. IEEE, 2011.

[5] N. F. Garcıa, J. A. Fisteus, and L. S. Fernández. Comparative evaluation of link-based approaches for candidate ranking in link-to-wikipedia systems. *Journal of Artificial Intelligence Research*, 49:733–773, 2014.

[6] D. Graus, T. Kenter, M. Bron, E. Meij, and M. De Rijke. Context-based entity linking–university of amsterdam at tac 2012. 2012.

[7] Y. Guo, G. Tang, W. Che, T. Liu, and S. Li. Hit approaches to entity linking at tac 2011. In *Proceedings of the Fourth Text Analysis Conference (TAC 2011)*, 2011.

[8] 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 *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792, 2011.

[9] H. Ji, H. Dang, J. Nothman, and B. Hachey. Overview of tac-kbp2014 entity discovery and linking tasks. In *Proceedings of Text Analysis Conference*, 2014.

[10] T. Joachims. Training linear svms in linear time. In *Proceedings of KDD*, pages 217–226, 2006.

[11] Y. Li, C. Wang, F. Han, J. Han, D. Roth, and X. Yan. Mining evidences for named entity disambiguation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1070–1078, 2013.

[12] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.

[13] A. Moro, A. Raganato, and R. Navigli. Entity linking meets word sense disambiguation: A unified approach. In *Transactions of the Association for Computational Linguistics*, volume 2, 2014.

[14] L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1375–1384, 2011.

[15] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC: Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.

[16] V. Stoyanov, J. Mayfield, T. Xu, D. W. Oard, D. Lawrie, T. Oates, and T. Finin. A context-aware approach to entity linking. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 62–67, 2012.

[17] W. Zhang, Y. C. Sim, J. Su, and C. L. Tan. Entity linking with effective acronym expansion, instance selection, and topic modeling. In *Proceedings of IJCAI*, volume 2011, pages 1909–1914, 2011.

[4] https://radimrehurek.com/gensim/wiki.html