Short Text Entity Linking Using Multiple Knowledge Bases: A Case Study of Wikipedia and Freebase Postprint

Authors: Zhou Pengcheng, Wuchuan, Lu Wei

Date: 2017-10-11T00:00:00+00:00

Abstract

【Purpose】To perform entity linking based on multiple knowledge bases, addressing the low coverage problem of entity linking based on a single knowledge base. 【Method】First, n-grams are generated from the text and candidate mentions are extracted using part-of-speech tags and multiple mention-entity dictionaries; mention combinations are then generated, retaining only those with maximal coverage that are not contained by other combinations; next, candidate entity sequences are generated and the relevance of each sequence is computed using information from multiple knowledge bases; finally, the entity sequence with maximal relevance is selected as the final result. 【Results】Experimental results using Wikipedia and Freebase as examples demonstrate that entity linking based on Wikipedia+Freebase achieves precision, recall, and F-score of 71.81%, 76.86%, and 74.25%, respectively. 【Limitations】Filtering n-grams based on part-of-speech lacks theoretical justification, and the FACC1 dataset is characterized by high precision but low recall. 【Conclusion】Leveraging entity information from multiple knowledge bases can enhance entity linking effectiveness.

Full Text

Preamble

Entity Linking Method for Short Texts with Multi-Knowledge Bases: A Case Study of Wikipedia and Freebase

Zhou Pengcheng[1], Wu Chuan[1], Lu Wei[1,2]
[1](School of Information Management, Wuhan University, Wuhan 430072, China)
[2](Center for the Studies of Information Resources, Wuhan University, Wuhan 430072, China)

Abstract

Objective: This paper proposes an entity linking method using multiple knowledge bases to address the low coverage problem associated with entity linking based on a single knowledge base. **Methods:** First, we generate n-grams from the input text and obtain candidate mentions using part-of-speech tagging and multiple mention-entity dictionaries. Next, we generate mention combinations and retain those with maximum coverage that are not contained by other combinations. We then generate candidate entity sequences and compute their relevance using information from multiple knowledge bases. Finally, we select the entity sequence with the highest relevance as the final result. **Results:** Using Wikipedia and Freebase as case studies, the experimental results demonstrate that entity linking based on Wikipedia+Freebase achieves precision, recall, and F-value of 71.81%, 76.86%, and 74.25%, respectively. **Limitations:** Filtering n-grams based on part-of-speech tagging lacks theoretical justification, and the FACC1 dataset exhibits characteristics of high precision but low recall. **Conclusions:** Utilizing entity information from multiple knowledge bases improves entity linking performance.

Keywords: Entity linking; Knowledge base; Wikipedia; Freebase

1. Introduction

An entity is an objectively existing and distinguishable object in the real world, encompassing both concrete things (e.g., person names, locations, organization names) and abstract concepts (e.g., concepts, relationships). Entity linking refers to the process of linking text fragments representing entities in a document, known as entity mentions (or simply mentions), to corresponding entries in a specific knowledge base (KB). This process is sometimes called named entity linking [1].

Entities are ubiquitous across various types of text. When encountering unknown entities, entity linking technology enriches the original text with semantic information from relevant knowledge base entries, helping readers deepen their understanding of the entity and enabling both humans and computers to better comprehend and process the text. Entity linking research has attracted significant attention due to its importance, with several international evaluation conferences introducing related tasks, such as the "Link the Wiki" task at INEX 2007, the "Knowledge Base Population" task at TAC 2009, and the "Knowledge Base Acceleration" task at TREC 2012. Entity linking demonstrates promising applications in information retrieval [2], knowledge base construction [3], and question answering systems [4].

The primary challenges in entity linking are polysemy and synonymy. Synonymy occurs when an entity has multiple mentions—its standard name, aliases, and abbreviations can all refer to the same entity. For example, "Michael Jordan," "MJ," and "Jordan" can all refer to the entity Michael Jeffrey Jordan. Polysemy occurs when a single mention can refer to multiple entities; for

instance, "MJ" may refer to Michael Jeffrey Jordan or Michael I. Jackson. Resolving polysemy requires leveraging entity information from knowledge bases for disambiguation. Since single knowledge bases contain relatively limited entity information, we hypothesize that utilizing entity information from multiple knowledge bases could better address the polysemy problem.

Knowledge bases serve as the foundation for entity linking research. Common knowledge bases include Wikipedia, Freebase [5], YAGO [6], and DBpedia [7]. Wikipedia is the most frequently used knowledge base in entity linking research, rich in textual semantic information where each entity page describes a specific entity. Freebase is also commonly used, offering more structured entity information compared to Wikipedia. In 2013, Google released the Freebase entity annotation dataset FACC1 [8], which has been applied in information retrieval [9]. FACC1 provides entity annotations for ClueWeb09 and ClueWeb12, enabling the statistical analysis of entity popularity.

This paper proposes a multi-knowledge-base entity linking method that leverages multiple mention-entity dictionaries for mention recognition and utilizes entity information from multiple knowledge bases for disambiguation, aiming to solve the low coverage problem associated with single-knowledge-base entity linking.

Entity linking comprises two steps: mention recognition and entity disambiguation [10]. Although some studies [11] adopt slightly different divisions, the essence remains the same. Traditional entity linking has primarily focused on long documents, but recent researchers [9,12-14] have begun investigating short text entity linking for microblogs, search queries, etc., with applications already emerging in information retrieval [9]. The main distinction is that short texts contain limited contextual information, making entity disambiguation more challenging. Additionally, short texts often exhibit non-standard writing conventions, such as missing capitalization and punctuation [9] and spelling errors [14], which further complicate mention recognition. Therefore, we argue that short text entity linking presents greater challenges and deserves more attention.

2.1 Long Document Entity Linking

The first step in entity linking is mention recognition, which requires constructing a mention-entity dictionary. Most researchers extract page titles from Wikipedia entity pages, disambiguation pages, and redirect pages as entity mentions to build such dictionaries. Alternative approaches exist, such as Sil et al. [15] extracting standard names and aliases of entities from Freebase. Mentions are then identified according to specific rules; for example, Cucerzan [11] uses capitalization rules and prior statistical information for mention recognition, selecting the entity sequence that maximizes consistency between entity context and Wikipedia homepage text as well as among candidate entities. Mihalcea et al. [16] utilize linking probability for mention recognition and combine knowledge engineering methods with naive Bayes classification to determine the

final entity sequence.

Since a mention may refer to multiple entities, methods are needed to determine the correct entity, i.e., entity disambiguation. Current approaches include machine learning [17-18], learning to rank [19-21], graph models [22-25], unsupervised methods [11,26], and ensemble methods [27-28]. Zhang et al. [17] adopt the method from [11] to construct a mention dictionary for mention recognition. If the candidate mention set is empty, they supplement it using Wikipedia's "Did You Mean"and search engine features. They frame entity disambiguation as a binary classification problem where mention-entity pairs formed by mentions and their referred entities are positive examples, while pairs with other candidate entities are negative examples. They use lexical features, word-category features, and entity types to train an SVM classifier. If multiple candidate entities are classified as positive, they compute mention-entity similarity using bag-of-words and entity co-occurrence features, selecting the candidate with highest similarity. Ratinov et al. [20] assume mentions are given and propose two feature types: local features (cosine similarity between mention context and entity homepage text, mention document and entity homepage text, mention context and entity context, etc.) and global features (normalized Google distance, pointwise mutual information for entity category similarity, in-link and out-link similarities), training a Rank SVM model to select the highest-ranked entity as the mention' s referent in context. Han et al. [22] also focus solely on entity disambiguation, constructing a mention-entity and entity-entity relationship graph with mentions and candidate entities as nodes, using a PageRank-like mechanism to identify entities.

2.2 Short Text Entity Linking

Ferragina et al. [13] were among the first to address short text entity linking. They construct a mention dictionary using the method from [11], filter it with manual rules, and identify candidate mentions using this dictionary. They then employ both machine learning and manual rule-based methods for disambiguation, leveraging features such as prior probability of mentions referring to entities and relevance among candidate entities. Meij et al. [14] aim to obtain as many candidate mentions as possible, proposing four feature categories: n-gram features, concept features, n-gram-concept features, and tweet features, using machine learning to identify concepts and link them to corresponding Wikipedia pages. Liu et al. [29] build upon Meij et al.' s work by incorporating mention-mention features and selecting the entity sequence with the highest similarity score.

We observe that current entity linking research relies on single knowledge bases. However, since certain entities exist only in specific knowledge bases, a single knowledge base may not fully cover all entities in a document. Moreover, single knowledge bases provide relatively limited entity information, which affects disambiguation performance. To address these issues, this paper proposes a multi-knowledge-base entity linking method that effectively leverages entity in-

formation from multiple knowledge bases and simultaneously performs entity linking across them.

3. Methodology

3.1 Problem Definition

Entity linking involves identifying mentions in a given text, determining the entities they refer to through disambiguation, and linking them to corresponding entries in specific knowledge bases.

The formal definition is as follows: Input is a text consisting of n words M = (w_1, w_2, ⋯, w). Output is a mention combination and its corresponding entity sequence E = (e_1, e_2, ⋯, e), where e represents an entry in a specific knowledge base. If |M| = 1, then the output is the set of possible entities that the mention may correspond to, denoted as Set e_1, e_2, ⋯, e .

3.2 Entity Linking Method

[Figure 1: see original paper] illustrates the steps of multi-knowledge-base entity linking, which consists of offline and online phases. The offline phase builds mention-entity dictionaries and entity mapping dictionaries. The online phase comprises the main steps of the entity linking method, including n-gram generation, candidate mention identification, mention combination generation, entity sequence generation, and entity relevance calculation.

(1) Dictionary Construction

For mention recognition, we collect standard names, aliases, and other information from knowledge bases as entity mentions, preprocess them, and construct mention-entity dictionaries. Each dictionary contains two fields: a mention field and an entity field, stored in the format "m → e_1(e_{count}$_1$) e_2(e_{count}$_2$) ⋯" , where m represents an entity mention and e (e_{count}) indicates that mention m may refer to entity e with occurrence count e_{count} , which is used to calculate the prior probability of candidate entities.

To leverage multi-knowledge-base information simultaneously, we construct entity mapping dictionaries following specific methods. The mapping dictionary contains n fields, stored in the format "e_1 → e_2 → ⋯→ e " , where e represents an entity from knowledge base i, and e_1, e_2, ⋯, e correspond to the same entity across different knowledge bases.

(2) Method Steps

N-gram Generation: Generate n-grams from the input short text. For example, the short text "obama family tree" yields six n-grams: {obama, family, tree, obama family, family tree, obama family tree}.

Candidate Mention Identification: For each generated n-gram, search directly across the mention fields of multiple mention-entity dictionaries. If

a record exists in any dictionary, the n-gram is considered a potential entity mention. N-grams containing no nouns are filtered out, as entities typically appear as nouns. For instance, "obama" exists in the mention field with noun part-of-speech, so it qualifies as an entity mention, as do "family," "tree," "obama family," and "family tree." However, "obama family tree" is filtered out as it lacks a corresponding record. Thus, "obama family tree" retains five potential entity mentions: {obama, family, tree, obama family, family tree}.

Mention Combination Generation: Candidate mentions from the previous stage may overlap. Some researchers [30] address this using a left-to-right longest matching strategy, but we argue this may cause recognition errors. Our approach generates candidate mention combinations through three steps: 1) Select combinations of non-overlapping candidate mentions; 2) Retain combinations with maximum coverage; 3) Retain combinations where at least one mention is not contained by other combinations. Here, "containment" means one mention is either part of another or identical to it. For example, "obama family tree" retains two mention combinations: {obama + family tree} and {obama family + tree}.

Entity Sequence Generation and Relevance Calculation: If n mention combinations are retained in the previous stage, with the i-th combination containing n mentions, we merge candidate entity records from all knowledge bases for each mention and select the top k entities with highest prior probability as candidates, generating $_1{}^n$ k^n entity sequences. We assume co-occurring entities are relevant, so we compute relevance scores for each entity sequence and return the highest-scoring sequence as the final result, as shown in formula (1):

$$\arg\max_{E\in \text{Set}} \sum_{i=1}^{l} \alpha \cdot h(\vec{a}, m_i, e_i) + \beta \cdot f(\vec{b}, e_i, e_j)$$

where Set represents all possible entity sequences; h(a, m , e) denotes the relevance function between mention m and its candidate entity e with weight vector a; f(b, e , e) denotes the relevance function between entities with weight vector b; α and β balance the two relevance functions with α, β (0,1) and $\alpha + \beta = 1$.

4. Implementation Details

Since Wikipedia and Freebase are widely used and representative in entity linking research, we conduct experiments based on Wikipedia, Freebase, and Wikipedia+Freebase separately.

4.1 Wikipedia-Based Entity Linking

(1) Mention-Entity Dictionary Construction: Following Bunescu et al. [31], we extract page titles from Wikipedia entity pages, disambiguation

pages, and redirect pages, as well as anchor texts from entity homepages. After preprocessing (lowercasing, etc.), we construct the mention-entity dictionary and use anchor text statistics to record mention-to-entity occurrence counts.

The prior probability of candidate entities is crucial disambiguation information used in many studies [23,29,31]. We calculate it using formula (2):

$$\text{Prior}(e_i|m) = \frac{\text{count}(m, e_i)}{\sum_{j=1}^{l} \text{count}(m, e_j)}$$

where count(m, e) represents the number of times mention m links to entity e in Wikipedia homepage texts.

String Similarity: Higher similarity between a mention and a candidate entity' s standard name indicates greater probability that the mention refers to that entity. Edit Distance measures string similarity as the minimum number of editing operations required to transform one string into another; smaller distance indicates higher similarity. We compute similarity using formula (3):

$$h(m_i, \text{CN}(e_i)) = 1 - \frac{\text{ED}(m_i, \text{CN}(e_i))}{\max\{\text{length}(m_i), \text{length}(\text{CN}(e_i))\}}$$

where CN(e) denotes entity e 's standard name (the Wikipedia entity page title), ED(m , CN(e)) is the edit distance, and max{length(m), length(CN(e))} is the length of the longer string.

Entity-Entity Features:

1) Textual Relatedness: Related entities likely share similar description texts. We preprocess Wikipedia entity homepage texts by lowercasing, removing special characters, and eliminating stopwords, then compute textual relatedness using formula (4):

$$f_{\text{text}}(e_i, e_j) = \frac{\sum_{k=1}^{l} w_{ik} \cdot w_{jk}}{\sqrt{\sum_{k=1}^{l} w_{ik}^2} \cdot \sqrt{\sum_{k=1}^{l} w_{jk}^2}}$$

where w represents the weight of the k-th word in entity e ' s document (using term frequency as weight). For entity sequences with more than two entities, we use the arithmetic mean of pairwise textual relatedness values.

2) Related Entity Overlap: Related entities often share common related entities. Wikipedia entity pages contain links to other entity pages, which we use to collect related entity sets. Wikipedia entities have three related entity

types: in-link related entities (e appears in e ' s page but not vice versa), out-link related entities (e appears in e ' s page but not vice versa), and mutual-link related entities (bidirectional links). We use Jaccard coefficient to measure related entity overlap, as shown in formula (5):

$$\mathrm{reo}_{\mathrm{out}}(e_i, e_j) = \frac{|\mathrm{Set}^i_{\mathrm{out}} \cap \mathrm{Set}^j_{\mathrm{out}}|}{|\mathrm{Set}^i_{\mathrm{out}} \cup \mathrm{Set}^j_{\mathrm{out}}|}$$

where Set and Set represent out-link related entity sets. Similar formulas apply for in-link and mutual-link relatedness, with the final related entity relatedness being a weighted average of the three types.

3) Category Relatedness: Related entities often share categories. Wikipedia editors assign categories to entities, which we extract from homepage information. We again use Jaccard coefficient for category relatedness:

$$f_{\mathrm{cat}}(e_i, e_j) = \frac{|\mathrm{cSet}_i \cap \mathrm{cSet}_j|}{|\mathrm{cSet}_i \cup \mathrm{cSet}_j|}$$

where cSet and cSet are category sets for entities e and e .

4.2 Freebase-Based Entity Linking

(1) Mention-Entity Dictionary Construction: We extract standard names and aliases from Freebase to build the mention-entity dictionary. Freebase' s highly structured entity information includes dedicated fields for entity properties, allowing direct extraction from name and alias fields. We apply the same lowercasing and special character removal, and use the ClueWeb09 Freebase annotation dataset FACC1 [8] to compute mention-to-entity occurrence counts. For example:

$$\text{baldwin\_vi} \rightarrow \text{/m/0129jf(3) /m/01\_dt9(6)}$$

where baldwin_{vi} is the mention, /m/0129jf and /m/01_{dt9} are possible entities (represented by Freebase IDs), with occurrence counts 3 and 6, respectively. Prior probability is calculated using formula (2).

(2) Freebase Entity Features:

Mention-Entity Features: We use the same prior probability and string similarity features as for Wikipedia entities. Note that Freebase entity standard names are extracted from the name field.

Entity-Entity Features:

1) Textual Relatedness: Freebase entity textual relatedness uses text from the Description field, with the same preprocessing and calculation as Wikipedia (formula 4).

2) Type Relatedness: Similar to Wikipedia categories, related entities likely share types. Freebase editors assign hierarchical types to entities (e.g., Barack Obama has 97 types including /people/person, /government/politician). We perform simple term frequency statistics on hierarchical type names, using frequency as weight, and compute type relatedness using formula (4).

4.3 Wikipedia+Freebase Entity Linking

Wikipedia+Freebase entity linking uses the mention-entity dictionaries from Sections 4.1 and 4.2 for candidate mention identification. Since Wikipedia and Freebase contain complementary entity information, we extract different features from each knowledge base. To leverage both, we construct a Wikipedia-Freebase entity mapping dictionary.

(1) Wikipedia-Freebase Entity Mapping Dictionary: Freebase entity pages contain an Equivalent Webpage field with links to corresponding entities in other knowledge bases. We extract equivalent Wikipedia page titles to establish one-to-one mappings, e.g.:

$$/\text{m}/03\text{kkbz} \rightarrow 873558 \rightarrow \text{Ivan\_Bella}$$

where /m/03kkbz is the Freebase ID, Ivan_{Bella} is the equivalent Wikipedia page title, and 873558 is the Wikipedia entity ID.

(2) Wikipedia+Freebase Entity Features: For mention-entity features, we use the arithmetic mean of Wikipedia and Freebase prior probabilities and string similarities. For entity-entity features, we fuse Wikipedia' s textual, related entity, and category relatedness with Freebase' s type relatedness using the mapping dictionary. Definitions follow Sections 4.1 and 4.2.

5. Experiments

5.1 Dataset and Experimental Setup

The experimental input consists of 200 query topics from the TREC Web Track tasks (2009-2012). We downloaded the December 2, 2013 Wikipedia dump containing over 4,450,000 entity pages, disambiguation pages, redirect pages, page link relationships, and category information. Using Java, we processed raw files, extracted entity mentions, computed prior probabilities, and built a Lucene index for mention search, extracting over 14,870,000 mentions from Wikipedia. We also downloaded the July 6, 2014 Freebase RDF file containing aliases, equivalent pages, descriptions, and other properties for over 42,660,000 entities, extracting over 22,130,000 mentions using the same tools.

Since Freebase contains mappings to Wikipedia, we extracted these relationships and built corresponding indexes with Lucene. After removing query topics that were themselves entity mentions (which cannot be disambiguated from context), we processed 179 query topics with 242 annotated entities.

5.2 Evaluation Metrics

We evaluate performance using precision, recall, and F-value:

$$\text{Precision} = \frac{|\text{Set}_R \cap \text{Set}_L|}{|\text{Set}_R|}$$

$$\text{Recall} = \frac{|\text{Set}_R \cap \text{Set}_L|}{|\text{Set}_L|}$$

$$\text{F-value} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where Set_R is the set of entities identified by our method, Set_L is the set of manually annotated entities, and $|\cdot|$ denotes set cardinality. Precision measures the proportion of correctly identified entities among all identified entities, recall measures the proportion of correctly identified entities among all annotated entities, and F-value is the harmonic mean of precision and recall.

5.3 Experimental Results

As shown in Table 1, Wikipedia-based entity linking achieves 62.68% precision, 71.49% recall, and 66.80% F-value. Freebase-based entity linking achieves 69.32% precision, 75.62% recall, and 72.33% F-value. Wikipedia+Freebase-based entity linking achieves 71.81% precision, 76.86% recall, and 74.25% F-value.

Evaluation results of entity linking based on different knowledge bases and combined knowledge bases

Method	Precision	Recall	F-value
Wikipedia	62.68%	71.49%	66.80%
Freebase	69.32%	75.62%	72.33%
Wikipedia+Freebase	**71.81%** (+14.57%) (+3.59%)	**76.86%** (+7.51%) (+1.64%)	**74.25%** (+11.15%) (+2.65%)

Note: Best results are bolded. Parentheses show improvement of Wikipedia+Freebase over Wikipedia or Freebase alone.

The results demonstrate that Wikipedia+Freebase outperforms individual knowledge bases, with precision improving by 14.57% and 3.59%, recall by 7.51% and 1.64%, and F-value by 11.15% and 2.65% compared to Wikipedia and Freebase, respectively, proving the effectiveness of the multi-knowledge-base approach.

Figures 2 and 3 show recall and precision for 15 query topics across the three experiments. Some topics perform well with Wikipedia (e.g., "pacific northwest laboratory," "arkadelphia health club"), others with Freebase (e.g., "condos in florida,""uss yorktown charleston sc"), while all 15 topics show good performance with Wikipedia+Freebase.

[Figure 2: see original paper] Recall of 15 query topics in three experiments
[Figure 3: see original paper] Precision of 15 query topics in three experiments

To analyze the reasons for improvement, we conducted four additional experiments (Table 2). Wikipedia-MF uses only the Wikipedia mention-entity dictionary for mention recognition but both Wikipedia and Freebase features for disambiguation, verifying the impact of multi-features. Wikipedia-MD uses both dictionaries for mention recognition but only Wikipedia features for disambiguation, verifying the impact of multiple mention-entity dictionaries. Freebase-MF and Freebase-MD are analogous.

Evaluation results of different knowledge bases and supplementary experiments

Method	Precision	Recall	F-value
Wikipedia	62.68%	71.49%	66.80%
Wikipedia-MF	63.26% (+0.93%)	69.01% (-3.5%)	66.01% (-1.2%)
Wikipedia-MD	71.43% (+13.96%)	76.45% (+6.94%)	73.85% (+10.55%)
Freebase	69.32%	75.62%	72.33%
Freebase-MF	69.47% (+0.22%)	75.21% (-0.54%)	72.22% (-0.15%)
Freebase-MD	69.26% (-0.09%)	77.27% (+2.2%)	73.05% (+1%)

Note: Wikipedia-MF = Wikipedia dictionary + both KB features; Wikipedia-MD = both dictionaries + Wikipedia features; Freebase-MF/MD are analogous.

The results show that using multiple mention-entity dictionaries is the primary reason for performance improvement, increasing recall by 6.94% and 2.2% and F-value by 10.55% and 1% for Wikipedia and Freebase, respectively. Multi-features provide only minor precision improvements (0.93% and 0.22%) without overall performance gains.

Error analysis of incorrectly identified query topics reveals three issues: (1) Correct mentions may be filtered during candidate mention identification. For example, in "old coins," the correct mention "coins" is filtered because "old coins" has greater coverage and matches the longest-match principle, though "old coins" appears only once as a mention in Wikipedia. Similar errors occur

for "espn sports" and "diabetes education." (2) Correct entities may not be retrieved when obtaining candidate entities. For instance, for "website design hosting," when retrieving candidates for "hosting," the top k entities by prior probability from Wikipedia do not include the correct entity, though Freebase' s top k does, with similar errors occurring for "lymphoma in dogs" and "fact on uranus." (3) Incorrect disambiguation, as seen with "obama family tree," where the combined approach fails to disambiguate correctly, suggesting that additional features might resolve such cases.

6. Conclusion

This paper proposes a multi-knowledge-base entity linking method. Experiments with Wikipedia and Freebase demonstrate that Wikipedia+Freebase outperforms individual knowledge bases. Two limitations remain: filtering n-grams by part-of-speech lacks theoretical foundation, and the FACC1 dataset has high precision but low recall [8]. Our method is applicable to other knowledge bases. For YAGO, the "HasWikipediaURL"relation can build YAGO-Wikipedia-Freebase mapping dictionaries; the "means" relation can collect YAGO entity mentions, while Wikipedia anchor texts can provide mention-to-entity counts [23] to construct similar mention-entity dictionaries. Disambiguation can use textual relatedness and type relatedness (computed from type and subClass relationships [23]). Based on our findings, Wikipedia+Freebase+YAGO should outperform individual knowledge bases.

Future work will explore better information fusion methods to further improve multi-knowledge-base entity linking performance.

References

[1] Zhang W, Sim Y C, Su J, et al. Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain. 2011: 1909-1914.

[2] Pantel P, Fuxman A. Jigs and Lures: Associating Web Queries with Structured Entities. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA. 2011: 83-92.

[3] Lin T, Etzioni O. Entity Linking at Web Scale. In: Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, Montreal, Canada. 2012: 84-88.

[4] Welty C, Murdock J W, Kalyanpur A, et al. A Comparison of Hard Filters and Soft Evidence for Answer Typing in Watson. In: Proceedings of the 11th International Conference on the Semantic Web. Springer-Verlag, 2012: 243-256.

[5] Bollacker K, Evans C, Paritosh P, et al. Freebase: A Collaboratively Created

Graph Database for Structuring Human Knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. ACM, 2008: 1247-1250.

[6] Suchanek F M, Kasneci G, Weikum G. YAGO: A Core of Semantic Knowledge. In: Proceedings of the 16th International Conference on World Wide Web. ACM, 2007: 697-706.

[7] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A Nucleus for a Web of Open Data. In: Proceedings of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, Busan, Korea. 2007: 722-735.

[8] ClueWeb09 Related Data: Freebase Annotations of the ClueWeb Corpora, v1 (FACC1). (2013-11-04). http://lemurproject.org/clueweb09/FACC1/.

[9] Brandão W C, Santos R L T, Ziviani N, et al. Learning to Expand Queries Using Entities. Journal of the Association for Information Science and Technology, 2014, 65(9): 1910-1925.

[10] Lu Wei, Wu Chuan. Literature Review on Entity Linking. Journal of the China Society for Scientific and Technical Information, 2015, 34(1): 105-112.

[11] Cucerzan S. Large-scale Named Entity Disambiguation Based on Wikipedia Data. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007: 708-716.

[12] Milne D, Witten I H. Learning to Link with Wikipedia. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. ACM, 2008: 509-518.

[13] Ferragina P, Scaiella U. Tagme: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities). In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, Ontario, Canada. 2010: 1625-1628.

[14] Meij E, Weerkamp W, De Rijke M. Adding Semantics to Microblog Posts. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining. ACM, 2012: 563-572.

[15] Sil A, Yates A. Re-ranking for Joint Named-entity Recognition and Linking. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, San Francisco, California, USA. 2013: 2369-2374.

[16] Mihalcea R, Csomai A. Wikify!: Linking Documents to Encyclopedic Knowledge. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management, Lisboa, Portugal. 2007: 233-242.

[17] Zhang W, Su J, Tan C L, et al. Entity Linking Leveraging Automatically Generated Annotation. In: Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, Beijing, China. 2010: 1290-1298.

[18] Pilz A, Paaß G. From Names to Entities Using Thematic Context Distance. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Glasgow, Scotland, UK. 2011: 857-866.

[19] Zheng Z, Li F, Huang M, et al. Learning to Link Entities with Knowledge Base. In: Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 483-491.

[20] Ratinov L, Roth D, Downey D, et al. Local and Global Algorithms for Disambiguation to Wikipedia. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011: 1375-1384.

[21] Shen W, Wang J, Luo P, et al. LINDEN: Linking Named Entities with Knowledge Base via Semantic Knowledge. In: Proceedings of the 21st International Conference on World Wide Web, Lyon, France. 2012: 449-458.

[22] Han X, Sun L, Zhao J. Collective Entity Linking in Web Text: A Graph-based Method. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, China. 2011: 765-774.

[23] Hoffart J, Yosef M A, Bordino I, et al. Robust Disambiguation of Named Entities in Text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 782-792.

[24] Hachey B, Radford W, Curran J. Graph-Based Named Entity Linking with Wikipedia. In: Proceedings of the 12th International Conference on Web Information System Engineering. 2011: 213-226.

[25] Guo Y, Che W, Liu T, et al. A Graph-based Method for Entity Linking. In: Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand. 2011: 1010-1018.

[26] Gottipati S, Jiang J. Linking Entities to a Knowledge Base with Query Expansion. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 804-813.

[27] Zhang W, Sim Y C, Su J, et al. NUS-I2R: Learning a Combined System for Entity Linking. In: Proceedings of Text Analysis Conference 2010 Workshop, Gaithersburg, Maryland, USA. 2010.

[28] Chen Z, Ji H. Collaborative Ranking: A Case Study on Entity Linking. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Scotland, UK. 2011: 771-781.

[29] Liu X, Li Y, Wu H, et al. Entity Linking for Tweets. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.

Association for Computational Linguistics, 2013.

[30] Wu C, Lu W, Zhou P. An Optimization Framework for Entity Recognition and Disambiguation. In: Proceedings of the International Workshop on Entity Recognition & Disambiguation. ACM, 2014: 105-110.

[31] Bunescu R C, Pasca M. Using Encyclopedic Knowledge for Named Entity Disambiguation. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy. 2006: 9-16.

Author Contributions

Zhou Pengcheng: Conceptual design, experimental implementation, manuscript drafting and revision.
Wu Chuan: Conceptual design and revision, multiple manuscript revisions.
Lu Wei: Supervised conceptual design and writing, multiple version and final manuscript revisions.

Conflict of Interest

All authors declare no conflict of interest.

Supporting Data

Supporting data is available in the journal's online version at http://www.infotech.ac.cn:
[1] Zhou Pengcheng, Wu Chuan, Lu Wei. Annotation data.xml. Entity annotation data for topics.
[2] Zhou Pengcheng, Wu Chuan, Lu Wei. freebase recognition results.txt. Freebase-based entity recognition results.
[3] Zhou Pengcheng, Wu Chuan, Lu Wei. wikipedia recognition results.txt. Wikipedia-based entity recognition results.
[4] Zhou Pengcheng, Wu Chuan, Lu Wei. wikipedia_{freebase} recognition results.txt. Wikipedia+Freebase-based entity recognition results.
[5] Zhou Pengcheng, Wu Chuan, Lu Wei. freebase mention-entity dictionary.txt. Mention-entity dictionary built from Freebase.
[6] Zhou Pengcheng, Wu Chuan, Lu Wei. wikipedia mention-entity dictionary.txt. Mention-entity dictionary built from Wikipedia.
[7] Zhou Pengcheng, Wu Chuan, Lu Wei. freebase entity-wikipedia entity mapping dictionary.txt. Mapping dictionary between Freebase and Wikipedia entities.

Received: 2016-01-13
Revised: 2016-03-20

Note: Figure translations are in progress. See original paper for figures.