Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 13, no. 8, December 2017, pp. 1293-1303
DOI: 10.23940/ijpe.17.08.p11.12931303

Entity Disambiguation with Markov Logic Network Knowledge Graphs

Jiangtao Ma[a,b], Tao Wei[c,d], Yaqiong Qiao [a], Yongzhong Huang[a,*], Weibo Xie[a], Chaoqin Zhang [a,b],Yanjun Wang [a,b], Rui Zhang[a,e]

[a]*State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450002, China*
[b]*Zhengzhou University of Light Industry, Zhengzhou 450002, China*
[c]*National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 451000, China*
[d]*Henan Institute of Engineering, Computer College, Zhengzhou 451000, China*
[e]*North China University of Water Resources and Electric Power, Zhengzhou 450002, China*

Abstract

Disambiguating named entities is an important problem in natural language processing, knowledge base, question answering systems. In the paper, we propose a Markov logic network knowledge graph solution for solving entity resolution problem. First, we employ knowledge graph to represent the entity relationship between linked entities in the knowledge base. Then, we utilize MLN to inference the inconsistent relationship in the knowledge graph, and disambiguate the entities in the process of entity disambiguation. As far as we know, inferencing with MLN is a first attempt for entity disambiguation in the knowledge graph. We evaluate the proposed solution with three real world knowledge bases and compare it with four baseline solutions. The experimental results demonstrate that our solution is 7% higher than other baseline methods with F1 measure. We also test our scheme and compare entity resolution systems on four datasets with three knowledge base corpora. Extensive experiments show that our solution achieves higher precision and recall than baseline solutions.

Keywords: Markov logic network; knowledge base; knowledge graph; entity resolution

(Submitted on July 25, 2017; Revised on August 30, 2017; Accepted on September 15, 2017)

(This paper was presented at the Third International Symposium on System and Software Reliability.)

1. Introduction

With the explosion of information on the World Wide web, the traditional methods are unable to effectively handle the massive data growth, manual identification efficiency is low and costly human and financial problems result. With the rapid development of information technology, automated named entity disambiguation method needs to adapt to the ever-expanding digital environment and make up for the lack of traditional methods. Knowledge fusion from different sources has received substantial attention in recent years [9] since knowledge fusion can let the knowledge be used repeatedly. However, building a knowledge base needs labor-intensive work and time cost. In order to avoid repetitive work, we need to reuse and share knowledge. Thus, we have to fuse data and information with different sources. In the information fusion process, we need to disambiguate the concept, instance, attribute, relation, and map the entity, concept and instance. Knowledge fusion can be completed through manually and automatically method. Manually method is suitable for small knowledge base; however, it needs labor-intensive work and is error-prone. On the other hand, large scale knowledge base needs automatically method, which is scalable as it is built on machine learning and ontology engineering.

Existing solutions to entity disambiguation includes DoSeR [30], DBpedia Spotlight [20], WAT [22], AIDA [30] and Wikifier [20]. There are knowledge bases such as YAGO [22], Probase [13] employ these entity resolutions methods to disambiguate entities. YAGO knowledge base fuses the Wikipedia, WordNet and Geo-Names into its disambiguation

* Corresponding author.
E-mail address: 18600200718@163.com

system. Wikipedia classifies the entities according to hierarchy structure based a directed acyclic graph. However, this structure only reflects topic information and error-prone. In order to disambiguate the entities, YAGO employs co-occurrences and the word frequency in WordNet to construct complete classification system.

Probase [4] utilizes a probabilistic-based entity resolution method to fuse knowledge base, which fuses Freebase [1] into its system. Probase views the data fusion problem as matching and mapping problem from known classification system. The first problem to be solved is entity resolution, that is to decide whether two entities belong to one entity or not. Probase classifies the outer data sources into positive evidence and negative evidence, and converts the entity resolution problem into the optimization on graph's multi-way cut problem. This entity resolution method can solve the problem of attribute information deficiency. In addition, in order to solve the high time cost on large scale entity resolution, Lee et al. [13] propose CnD and BoF principles to insure the scalability of entity resolution.

On behalf of helping customers find new data and knowledge quickly and easily, Google search publishes a knowledge graph that can systematize search results and achieve a complete knowledge of any keyword. The knowledge graph gets professional information from the Freebase or Wikipedia and improves the depth and breadth of search results through large-scale information search analysis. Compared with the previous search results, the knowledge graph will be in three areas to greatly improve the final results of Google search to find the correct results. Because a keyword may represent multiple meanings, the knowledge graph will show the most comprehensive information that allows the user to find the meaning that is most wanted. With the knowledge graph, Google can better understand the user search information and summed up the relevant content and topics. Since the knowledge graph will give the search results of the complete knowledge system, users tend to find a lot of things they do not know before searching with Google search engine. We can apply knowledge graph to represent the entity relationship in the knowledge base.

Markov logic network [16] (MLN) combines the first-order logic and the probability graph model into a representation. MLN is a first-order logical knowledge base with weights for each criterion or statement, where the constants represent objects in the library. It also states that in a basic Markov network knowledge base, each possible primitive of a first order logic criterion carries a corresponding weight. Markov's logical reasoning is achieved by applying the Markov chain Monte Carlo method to the minimum subset of elements needed to answer the question. The weight is obtained from the relational database through the iterative efficient learning of the quasi-likelihood metric. The extra clause can be learned using the inductive logic program technique. MLN can be utilized to solve the inconsistent information in the knowledge base.

We provide a new solution to disambiguate entities and combine knowledge graph and MLN to improve the accuracy and performance. To achieve that, we employ knowledge graph to represent the linking relationship between entities, and utilize MLN to represent the uncertainty of relationship between entities and inference the inconsistent to disambiguate the entities. Consequently, we combine the knowledge graph and MLN to disambiguate named entities in the text. Compared with existing baseline methods, the contributions of proposed method are summarized as follows:

- We propose a new solution that utilizes knowledge graph to represent the entity relationship between linked named entities in the text.
- We utilize MLN to inference the inconsistent relationship in the knowledge graph, and disambiguate the entities in the process of entity resolution. As far as we know, this is the first trial on entity disambiguation.
- We test our solution on three knowledge base corpora and compare it with existing baseline methods. Experimental results demonstrate that the proposed solution achieves 7% higher F1 than baseline solutions. We also test our scheme and compare entity resolution systems on four data sets with three knowledge bases. The results show that our framework achieves higher precision and recall than baseline solutions.

We review the background on knowledge base, knowledge graph and Markov logic networks in Section 1. Section 2 summarizes the related work on entity resolution and entity alignment. Section 3 describes and formulates the entity resolution problem in the knowledge base. Section 4 presents our entity resolution framework and Section 5 reports the experiments results for entity disambiguation and compares it with some baseline methods. Section 6 makes the conclusion of the whole paper.

2. Related Work

2.1. Entity Resolution

Researchers propose several solutions for solving named entities disambiguation (NED) problem. Cucerzan [29] tackles NED problem with data extracted from Wikipedia. The researcher employs a local method to maximize the similarity

between the input information and the concept in Wikipedia, and proposes a large-scale entity disambiguation system. Ratinov et al. [2] use Wikipedia knowledge base to disambiguate entities. Their experiments show that employing Wikipedia or other knowledge base methods achieve higher performance than using text similarity-based solutions for disambiguating entities. Wikipedia Miner [24] utilizes machine learning algorithms to disambiguate named entities. The researchers test their solution with Wikipedia and an AQUAINT subset. The AIDA solution [30] employs complicated graph algorithms and YAGO2 knowledge base for NED tasks. This solution utilizes dense sub-graphs to disambiguate named entities and uses a greedy algorithm for the scalability of Web scale. Additionally, AIDA can disambiguate the similar contexts, entities and context windows. DBpedia Spotlight [8] can mark and wikification relational entities. It employs vector-space model and vector's cosine similarity to solve NED tasks. By comparison to other solution, Spotlight can disambiguate all kinds of the DBpedia ontology. Spotlight is popular in many developing communities and has been adopted by many projects.

Ferragina and Scaiella [23] propose an updated version of their disambiguation system named TagMe2. TagMe2 can handle short texts around 30 words. TagMe2 is based on an anchor catalog, a page catalogue and an in-link graph. TagMe2 employs anchor catalog to match terms and identifies named entities. It combines in-link graph with page catalog and identified anchors entities to disambiguate the candidate entities. TagMe2 deletes the non-relevant named entities from the input short texts. Cornolti et al. [7] propose a benchmark test for named entity disambiguation methods. They compare six existing approaches in five well-known datasets. Furthermore, they defined different kinds of named entity annotation task.

2.2. Entity alignment

The entity alignment is mainly used to eliminate entity collisions in heterogeneous data source. Herzog et al. [11] employs the probability of entity linking model, assigns a different weight for each matching attribute pair, thereby improving the accuracy of the entity linking. Christen [7] proposes a novel SVM classification method based on the two-stage entity link analysis model. The matching accuracy is much higher than that of TAILOR. A novel adaptive entity name matching and clustering algorithm is proposed [12], which can generate an adaptive distance function by training the sample. In the paper, authors employ supervised learning method to train the distance function in the conditional random field based entity alignment model, and then adjust the weight to maximize the product of the characteristic function and the learning parameter.

In the active learning methods, staff continuous interact to solve the deficiency of training data problems. Sarawagi and Bhamidipaty [5] propose an ALIAS system that can be constructed through the human-computer interaction to complete the entity linking and dereliction of the task. Tejada et al. [6] build an Active Atlas system in a similar way. Based on the above solutions, Lacoste-Julien et al. [25] propose an algorithm called SiGMa for large-scale knowledge base entity alignment. The algorithm regards the entity alignment problem as an optimization of the global matching score objective function. The problem is belonged to the quadratic assignment problem. The approximate solution can be obtained by the greedy optimization algorithm.

McCallum and Wellner propose [27] a conditional random fields(CRF) entity analysis model based on graph partitioning technology. The model makes the decision of entity identification based on the observed value, which is helpful when dealing with the data of dependency relationship between attributes. Singla and Domingos [15] propose an entity analysis method based on Markov logical network, which transforms the maximum likelihood calculation problem in the probability graph model into a typical maximization weighted satisfying problem. However, based on the Markov network, a series of equivalent predicate axioms need to be defined, through which the entities are aligned in the knowledge base.

3. Problem Statement

Figure 1 is an illustration of the example showing how MLN knowledge graph can resolve entity resolution problem. Entities are shown in nodes, dotted line represents co-referent entities found with entity resolution, and dashed line means homonymy entities. The weight of edge represents the probability of two entities are same entity, weight$\in$[*0,1*]. When $weight_{(Apple,Apple)}$=0, which represents Apple and Apple are homonymy entities, they are not the same entity. When $w_{(Apple,Mac)}$=1, which represents Apple and Mac are similar or same entities. The aim of entity resolution is find the probability of two nodes in the MLN knowledge graph.

Many entities are homograph in knowledge bases; different words share the same entity in the real world while many entities share same words refer to different entities. Take Figure 1 for example. The two apples are not same entity in the MLN knowledge graphl the edge's weight between Apple and Mac is 0. While Apple and Mac are similar or the same entity,

the weight of edge between Apple and Mac is 1. Our solution utilizes entity resolution to determine co-referent entities and disambiguate homonymy entities with MLN knowledge graph, producing a probability between to nodes in the MLN knowledge graph.

4. Scheme Details

The details of proposed solution will be discussed in this section. First, we utilize knowledge graph to construct knowledge base. Then, we utilize MLN to the probability of two nodes in the knowledge graph. Third, we employ tensor decomposition to do the entity resolution. Finally, we apply reinforcement learning to improve the entity resolution in our solution.

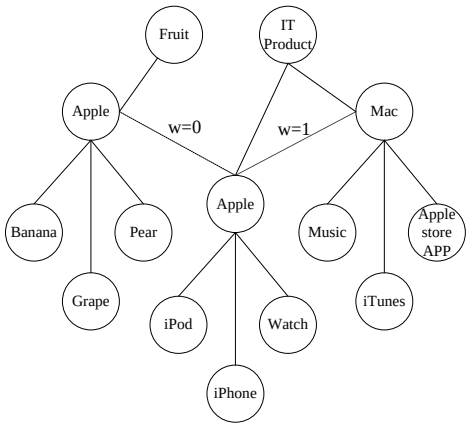

Figure 1. An illustration of the example MLN knowledge graph for entity resolution

4.1. Knowledge Graph

The knowledge graph is a semantic network that reveals the relationship between entities, and can be formally described in real world things and their relationships. We use triple tuples to represent knowledge graph G = (*E*, *R*, *S*), where E= {e_1, e_2, ⋯, $e_{|E|}$} represent the set of linked entities within the knowledge graph, |*E*| represents the quantity of entities. R= {r_1, r_2, ⋯, $r_{|E|}$} represents the relationship among linked entities. |*R*| represents the quantity of relationships. The basic form of triple tuples mainly includes entities, relationships, entities and concepts, attributes, attribute values, etc. Entities are the most basic elements in the knowledge graph, and there are different relationships between different entities. Concepts mainly refer to collections, categories, object types, and types of things, such as fruit, IT product in the Figure 1. Attributes mainly refer to objects that may have properties, characteristics, features, and parameters, such as the size, taste of the apple fruit. Each entity can be described as a unique ID. Each attribute (attribute-value pair, AVP) can be used to characterize the intrinsic properties of an entity. The relationship can be used to connect two entities, each of which can be used to link two entities, which characterize the association between them.

4.2. Markov Logic Network

In a learned Markov logic network, the weight of the rule can be calculated by the empirical probability of containing the rule. Only we can see that if a relationship violates the rules and the weight of the rules is small, then the relationship is likely to exist. Therefore, Markov logic network can effectively deal with the contradiction of relationship in the knowledge base, and tolerate its inherent flaws. It needs to be pointed out that the contradiction of relationship in the knowledge base is not necessarily its inherent flaw because the knowledge base is often only a small part of real world and the real world is always a contradictory complex system. MLN combines statistical machine learning and first order logic; thus, it has an advantage over logic or statistical method in solving entity resolution problem. MLN is a statistical learning framework, with a strong ability to describe, logical reasoning ability and ability to deal with uncertainty. From the point of dealing with uncertainty, MLN employs first-order to predicate the weight of edge between entities, which can tolerate the knowledge of incomplete and contradictory (inconsistent) in the knowledge base, and has the ability to deal with the problem of uncertainty. In terms of probability and statistics, MLN provides a simple and effective method for describing Markov networks.

L is a set of binary tuples (F_i, ω_i), the set of rules are represented by F_i, ω_i is a real number. The two tuples (F_i, ω_i) and a set of constant C={c_1,c_2,…,c_n} construct MLN. Each ground atom corresponds a binary value node in the L. If the ground atom is true, the binary value is 1, otherwise, the binary value is 0. Each ground formula has a feature value, if the ground

formula is real, the correspond value is 1; otherwise, the value is 0. And ω_i is the weight of feature value corresponding to the rule F_i in the binary tuples. Since the node in MLN is derived from ground atom, the edge is derived from the relationship of ground atom. Therefore, template can be made for deriving Markov network. Given the same Markov logic network and different finite constants set C, different Markov networks can be generated, and the difference in size of these Markov networks can be very large. However, different Markov networks generated from the same MLN have some commonalities in the structure and parameters. For example, given the same number of groups, all the possible constants of the same rule have the same weight, etc. Each Markov network generated in this way may be referred to as a ground Markov network. Equation 1 gives the probability distribution of a possible ground formula *x* within Markov logic network:

$$p(\mathrm{X}=\mathrm{x})=\frac{1}{z}\exp\{\sum_i w_i n_i(x)\}=\frac{1}{z}\prod_i \phi_i(x_{\{\mathrm{i}\}})^{n_i(x)} \tag{1}$$

Where, $n_i(x)$ represent the number of true rules corresponding ground formula; $x_{\{i\}}$ represents the state of atom in formula F_i, and $\phi(x_{\{\mathrm{i}\}})=\mathrm{e}^{w_i}$.

The first equation in (1) gives the logarithmic linear model of the Markov logical network, and the second equation uses the equivalent form of the potential function. Intuitively, the rules with great weight represent the rule (1), if a world violates this rule, the probability of the existence of the world will tend to 0. In fact, the variables involved in rule F_i are usually found in other rules, When the value of the rule F_i is changed, there is no guarantee that the values of the other rules remain unchanged. Therefore, there is no one-to-one correspondence between the weight of the rule F_i and its probability. The weight is regarded as the maximum entropy distribution, or the weight of the rule is regarded as the empirical probability, then the probability of all the rules together determines the weight of the rule F_i.

4.3. Entity Resolution with Tensor Factorization

In order to compute the relationship in the MLN knowledge graph efficiently, we use tensor decomposition method to analyse the MLN knowledge graph. As Figure 2 illustrates, we use tensors to represent MLN knowledge graph and utilize CANDECOMP/PARAFAC (CP) tensor decomposition [19] to inference the probability of relationship between nodes.

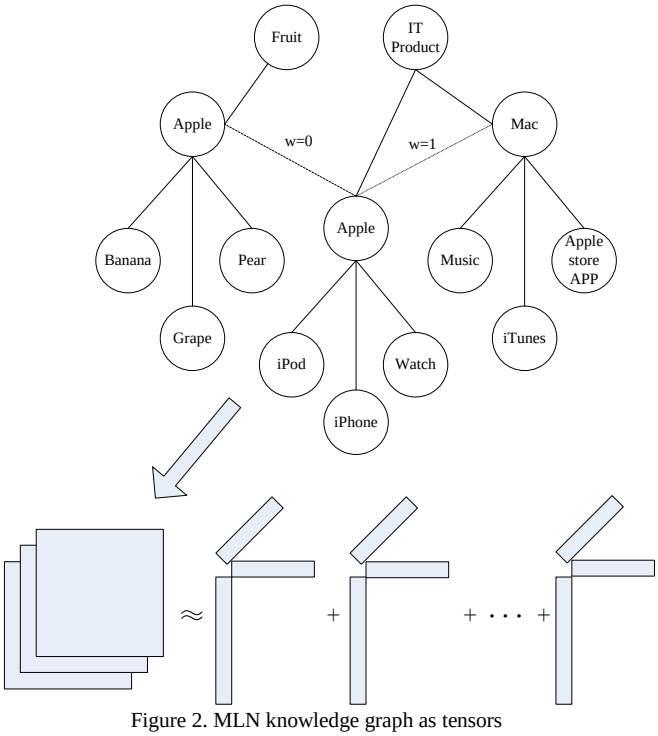

Figure 2. MLN knowledge graph as tensors

Nonnegative tensor decomposition is the tensor decomposition constraint factor is nonnegative. The multiplicative update rule is the most widely used nonnegative matrix decomposition method, which uses Kullback-Leibler divergence and

Euclidean distance to measure the cost function. Equation (2) is the Euclidean distance description of non-negative matrix decomposition on multiplicative update rule.

$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T WH)_{a\mu}}, W_{a\mu} \leftarrow W_{a\mu} \frac{(V^T H)_{a\mu}}{(WHH^T)_{a\mu}} \tag{2}$$

When the matrix $\boldsymbol{V}$ is decomposed into $\boldsymbol{WH}$, the Euclidean distance $\| \boldsymbol{V}\text{-}\boldsymbol{WH} \|$ does not grow under the formula (2). The Euclidean distance is a constant when $\boldsymbol{W}$ and $\boldsymbol{H}$ has a fixed distance.

Since this rule only contains multiplication and division operations, if the initial matrix is nonnegative, all intermediate results are nonnegative.

Given a tensor Z of size $\boldsymbol{O} \times \boldsymbol{P} \times \boldsymbol{Q}$, its *k*-component CP is decomposed into [26]:

$$Z \approx \sum_{k=1}^{K} \lambda_k l_k \circ m_k \circ n_k \tag{3}$$

Where $\circ$ indicates the outer product, $X = l \circ m \circ n$, and $\lambda_k \in \mathbb{R}_+$, $a_k \in \mathbb{R}^I$, $b_k \in \mathbb{R}^N$, $c_k \in \mathbb{R}^T$, k=1,…,K. Each $\lambda_k \circ l_k \circ m_k \circ n_k$ indicates one component, each component indicates one factor. Here, $\|l_k\|=\|m_k\|=\|n_k\|=1$, where λ_k is the coefficient of k^{th} component.

The greatest advantage of the CP model is its interpretability and that there is no restriction (there is no orthogonality limitation in SVD decomposition). The factors a_k and b_k represent the potential relationships between nodes in a MLN knowledge graph, and c_k represent the relationships among nodes in multiple networks. Therefore, the CP model can inference the potential relationships between nodes in MLN knowledge graph.

4.4. The Proposed Solution for Entity Resolution

Figure 3 shows our framework for entity resolution. When we have an entity-based search in knowledge base, we first build a MLN knowledge graph according knowledge base. Then, the framework enumerates candidate entities in MLN knowledge graph. After that, the framework identifies the potential matching entities. Furthermore, CP tensor factorization is adopted in proposed framework for improving the efficiency of inference in the MLN knowledge graph. After ranking the similar entities, the candidate results will be tested according to a threshold. If the candidates are not qualified, the framework will enumerate candidate entities in MLN knowledge graph again. Otherwise, the matched entities will be returned.

5. Evaluation

Extensive experiments are conducted on real-world datasets in this section. We employ YAGO, DBpedia, Wikidata to test our solution, and we compared our solution to AIDA [30], DBpedia Spotlight [8], WAT [9] and Wikifier [20]. Experiments are performed on a cluster with Intel Xeon E5-2620 V3 CPU, NVIDIA Tesla K80 GPU, Intel Xeon Phi 7120P, 128 GB main memory, 1T SSD, 6T SAS disk and CentOS release 6.4 64-bit version. We also utilize GERBIL- General Entity Annotator Benchmark [3] which include Spotlight and WAT that can offer comparison to other methods using multiple datasets. We utilize Precision and Recall to assess the performance of baseline solutions. Precision is the ratio of true entities among the returned result. Recall is the proportion of the entities resolution that are included in the returned correct result.

5.1. Corpus and Datasets

5.1.1 Corpus and Datasets

Yet Another Great Ontology [14] (YAGO) is designed by Max Planck Institute started from 2006. YAGO includes most instances of Wikipedia, such as singers, movies, cities. However, the hierarchy of these categories do not apply directly to word hyponymy. In contrast, WordNet has an accurate hyponymy but fewer instances. Therefore, the merger of the two

resources will play their respective advantages. YAGO also cooperates with DBpedia, Wikidata, Geo-names and the Google Knowledge Vault. YAGO employs manually test to ensure the quality of data. YAGO makes special efforts on anchoring to time and spacial facts in knowledge base.

DBpedia [28] project was started from University of Berlin and Leipzig University; the first dataset was opened to public in 2007. It is available to users reusing the dataset. The entities and facts information are extracted from Wikipedia pages and info box tables inside the Wikipedia pages. The 2014 version of DBpedia has more than 4.58 million items, including 1.45 million people, 735,000 locations, 123,000 records, 87,000 movies, 16,000 computer games and 240,000 organizations. The knowledge base is not only used by BBC, Reuters, the New York Times, but also Google, Yahoo and others search engines.

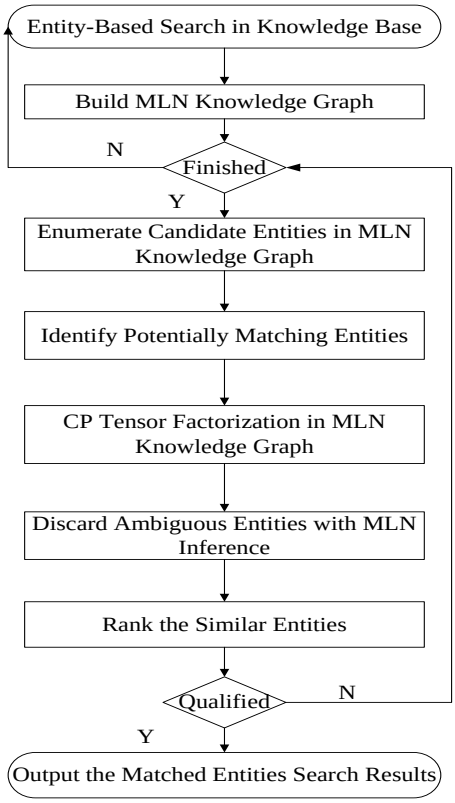

Figure 3. Entity resolution with MLN knowledge graph

Wikidata [18] is first proposed by the Wikipedia Foundation in Germany. The goal is to design a "world knowledge of a collaborative editor of the database", which will provide support for more than 280 language versions of Wikipedia. Designers employ the high quality of Wikipedia content through such a shared database to ensure the consistency of Wikidata in different languages. Wikidata was extracted from different language versions of the Wikipedia which have a common understanding entities and facts. The structured data were extracted from Wikipedia pages. Wikidata is a large knowledge database that can be read and edited by people and computer programs. Wikidata allows users to participate in data management. The distinguished feature is that data are entered in any language, then it will be displayed in other languages immediately. More importantly, Wikidata pays more attention to the quality of data source. In short, compared with other data sources, Wikidata has open, collaborative, multi-lingual and structured features.

5.1.2 Datasets

In this section, we introduce four publicly available datasets which are used in our experiments. All data sets are contained in General Entity Annotator Benchmark, which evaluates disambiguation system with several datasets. Table 1 describes the statistics on the test datasets.

ACE corpus includes annotated entities and relations and was created by Linguistic Data Consortium. The data we utilize is a proportion of the ACE2004 coreference papers which has 63 articles along with 282 entities.

Table 1. Statistics on test datasets

Dataset	Topic	Document	Entity	Entity/Document
ACE	news	63	282	4.47
AIDA/CO-NLL	news	228	4317	18.93
AQUAINT	news	48	716	14.92
MSNBC	news	19	632	33.26

AIDA/CO-NLL dataset was extracted from an evaluation competition [30]. The corpus was partitioned into one training dataset and two test datasets. There are 228 documents in the test datasets, and there are 18.93 entities in each document on average.

AQUAINT is offered by [24], which includes 48 documents and 14.92 entities in each document on average from a news dataset from news agency, such as New York Times.

MSNBC was offered by [29], which contains business news, health report, sports and travel information, and includes 19 news documents and 632 entities.

5.2. Compared Methods

We make comparison our solution with following solutions for entity resolution.

Wikifier [20] is a famous wikification framework developed in 2013, which employs statistical methods to recognize named entity from text. Wikifier utilizes Wikipedia as background knowledge base and can link entities from text to Wikipedia pages.

DBpedia Spotlight [8] is aimed for annotating entities from text documents with DBpedia URIs. It is an open source Web Service and is freely available for every user. Users can choose its configure according to their real needs with ontology or quality measures, such as relative importance, topic relevant degree, context ambiguity and entity disambiguation. DBpedia Spotlight rely on the RDF knowledge base DBpedia, YAGO, and Wikipedia knowledge base. Since entities within DBpedia, Wikipedia and YAGO offer same relations, it is easily to evaluate the disambiguation accuracy while utilizing same datasets.

AIDA [30] is a robust solution for collective disambiguation by annotating entity from knowledge bases and employing coherence graph method. It merges probability distribution, text similarity and the concurrence of entities in the text into a framework to solve the entity disambiguation in the context. A weighted graph of candidate entities is built within AIDA solution, and a dense subgraph that reflects the mention-entity mapping is searched in the weighted graph. AIDA also rely on the RDF knowledge bases DBpedia and YAGO and Wikipedia knowledge.

WAT also utilizes Wikipedia as background knowledge base and link entities in the text directly to Wikipedia pages. WAT redesigns TagMe [23] components and employs graph-based method and vote-based method for entity disambiguation. WAT has better modular of software and more efficient than TagMe. WAT re-designs spotting, disambiguation and pruning modules to improve the annotation pipeline.

5.3. Experiment Results

In this section, we give the experimental results on compared with DoSeR framework and Wikifier, DBpedia Spotlight, AIDA and WAT entity resolution systems, which utilize Wikipedia corpora.

5.3.1 Compared with DoSeR framework

We compare our solution against DoSeR [17] on DBpedia corpus. We also test whether our solution performs better on YAGO3 and Wikidata. Table 2 shows the experiments results compared with DoSeR. The results demonstrate that the proposed framework achieves higher performance than baseline methods on these datasets. Despite employing the same candidate selection method as introduced in DoSeR, our method achieves 8% F1 value higher than DoSeR on AQUAINT dataset with Wikidata corpus. Compared to other datasets, the advantage is 7% on average with F1 measure. Our method achieves best performance on precision among these datasets, which is 10% higher than DoSeR in ACE2004 with YAGO3 MSNBC data set. Our method also achieves 5% higher than DoSeR with recall on average.

5.3.2 Compared with entity resolution system

We test our method against Wikifier, DBpedia Spotlight, DoSeR, AIDA and WAT entity resolution systems, which utilizes Wikipedia corpora.

Table 2. Performance of comparison methods on F1, precision and recall

Data set	Corpus	Method	F1	P	R
ACE	YAGO3	Our method	**0.711**	**0.815**	**0.631**
		DoSeR	0.659	0.706	0.617
	DBpedia	Our method	**0.693**	**0.803**	**0.609**
		DoSeR	0.656	0.729	0.597
	Wikidata	Our method	**0.704**	**0.812**	**0.621**
		DoSeR	0.634	0.737	0.556
AIDA/CO-NLL	YAGO3	Our method	**0.710**	**0.821**	**0.625**
		DoSeR	0.646	0.742	0.572
	DBpedia	Our method	**0.712**	**0.817**	**0.631**
		DoSeR	0.647	0.742	0.573
	Wikidata	Our method	**0.706**	**0.831**	**0.614**
		DoSeR	0.630	0.753	0.542
AQUAINT	YAGO3	Our method	**0.704**	**0.827**	**0.613**
		DoSeR	0.651	0.749	0.576
	DBpedia	Our method	**0.715**	**0.824**	**0.631**
		DoSeR	0.661	0.759	0.585
	Wikidata	Our method	**0.726**	**0.831**	**0.645**
		DoSeR	0.647	0.742	0.573
MSNBC	YAGO3	Our method	**0.709**	**0.808**	**0.632**
		DoSeR	0.645	0.721	0.583
	DBpedia	Our method	**0.729**	**0.826**	**0.652**
		DoSeR	0.658	0.743	0.591
	Wikidata	Our method	**0.712**	**0.815**	**0.632**
		DoSeR	0.661	0.743	0.596

We use YAGO3, DBpedia, and Wikidata as core knowledge bases separately, and let compared solutions disambiguate all entities in DBpedia. Table 3, 4, 5 show the F1 values of comparison systems utilizing different data sets with YAGO3, DBpedia, and Wikidata corpus. Overall, the results of our solution achieve best performance in the comparison systems. Utilizing entities from YAGO3 corpus our solution outperforms WAT by 8.4% F1 values on average. Utilizing entities from DBpedia corpus our solution outperforms AIDA by 8.9% F1 values on average. Utilizing entities from Wikidata corpus our solution outperforms DBpedia spotlight by 7.4% F1 values on average.

Table 3. F1 values of comparison systems on four data sets with YAGO3 corpus

DATASET	**Our method**	**DoSeR**	**DBS**	**AIDA**	**WAT**
ACE	0.711	0.659	0.641	0.636	0.628
AIDA/CO-NLL	0.710	0.646	0.639	0.634	0.627
AQUAINT	0.704	0.651	0.612	0.631	0.625
MSNBC	0.709	0.645	0.614	0.621	0.619
Average	0.709	0.650	0.627	0.631	0.625

Table 4. F1 values of comparison systems on four data sets with DBpedia corpus

DATASET	**Our method**	**DoSeR**	**DBS**	**AIDA**	**WAT**
ACE	0.693	0.656	0.632	0.624	0.616
AIDA/CO-NLL	0.712	0.647	0.619	0.623	0.608
AQUAINT	0.715	0.661	0.617	0.627	0.625
MSNBC	0.729	0.658	0.621	0.619	0.611
Average	0.712	0.656	0.622	0.623	0.615

Table 5. F1 values of comparison systems on four data sets with wikidata corpus

DATASET	**Our method**	**DoSeR**	**DBS**	**AIDA**	**WAT**
ACE	0.704	0.634	0.621	0.632	0.613
AIDA/CO-NLL	0.706	0.630	0.619	0.615	0.621
AQUAINT	0.726	0.647	0.627	0.624	0.626
MSNBC	0.712	0.661	0.686	0.622	0.617
Average	0.712	0.643	0.638	0.623	0.619

We compared our solution to baseline entity disambiguation systems on ACE dataset with DBpedia corpus. The precision-recall curves in Figure 4 show that our method performs particularly well in the tail of high recall values. The mean average precision curve reflects the advantage of our method to comparison systems.

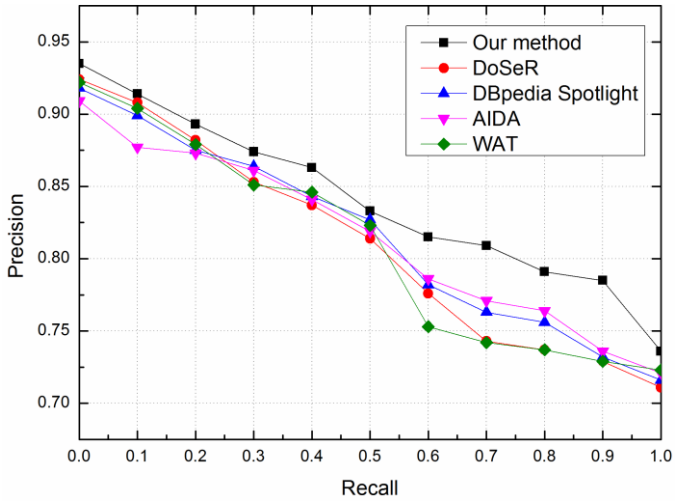

Figure 4. Experimental results on DBpedia with comparison systems: precision-recall curves

6. Conclusions

In the paper, a MLN knowledge graph model is proposed for solving entity resolution problem. We present a model that utilizes knowledge graph to represent the entity relationship between linked entities in the knowledge base. We utilize MLN to inference the inconsistent relationship within the knowledge base, and disambiguate the entities in the process of entity resolution. Our method outperforms baseline approaches for entity resolution by up to 7% with F1 measure. We also compare with the baseline entity resolution systems on three real knowledge bases. Experimental results demonstrate that our method achieves the best performance among baseline methods.

Acknowledgements

This research is supported by the National Natural Science Foundation of China, in part by Henan Province Science and Technology Committee, in part by Henan Province Educational Committee.

References

1. J. Biega, E. Kuzey, and F.M. Suchanek, “Inside YAGO2s: A Transparent Information Extraction Architecture,” in *proceedings of the 22nd International World Wide Web Conference* (WWW), pp. 325–328, Rio Janeiro, Brazil, May 2013
2. K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, “Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge,” in *proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (SIGMOD), pp. 1247–1250, Vancouver, Canada, June 2008
3. J.D. Carroll and J.-J. Chang, “Analysis of Individual Differences in Multidimensional Scaling Via an N-way Generalization of “Eckart-Young” Decomposition,” *Psychometrika*. vol.35, no. 35, pp. 283–319, July 1970
4. X. Cheng and D. Roth, “Relational Inference for Wikification,” in *proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pp. 1787–1796, Washington, USA, August 2013
5. P. Christen, “Automatic Training Example Selection for Scalable Unsupervised Record Linkage,” in *proceedings of the 12th Advances in Knowledge Discovery and Data Mining, Pacific-Asia Conference,* (PAKDD), pp. 511–518, Osaka, Japan, May 2008
6. W.W. Cohen, and J. Richman, “Learning to Match and Cluster Large High-dimensional Data Sets for Data Integration,” in *proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining* (SIGKDD), pp. 475–480, Alberta, Canada, July 2002
7. M. Cornolti, P. Ferragina and M. Ciaramita, “A Framework for Benchmarking Entity-Annotation Systems,” in *proceedings of the 22nd International World Wide Web Conference* (WWW), pp. 249–260, Rio Janeiro, Brazil, May 2013
8. S. Cucerzan, “Large-Scale Named Entity Disambiguation Based on Wikipedia Data,” in *proceedings of the Conference on Empirical Methods in Natural Language Processing Conference and Computational Natural Language Learning Joint Meeting following ACL 2007* (EMNLP-CoNLL 2007), pp. 708–716, Prague, Czech Republic, June 2007
9. X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun and W. Zhang, “Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion,” in *proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (SIGKDD), pp. 601–610, New York, NY, USA, August 2014
10. F. Erxleben, M. Günther, M. Krötzsch, J. Mendez and D. Vrandecic, “Introducing Wikidata to the Linked Data Web,” in *proceedings of the 13th International Semantic Web Conference* (ISWC), pp. 50–65, Riva del Garda, Italy, October 2014
11. P. Ferragina and U. Scaiella, “Fast and Accurate Annotation of Short Texts with Wikipedia Pages,” *IEEE Software*. vol. 29, no.

1, pp. 70–75, January 2012
12. T.N. Herzog, F.J. Scheuren and W.E. Winkler, "Data Quality and Record Linkage Techniques," Springer, 2007
13. 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 2011 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pp. 782–792, Edinburgh, UK, July 2011
14. T.G. Kolda and B.W. Bader, "Tensor Decompositions and Applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, May 2009
15. S. Lacoste-Julien, K. Palla, A. Davies, G. Kasneci, T. Graepel and Z. Ghahramani, "SIGMa: Simple Greedy Matching for Aligning Large Knowledge Bases," in *proceedings of the 19th International Conference on Knowledge Discovery and Data Mining* (KDD), pp. 572–580, Chicago, USA, August 2013
16. T. Lee, Z. Wang, H. Wang and S. Hwang, "Web Scale Taxonomy Cleansing," in proceedings of the VLDB Endowment 2011, vol. 4, no. 12, pp. 1295–1306, September 2011
17. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer and C. Bizer, "DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, August 2015
18. F. Mahdisoltani, J. Biega and F.M. Suchanek, "YAGO3: A Knowledge Base from Multilingual Wikipedias," *online proceedings of the Seventh Biennial Conference on Innovative Data Systems Research* (CIDR), Asilomar, CA, USA, January 2015
19. A. McCallum and B. Wellner, "Conditional Models of Identity Uncertainty with Application to Noun Coreference," in *proceedings of the 17th International Conference on Neural Information Processing Systems* (NIPS), pp. 905–912, Vancouver, Br. Columbia, Canada, December 2004
20. P.N. Mendes, M. Jakob, A. Garcia-Silva and C. Bizer, "DBpedia Spotlight: Shedding Light on the Web Of Documents," in *proceedings the 7th International Conference on Semantic Systems* (I-SEMANTICS), pp. 1–8, Graz, Austria, September 2011
21. D.N. Milne and I.H. Witten, "Learning to Link with Wikipedia," in *proceedings of the 17th Conference on Information and Knowledge Management* (CIKM), pp. 509–518, Napa Val. California, USA, October 2008
22. F. Piccinno and P. Ferragina, "From TagME to WAT: A New Entity Annotator," in *proceedings of the First International Workshop on Entity Recognition & disambiguation* (ERD '14), pp. 55–62, Gold Coast, Queensland, Australia, July 2014
23. 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* (HLT '11), pp. 1375–1384, Portland, Oregon, June 2011
24. M. Richardson and P. Domingos, "Markov Logic Networks," *Machine Learning*, vol. 62, no. 1, pp. 107–136, February 2006
25. S. Sarawagi and A. Bhamidipaty, "Interactive Deduplication Using Active Learning," in *proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining* (SIGKDD), pp. 269–278, Edmonton, Alberta, Canada, July 2002
26. P. Singla and P.M. Domingos, "Entity Resolution with Markov Logic," in *proceedings of the 6th International Conference on Data Mining* (ICDM), pp. 572–582, Hong Kong, China, December 2006
27. S. Tejada, C.A. Knoblock and S. Minton, "Learning Domain-Independent String Transformation Weights for High Accuracy Object Identification," in *proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining* (SIGKDD), pp. 350–359, Edmonton, Alberta, Canada, July 2002
28. R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis and L. Wesemann, "GERBIL: General Entity Annotator Benchmarking Framework," in *proceedings of the 24th International Conference on World Wide Web* (WWW'15), pp. 1133–1143, Florence, Italy, May 2015
29. W. Wu, H. Li, H. Wang and K.Q. Zhu, "Probase: A Probabilistic Taxonomy for Text Understanding," in *proceedings of the International Conference on Management of Data* (SIGMOD), pp. 481–492, Scottsdale, AZ, USA, May 2012
30. S. Zwicklbauer, C. Seifert and M. Granitzer, "DoSeR - A Knowledge-Base-Agnostic Framework for Entity Disambiguation Using Semantic Embeddings," in *proceedings of the Semantic Web. Latest Advances and New Domains: 13th International Conference* (ESWC 2016), pp. 182–198, Heraklion, Crete, Greece, May 2016