Distant Supervision for Entity Linking

Miao Fan*, **Qiang Zhou and Thomas Fang Zheng**
CSLT, Division of Technical Innovation and Development
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology,
Tsinghua University, Beijing, 100084, China
`*fanmiao.cslt.thu@gmail.com`

Abstract

Entity linking is an indispensable operation of populating knowledge repositories for information extraction. It studies on aligning a textual entity mention to its corresponding disambiguated entry in a knowledge repository. In this paper, we propose a new paradigm named distantly supervised entity linking (DSEL), in the sense that the disambiguated entities that belong to a huge knowledge repository (Freebase) are automatically aligned to the corresponding descriptive webpages (Wiki pages). In this way, a large scale of weakly labeled data can be generated without manual annotation and fed to a classifier for linking more newly discovered entities. Compared with traditional paradigms based on solo knowledge base, DSEL benefits more via jointly leveraging the respective advantages of Freebase and Wikipedia. Specifically, the proposed paradigm facilitates bridging the disambiguated labels (Freebase) of entities and their textual descriptions (Wikipedia) for Web-scale entities. Experiments conducted on a dataset of 140,000 items and 60,000 features achieve a baseline F1-measure of 0.517. Furthermore, we analyze the feature performance and improve the F1-measure to 0.545.

1 Introduction

To build the "Digital Alexandria Library" for our human race, researchers in the NLP community have dedicated themselves to *Information Extraction* (Sarawagi, 2008) over the past decades. Information extraction focuses on processing natural language text to produce structured knowledge, which is usually represented as triples (two entities and their relation) for the convenience of storage in a database, retrieval, or even automatic reasoning. For example, if we send a natural language sentence, *Michael Jordan visited CMU yesterday*, to the pipeline of information extraction machine, it will be processed by three operations in advance, i.e.,

- **Named Entity Recognition** (Nadeau and Sekine, 2007): Entities should firstly be identified and classified into predefined categories, such as person (PER), location (LOC) and organization (ORG). The sentence will be annotated as *[Michael Jordan]/PER visited [CMU]/ORG yesterday*, after being processed by this operation.

- **Coreference Resolution** (Ng, 2010): Some entities may have alias or abbreviations. It is well known that *CMU* is the abbreviation for *Carnegie Mellon University*. The knowledge repository may only store the regularized name, e.g., *Carnegie Mellon University*, for this named entity, so coreference resolution is indeed necessary.

- **Relation Extraction** (Bach and Badaskar, 2007): After both of the named entities (*[Michael Jordan]/PER* and *[Carnegie Mellon University]/ORG*) are recognized and regularized, we begin to study on the relation between them. In this case, we extract the verb *visited* and map it to the relation *visit*. Then the output will be a triple, i.e., (*Michael Jordan [PER], visit, Carnegie Mellon University [ORG]*).

So far, we only abstract the triple as the structured knowledge from the natural language sentence. However, it devotes nothing to increasing the scale of the knowledge repository such as Free-

base (Bollacker et al., 2007) which is a huge[1], public[2], collaborative[3](Bollacker et al., 2008) and online knowledge base with billions of triples and millions of disambiguated entities, and is primarily maintained by Google Inc., because we even do not know which exact *Michael Jordan* the triple (*Michael Jordan [PER], visit, Carnegie Mellon University [ORG]*) refers to in Freebase. As illustrated in Figure 1, there are three different persons named *Michael Jordan* in Freebase and each of them may be the protagonist of that news. Therefore, to populate knowledge repositories (Ji and Grishman, 2011), we need the *fourth operation*:

- **Entity Linking** (Rao et al., 2013): It concerns about the study of aligning a textual entity mention to the corresponding disambiguated entry in a knowledge repository. More specifically, since there are several *Michael Jordan* disambiguated by different MIDs (machine identifiers) as illustrated in Figure 1, we may build a classifier that can help assign the *Michael Jordan* in the extracted triplet (*Michael Jordan [PER], visit, Carnegie Mellon University [ORG]*) to the exact named entity in Freebase or find out that this *Michael Jordan* is a newly discovered named entity (NIL).

Hachey et al. (2013) and Rao et al. (2013) elucidate that most of the literatures (Bunescu and Pasca, 2006; Mihalcea and Csomai, 2007; Cucerzan, 2007; Milne and Witten, 2008; Ratinov et al., 2011) and the entity linking tracks[4] in TAC-KBP (McNamee and Dang, 2009; Ji et al., 2010) concentrate on linking ambiguous entities to the entries in Wikipedia, whereas our ultimate goal is to populate the structured knowledge repository, e.g., Freebase. However, to the best of our knowledge, few works (Zheng et al., 2012) concern about disambiguating named entities using Freebase which contains much more entries but less text information for each entry than Wikipedia.

Overall, Hachey et al. (2013) and Zheng et al. (2012) represent two research directions leverag-

---

[1]According to the statistics released on 10th March, 2014 by Google Inc., there are about 1.9 billion Freebase triples and 43 million entities.

[2]The whole dump of Freebase can be downloaded from `https://developers.google.com/freebase/data`

[3]One can access to Freebase and contribute more knowledge.

[4]`http://www.nist.gov/tac/2013/KBP/EntityLinking/index.html`

Figure 1: The disambiguated entities with the same name *Michael Jordan* in Freebase. The entities in Freebase are disambiguated by a unique machine identifier, e.g., the famous basketball player, Michael Jordan labeled by 054c1 (MID).

ing Wikipedia and Freebase, respectively. As both of the two collaborative web resources have their respective superiorities, i.e., more context information and more disambiguated entities, we begin to study a new paradigm that could bridge the gap between those two separated repositories and benefit from their respective advantages. From the perspective of supervised learning, entity linking can be naturally regarded as a classification problem. To build a training dataset for disambiguating a set of entities with the same name, we can firstly collect the sentences that mention that name from webpages, such as Wiki pages[5], and then manually annotate each entity mention with its unique machine identifier (MID) in Freebase given the contexts of sentences that it occurs in. However, hand-labeled data is time consuming and usually applicable to some specific classes of entities, such as person (*PER*), location (*LOC*) and organization (*ORG*). Therefore, we look forward to an approach that averts the tedious and laborious work.

Inspired by the idea of weak labeling (Fan et al., 2014; Craven et al., 1999), we contribute a new paradigm called distantly supervised entity linking (DSEL) without manual annotation in this paper. More specifically, we take advantage of a heuristic alignment assumption based on crowd sourcing to connect a certain disambiguated entity in Freebase with its related webpages. In these webpages, feature vectors can be extracted from the sentence-level textual contexts of that entity mention, and be labeled by its corresponding MID in

---

[5]The Wiki page for the famous basketball player, Michael Jordan, is `http://en.wikipedia.org/wiki/Michael_jordan`.

Topic equivalent webpage /common/topic/topic_equivalent_webpage

Topic equivalent webpage

http://en.wikipedia.org/wiki/index.html?curid=20455

http://ja.wikipedia.org/wiki/index.html?curid=30336

http://es.wikipedia.org/wiki/index.html?curid=10553

http://de.wikipedia.org/wiki/index.html?curid=32444

http://pt.wikipedia.org/wiki/index.html?curid=71267

http://fr.wikipedia.org/wiki/index.html?curid=50915

Figure 2: The topic equivalent webpages of the famous basketball player, *Michael Jordan* in Freebase.

Freebase. Then we can produce a large scale of weakly labeled[6] dataset in this way. Moreover, it is unrealistic to learn a specific classifier for each entity, as there are about 43 million disambiguated entities in Freebase. To tackle with those challenges, we propose a strategy of training a general classifier for disambiguating multiple entities and select a well known classifier, i.e., *liblinear* (Fan et al., 2008) to self-learn the weights among the high-dimensional sparse and noisy features. Experiments are conducted on a dataset of 140,000 items and 60,000 features. DSEL achieves a baseline F1-measure of 0.517. Furthermore, we analyze the performance influenced by other different features, and finally the F1-measure is improved to 0.545.

2 Paradigm

Traditional supervised learning methods for entity disambiguation require tedious labor on manual annotation to build training datasets. Manual annotation costs a lot, and can only cover some specific category, e.g., person names (Christen, 2006) as well. Therefore, we look forward to exploring a paradigm that could automatically generate large scale of open-category training datasets without manual annotation. Based on the dataset, we aim to build a practical classifier and generalize it to disambiguate more unlinked entity mentions in free texts.

Freebase contains 43 million disambiguated entities falling into 76 categories. Each entity is assigned by a unique machine identifer (MID). Those MIDs are the natural labels for the newly identified entity mentions linking to. However, there are inadequate free texts *locally* for extracting features, as Freebase is a well-structured

knowledge repository with billions of triples. Therefore, we resort to other free-text corpus that could be *distantly supervised* by Freebase and the key challenge is to find the bridge of supervision.

Fortunately, every entity in Freebase maintains a list of links to its topic equivalent webpages via crowd sourcing (Howe, 2006) as shown in Figure 2. These links will guide us to find the description webpages for that entity. Even though those links involves in different languages, we only choose the English Wiki pages to conduct experiments. Overall, we jointly exploit Freebase and Wikipedia to automatically construct the data for training a classifier.

3 Feature

For each entity in Freebase, we find its topic-equivalent Wiki page and extract the contextual features of its mention at sentence level.

Generally, we simultaneously choose K ($K = 1, 2, 3$) open-class words (Van Petten and Kutas, 1991), namely nouns, verbs, adjectives and adverbs, in front and behind the given entity mention. If we ignore the sequence of these words, we can gain the bag-of-words feature, whereas the word sequence feature. Furthermore, we use Stanford NLP core[7] and add the part-of-speech tagging feature which may help disambiguate those contextual words. Therefore, for each K size window surrounding the entity mention, we could extract four kinds of different features, i.e., bag of words (BOW), word sequence (WS), bag of words plus part-of-speech tagging (BOW + POS) and word sequence plus part-of-speech tagging (WS + POS). In total, there are twelve kinds of lexical features.

To elucidate the various kinds of contextual features, we randomly pick up a sentence from the Wiki page of the famous basketball player as example, i.e.,

His biography on the National Basketball Association (NBA) website states, "By acclamation, **Michael Jordan** *is the greatest basketball player of all time."*

The twelve kinds of lexical features for the sentence above are listed in Table 1. We will compare the performance among these features in Section 5.

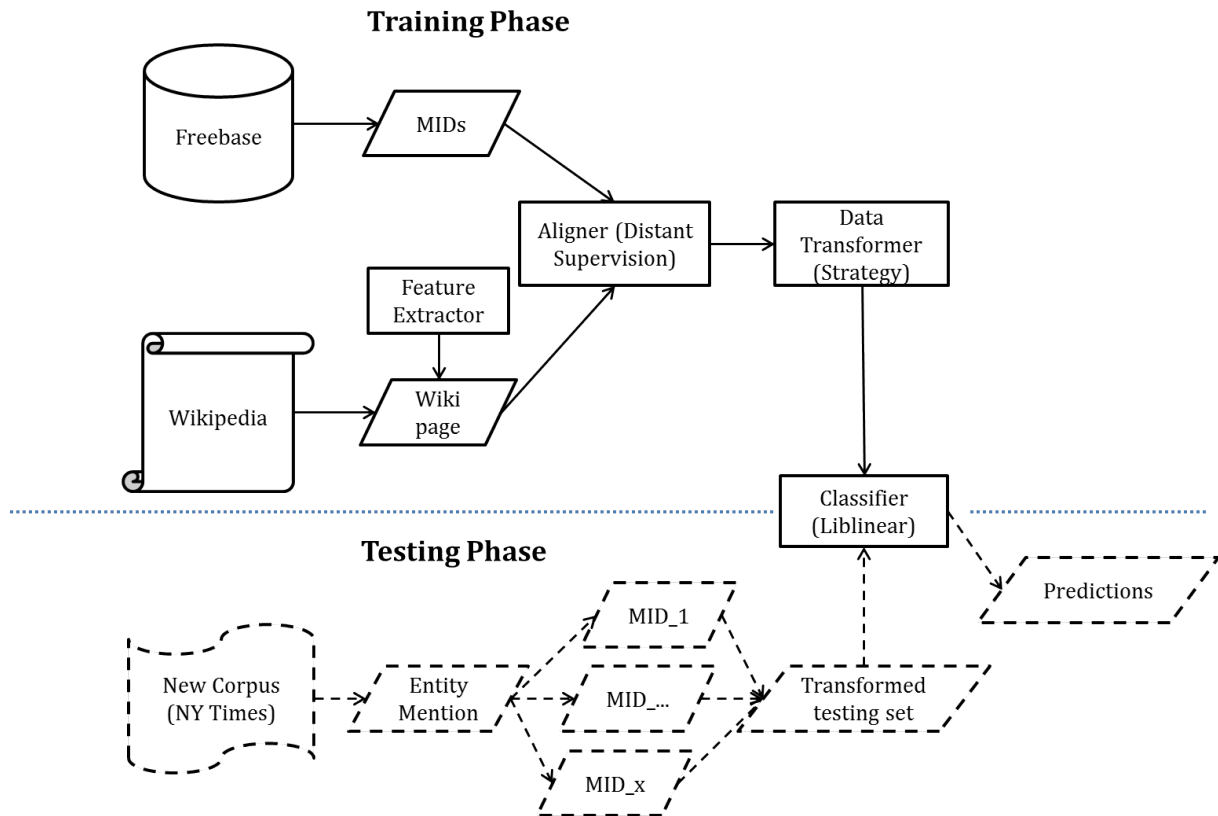---

[6]Auto-labeling via crowd sourcing may naturally bring about noise. Therefore, we regard the dataset weakly labeled.

[7]http://nlp.stanford.edu/software/corenlp.shtml

Training Phase

Testing Phase

Figure 3: The architecture of DSEL system.

Sentence	*His biography on the National Basketball Association (NBA) website states, "By acclamation, **Michael Jordan** is the greatest basketball player of all time."*
BOW ($K = 1$) **BOW** ($K = 2$) **BOW** ($K = 3$)	$<\{$acclamation$\}, \{$is$\}>$ $<\{$states$\}, \{$acclamation$\}, \{$is$\}, \{$greatest$\}>$ $<\{$website$\}, \{$states$\}, \{$acclamation$\}, \{$is$\}, \{$greatest$\}, \{$basketball$\}>$
WS ($K = 1$) **WS** ($K = 2$) **WS** ($K = 3$)	$<\{$acclamation$\}, \{$is$\}>$ $<\{$states-acclamation$\}, \{$is-greatest$\}>$ $<\{$website-states-acclamation$\},\{$is-greatest-basketball$\}>$
BOW + POS ($K = 1$) **BOW + POS** ($K = 2$) **BOW + POS** ($K = 3$)	$<\{$acclamation/NN$\}, \{$is/VBZ$\}>$ $<\{$states/NNS$\}, \{$acclamation/NN$\}, \{$is/VBZ$\} \{$greatest/JJS$\}>$ $<\{$website/NN$\}, \{$states/NNS$\}, \{$acclamation/NN$\}, \{$is/VBZ$\}, \{$greatest/JJS$\}, \{$basketball/NN$\}>$
WS + POS ($K = 1$) **WS + POS** ($K = 2$) **WS + POS** ($K = 3$)	$<\{$acclamation/NN$\}, \{$is/VBZ$\}>$ $<\{$states/NNS-acclamation/NN$\}, \{$is/VBZ-greatest/JJS$\}>$ $<\{$website/NN-states/NNS-acclamation/NN$\},\{$is/VBZ-greatest/JJS-basketball/NN$\}>$

Table 1: Twelve kinds of lexical features for the given sentence. A pair of angle brackets stands for a feature vector, e.g., $<\{$*states*$\}, \{$*acclamation*$\}, \{$*is*$\} \{$*greatest*$\}>$. A feature item is marked by a pair of braces, e.g., $\{$*states-acclamation*$\}$.

# of MIDs with the same name	# of names	# of MIDs with the same name	# of names	# of MIDs with the same name	# of names
2	4,467,216	5	180,489	8	60,273
3	740,530	6	134,012	9	41,256
4	440,261	7	76,459	10	33,628

Table 2: The distribution of ambiguous entities in Freebase.

4 Implementation

As we have already automatically produced a training dataset based on the proposed distant supervision paradigm, an intuitive idea is to feed a specific classifier for each ambiguous name with its unambiguous MIDs and the corresponding feature vectors. However, Table 2 shows that there are at least 5.5 million names that denominate more than one entity (MID) in Freebase. Therefore, it is infeasible to build 5.5 million specific classifiers. To train a general classifier that does not restrict itself to disambiguating a certain name, we adopt a strategy that merges those specific classifiers. Concretely, we transform MIDs, the original labels into features and use 1/0 to indicate whether the contextual features from Wiki pages and MIDs in Freebase match or not with each other. If we choose the BOW ($K = 3$) feature in Table 1 for instance, one positive training sample will contain a new feature vector ($<$\{website\}, \{states\}, \{acclamation\}, \{is\}, \{greatest\}, \{basketball\}, \{MID:054c1\}$>$) labeled by 1. To balance the training dataset, we randomly pick up features from other entities uniformly named to generate negative samples. For example, another well-known *Michael Jordan* (MID:0bby3vs) is an English mycologist. We can extract a BOW ($K = 3$) feature vector, i.e., $<$\{is\}, \{English\}, \{mycologist\}$>$, and it concatenates \{MID:054c1\} to construct a negative sample labeled by 0.

The distant supervision paradigm and the strategy of building the training set for a general classifier lead to high-dimensional noisy and sparse features. Moreover, given the millions of training samples produced by aligning Freebase and Wikipedia, we choose a linear classifier that is based on logistic regression approach, i.e., Liblinear (Fan et al., 2008), to rapidly self-learn the weights among the high-dimensional sparse and noisy features.

For a newly discovered entity mention in the testing corpus, we firstly extract its contextual feature, e.g., bag of words as above. Then the feature concatenates all the candidate MIDs that share the same name with that entity mention. Each testing sample within the same name collection will predict a score indicating the strength of linking. For each collection, the Top-N predictions with higher probabilities are selected for evaluation.

We summarize the procedures of implementing our proposed paradigm and use Figure 3 to demonstrate the architecture of DSEL system.

5 Experiments

In this section ,we report the experimental results following the procedures described in Section 4. To evaluate the performance of different features, we adopt three widely used metrics (Meij et al., 2013), namely precision, recall and F1-measure.

5.1 Dataset

We randomly select 20,000 ambiguous names (collections) in Freebase. About 82,000 sentences that contain at least one entity mention are extracted from the topic-equivalent Wiki pages. For each collection, 80% sentences are randomly picked up for constructing the training set and 20% remains are for held-out evaluation. Following the procedures of building training samples described in Section 4, we gain a dataset including around 140,000 items and 60,000 features.

5.2 Evaluation metrics

Precision and recall are widely used metrics to evaluate different rank-based approaches on entity linking. F1-measure synthetically measures precision and recall by calculating the harmonic mean of them. Suppose that C denotes the whole collection set for testing. $C_{i,j}$ represents the set of Top-j predictions with higher probabilities in the i-th collection. G_i stands for the set of gold standards of the i-th collection. $\#(S)$ is the function that counts the entries in set S. Then the formulae to calculate precision, recall and F1-measure are as follows,

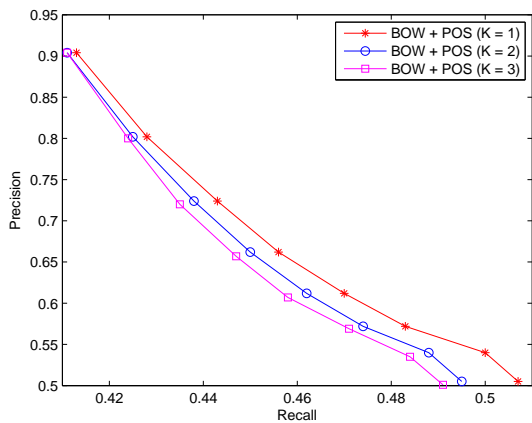| Feature type	Avg. F1-measure	Feature type	Avg. F1-measure
BOW ($K = 1$)	**0.539**	**WS** ($K = 1$)	**0.544**
BOW ($K = 2$)	0.531	**WS** ($K = 2$)	0.532
BOW ($K = 3$)	0.529	**WS** ($K = 3$)	0.518
BOW + POS ($K = 1$)	**0.540**	**WS + POS** ($K = 1$)	**0.545**
BOW + POS ($K = 2$)	0.532	**WS + POS** ($K = 2$)	0.531
BOW + POS ($K = 3$)	0.529	**WS + POS** ($K = 3$)	0.517

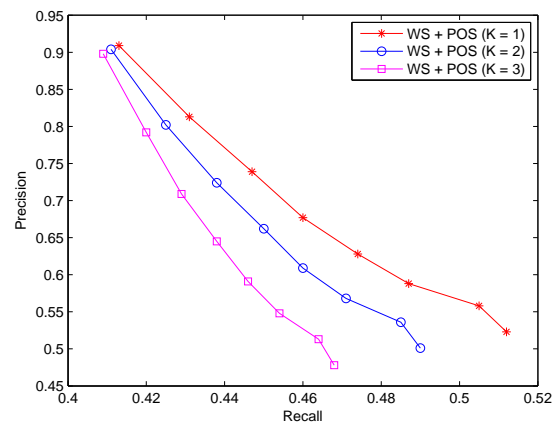Table 3: The F1-measure comparison among different features.
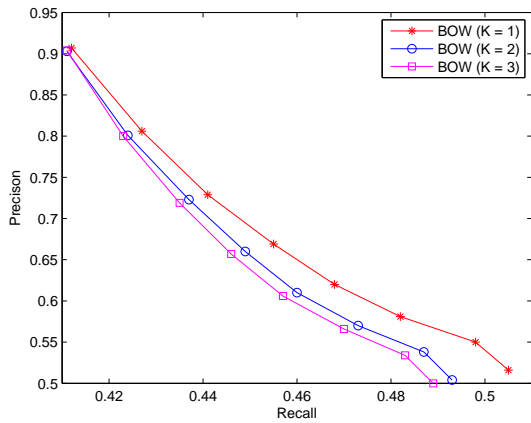
(a) Precision-Recall curves for BOW features.

(b) Precision-Recall curves for BOW + POS features.

Figure 4: Precision-Recall curves for the BOW-class lexical features.
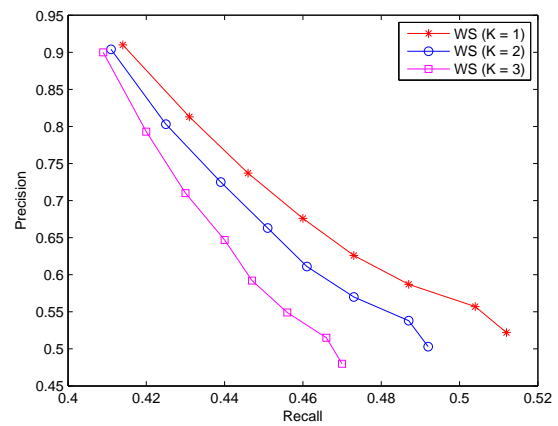
(a) Precision-Recall curves for WS features.

(b) Precision-Recall curves for WS + POS features.

Figure 5: Precision-Recall curves for the WS-class lexical features.

$$\text{Precision} = \sum_i \sum_j \frac{\#(C_{i,j} \bigwedge G_i)}{\#(C_{i,j})},$$

$$\text{Recall} = \sum_i \sum_j \frac{\#(C_{i,j} \bigwedge G_i)}{\#(C)},$$

$$\text{F1-measure} = \frac{2 \times Precision \times Recall}{Precision + Recall}.$$

5.3 Feature comparison

For each type of feature, we conduct one trial and tune the parameters for the logistic classifier using 5-fold cross validation. Then we adopt held-out testing taking advantage of the 20% sentences left.

Figure 4 and Figure 5 show the precision-recall curves for the twelve lexical features, and Table 3 displays the average F1-measure comparison among different features. We find out that the WS-class features generally outperform the BOW-class features, and the short-distance contextual features ($K = 1$) are more effective than the long-distance ones ($K = 2, 3$).

6 Conclusion and Future Work

As far as we know, it is the first attempt to deal with the task of entity linking based on the idea of distant supervision. We leverage a heuristic alignment assumption, i.e., the topic equivalent pages, to bridge the gap between Freebase and Wikipedia and jointly use those two knowledge bases to automatically produce training data without manual annotation. Moreover, we propose a strategy that transforms labels into features and feed them to a general classifier, rather than building an individualized classifier for each ambiguous name for millions of entities.

For the future work, we believe that this new paradigm leaves several open questions:

- Besides the entities (MIDs) that have already been stored in knowledge repositories (Freebase), new entity instances (NIL) with the same name need to be discovered. Therefore, further study could focus on extending paradigm to identify unknown entities.

- The link for many other webpages in different languages are also provided in Freebase, as illustrated in Figure 2. It may facilitate the research of cross-lingual entity linking.

- The alignment assumption is simple and heuristic. Further studies may dedicate on discovering other reasonable alignment principles.

- Even though the strategy for generating training data that fits a general classifier, it rises the problem that high-dimensional sparse and noisy features impact the effectiveness and efficiency of the proposed paradigm.

Generally speaking, the experiments prove that our new proposed paradigm is promising and it is worthy of being further studied.

Acknowledgements

References

Nguyen Bach and Sameer Badaskar. 2007. A review of relation extraction. *Literature review for Language and Statistics II*.

Kurt Bollacker, Robert Cook, and Patrick Tufts. 2007. Freebase: A shared database of structured general human knowledge. In *Proceedings of the national conference on Artificial Intelligence*, volume 22, page 1962. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

Razvan C Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL*, volume 6, pages 9–16.

Peter Christen. 2006. A comparison of personal name matching: Techniques and practical issues. In *Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on*, pages 290–294. IEEE.

Mark Craven, Johan Kumlien, et al. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, volume 7, pages 708–716.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Miao Fan, Deli Zhao, Qiang Zhou, Zhiyuan Liu, Thomas Fang Zheng, and Edward Y Chang. 2014. Distant supervision for relation extraction with matrix completion. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 839–849.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R Curran. 2013. Evaluating entity linking with wikipedia. *Artificial intelligence*, 194:130–150.

Jeff Howe. 2006. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1148–1158. Association for Computational Linguistics.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the tac 2010 knowledge base population track. In *Third Text Analysis Conference (TAC 2010)*.

Paul McNamee and Hoa Trang Dang. 2009. Overview of the tac 2009 knowledge base population track. In *Text Analysis Conference (TAC)*, volume 17, pages 111–113.

Edgar Meij, Krisztian Balog, and Daan Odijk. 2013. Entity linking and retrieval. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 1127–1127. ACM.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM.

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

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1396–1411. Association for Computational Linguistics.

Delip Rao, Paul McNamee, and Mark Dredze. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*, pages 93–115. Springer.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. 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. Association for Computational Linguistics.

Sunita Sarawagi. 2008. Information extraction. *Foundations and trends in databases*, 1(3):261–377.

Cyma Van Petten and Marta Kutas. 1991. Influences of semantic and syntactic context on open- and closed-class words. *Memory & Cognition*, 19(1):95–112.

Zhicheng Zheng, Xiance Si, Fangtao Li, Edward Y Chang, and Xiaoyan Zhu. 2012. Entity disambiguation with freebase. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 82–89. IEEE Computer Society.