Entity Disambiguation Algorithm for Literature in Biomedical Field

Jing Wang[1,a,*], Jianzhuo Yan[1,b] and Ruying Lv[1,c]

[1]Faculty of Information Technology,Beijing University of Technology
[a] dkwangjing@emails.bjut.edu.cn, [b] yanjianzhuo@bjut.edu.cn, [c] 920774190@qq.com

Keywords: Domain literature, Entity disambiguation, Contextual characteristics, Probability model, Language model.

Abstract. Based on the requirements of knowledge learning and application in the domain of biomedical, a kind of entity disambiguation algorithm is proposed to solve the problem of entity ambiguity. Entity disambiguation is usually divided into two parts: candidate generation and entity disambiguation. In this paper, candidates of name mention are generated based on the knowledge base method and candidate entities are filtered based on the rule in the candidate generation stage, which ensures the recall rate of the candidate entity set and reduces the computational complexity and noise of the disambiguation stage effectively. In the stage of entity disambiguation, we propose a kind of entity disambiguation method based on probability model, estimating the probability that an entity becomes the target entity through the language model and selecting the entity with the highest probability as the target entity. The result of the method proposed in this paper shows the accuracy rate is 83%, higher than that of other algorithms. The method of entity disambiguation proposed in this paper is the best in the field of biomedical.

1. Introduction

In the Information Age, the information explosion has brought us numerous information, however, in the meantime, it also gives us a challenge to obtain target information swiftly and precisely. In biomedical domain, there is a huge amount of research achievements. The experiment design paradigm, data analysis method and experiment information included in these achievements have also been used in the unstructured biomedical literature which, therefore, is a vital knowledge source of medical domain research. In today's world, with the rapid growth in biomedical literature amount, literature mining is not only an effective method to collect and integrate professional knowledge, but also a feasible way to identify potential knowledge and advance the breakthrough of biomedical research. Biomedical literature mining is to obtain specific knowledge including research hot spots like entity, attribute, attribute value identification and semantic relation mining from the biomedical literature, among which the foundation is entity identification. Entity identification is to recognize terminologies in specific domains, which can generally be considered

into two categories: method based on heuristic regulation and method based on statistical.

Since biomedicine is an interdiscipline, the frequency of ambiguous terms' occurrence is rather high. For example, term "stroke" in the literature means a kind of cerebrovascular disease, but besides that, it may also carries a meaning of a band's or a place's name or other entity, which may be an obstacle to learn and apply the knowledge of biomedical literature. Thus, entity disambiguation is an essential procedure for entity identification of biomedical literature. If not, the accuracy of knowledge based application like research result and literature recommendation may be affected.

The present entity identification based on biomedical literature is mainly focused on identifying entities while the research about entity disambiguation is less concerned. Aiming at this point, this thesis explains a context-based entity disambiguation algorithm according to the analysis of data characteristic and the investigation of relevant algorithms. The entity disambiguation algorithm can identify the entity by analyzing the contextual situation of the mention though the context model, which has been proved to have a better result.

2. Relevant work

2.1. Research Status of Candidate Entity Generation

To generate a candidate entity is the key procedure of entity disambiguation, which can select the scope of candidate entity and add information to the second procedure. Then the target entities of the name mentions in the literature will be selected from the set of candidate entities. The generation of the candidate entity is important to the result of disambiguation because if the target entities are not contained in the candidate set at this procedure, no matter how advanced the algorithm is, we cannot find out a corresponding entity of a name mention; or if the candidate set is oversize, it will also affect the efficiency.

Therefore, it remains a major problem that how to shrink the candidate set without missing the target entities. Zhang[1].etc have adapted a supervised machine learning method which has taken not only text markup, verbal cues but also semantic information between texts and entities into consideration. But a further filtration is needed because the huge amount candidate entities generated by the referred method may increase the workload of entity disambiguation. And Nguyen[2] has used a series of heuristic regulations to generate candidate entities. He uses the frequency of a entity`s occurrence to select the candidates in the knowledge base. Zhou[3] has used the implicit semantic index to compute and rank the similarities between the source document and the Wikipedia text of the entity, thus selecting the top 50 as the final candidate entity. Some of these methods have achieved great recall rate but failed to meet the need of minimizing the candidate entities set.

2.2. Research Status of Entity disambiguation

Entity disambiguation plays an important role in the application of natural language processing, which can effectively solve the natural language processing tasks such as semantic network, information retrieval and knowledge recommendation. The essence of entity disambiguation is to

calculate the similarity between entity mention and candidate entity, and select the entity with the most similarity as the target entity of the name mention. In the literature[4] , the words with similar meanings usually appear in similar contexts, so the process of disambiguation is the process of selecting the entities with the largest context similarity from the candidate set, and the focus of disambiguation is how to calculate the context similarity.Mihalcea & Csomai[5]proposed an entity disambiguation algorithm based on the word pocket model (BOW), which uses the number of overlapping words in the context of the mention and of the entity as a measure of context similarity. Cucerzan[6],Bunescu&Pasca[7]and others enriched and advanced it. Bunescu &Pasca[7]measured the context similarity by cosine similarity, Cucerzan[6] used the dot product to measure the similarity. The above methods can only compare the exact match of the words in different texts. The selected feature is too limited which can not express the similarity between the mentions and the context of the candidate entity accurately .The subject model is that the entities appearing in the same text are about a topic, and there is a semantic association between them, revealing some kind of cohesion. So the entity disambiguation decisions are interdependent.

Zhang .etc[8] regarded the label of the entity category in the knowledge base as a marked theme, through which the theme model would be trained. Then the well-trained theme model will be used for the disambiguation task. Han and Sun [9] assumed that the query document consists of a number of abstract themes, each subject generates the target entity with a certain probability, and the target entity generates the query name and context with a certain probability. And finally the probability that the query document contains the entity would be the basis for sorting the entity. The result of this entity disambiguation method based on the text theme model is better when the query document contains a large number of entities and the subject is a single case. However, when the query document is short, the effect of disambiguation is reduced. Traditional entity disambiguation is generally aimed at web pages, news lines and other text rather than the field of literature research, because the terminology in biomedical field have fast updates, a variety of writing forms, high proportion of acronyms, nestling and other characteristics. The traditional entity disambiguation method cannot reflect the characteristics of terminology in the field of biomedicine. Therefore, we propose a kind of entity disambiguation method based on context characteristics of strong context correlation and context relevance among entities in biomedical field.

3. Entity disambiguation based on literature

The method proposed in this paper is divided into three parts: entity mention extraction, candidate entity generation and entity disambiguation based on probability model. In the generation of mention list, we mainly use stanford word segmentation tool to complete word segmentation processing for word data, then do some text preprocessing work like deleting stop words repeated words. The specific process is as shown below:

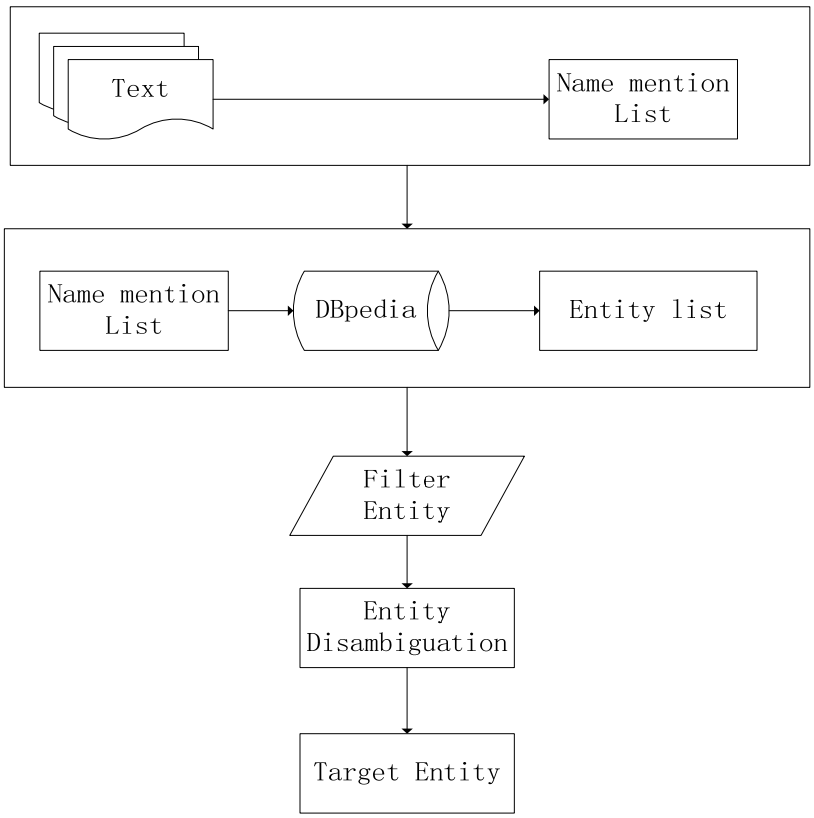

Figure1 The process of Entity disambiguation

3.1. Rule-based method for generating candidate entities

The generation of candidate entities has delineated the range of candidate entities from which the target entity of name mention are selected. The generation of candidate entities is important for the results of entity disambiguation. If the target entity can`t be included in the candidate entity set on the candidate generation stage, correct results won`t come out in the disambiguation phrase. On the contrary, if too many candidate entities have been selected in the candidate generation period, the required computing space and time would be larger and longer, reducing disambiguation efficiency. Therefore, the candidate generation is of great importance to the whole process of disambiguation. How to decrease the candidate entity set in the case of guaranteeing the inclusion of the target entity is the main problem of candidate generation.

With the term recognition technology, we identify the name mention in the literature as the named entity at first, and then use the candidate generation method to generate the candidate entity. There is a disambiguations_en.nt dataset which has mentions with multiple meanings in the DBpedia Knowledge Base. If we list all of their entities, the expression would often be entity name followed by "_ (disambiguation)". For example, the mention of the "Anxiety" in the data set is represented by Anxiety，Anxiety_(film)，Anxiety_(Munch), etc. So the disambiguation data set can provide all possible sets of entities. Candidate entities selected by this method have a large coverage, but their number is too large, so they need to be screened.

3.1.1. Screening by entity popularity

Often, the probability of finding an unpopular entity in the literature is relatively small. Therefore, we use the entity's reputation information to filter the candidate entity. For a mention m_0, the corresponding set of entities is expressed as $E=\{(e_1,c_1),(e_2,c_2),...(e_n.c_n)\}$. The e_i in the set indicates an entity of the mention, c_i refers to the number of links of e_i in the knowledge base, so the popularity of an entity can be expressed as:

$$p(e_i)=\frac{c_i}{\sum_{i=1}^{n} c_i} \quad (1)$$

3.1.2. Screening by text similarity

If we want to calculate the similarity between name mention m and entity e, we can select the words near the mention m as the background text, the interpretation document of the mention m in the knowledge base as an entity document, abstracting the two texts into word vectors $\vec{V}_1=\{x_1,x_2,...,x_n\}$ and $\vec{V}_2=\{y_1,y_2,...y_n\}$ respectively, the similarity between mention m and entity e can be calculated with the following formula:

$$Sim(m,e)=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{(\sum_{i=1}^{n}(x_i)^2)}\times\sqrt{(\sum_{i=1}^{n}(y_i)^2)}} \quad (2)$$

Where the values of x_i and y_i depend on their TFIDF value.

3.1.3. Screening by the numbers of candidate entities

Filtering by the numbers of candidate entities means that no new candidate is added after the number of entities in the candidate entity set exceeds a certain threshold. The specific algorithm flow is as follows:

Table 1:Algorithm of candidate entity generation

Input: name mention n, similarity threshold value ST, threshold value of the number of candidate entities NT
Output: the set of candidate entities C
Step 1: Set the candidate entity set C to an empty set

Step 2: when $Q_N \neq \phi$, for(n in Q_N) Q_E=SortbyPop(E(n)). Then add the corresponding candidate entities of mention n to the sequence QE in the order of popularity.

Step 3: define the count variable i=0

*Step4:*while $Q_E \neq \phi$ and $i<NT${

for(e in Q_E){
if {
Sim(m_0,e)>ST
}
$C \leftarrow e$
$i=i+1$
}
The algorithm ends

As we can see from the algorithm above, we rank the candidate entities according to an entity`s popularity in the knowledge base in the process of selecting a candidate entity set, add the sorted initial candidate entity collection to the queue, and select the candidate entity from the queue one after another. Then we compare the similarity between the candidate entity e and the name mention m0, adding the entity with a similarity greater than the threshold to the final set of candidate entities until the number of entities in the candidate entity set reaches the candidate entity threshold. In this paper, the similarity threshold is set to 0.7 and the candidate entity threshold is set to 30.

3.2. Entity disambiguation based on the feature of context

For each target entity of the name mention that is to be disambiguated in the disambiguation task of named entity, built on probabilistic model, we present a method which is mainly based on language model to fully describe the information of the candidate entity. The language model is a text probability distribution, that is to say, the article consists of words, and it`s a random event for the words to appear in order in the article, and the occurrence of this random event obeys a certain probability distribution, i.e the language model. We can estimate a random event (That is, the probability of occurrence of the article constituted by words arranged in a certain order) through the language model. The event is called the probability that this article is generated by the language model. As is mentioned by HanX[10], an entity's context knowledge can usually be expressed in the following one-dimensional language model:

$$M_e(t) = \{p_e(t)\} \quad (3)$$

$P_e(t)$ refers to the frequency at which each word t in the context of an entity appears. $M=\{m_1,m_2,...,m_k\}$ is defined to be a collection of name mention abstracted from the biomedical literature, the context of each mention m in M is expressed as $c.E=\{e_1,e_2,...e_m\}$ is defined as a set of entities in the knowledge base which has a descriptive text for each entity. We describe the text as the context of the entity, expressing as e. In this paper, variate $p(e|c)$ represents the entity`s contextual knowledge, it`s value would be greater when the correlation between the named mention and the entity is closer, vice versa. The word collection contained in the context where the name mention lies is regarded as a corpus of the one-dimensional language model, it is noted as c, and the word order of the entity context is marked as $e=t_1t_2...t_k$. As a result, the the entity`s contextual knowledge $p(e|c)$ is calculated as below:

$$p(e \mid c) = p(t_1 t_2 ... t_k \mid c) \quad (4)$$

Each term t in the context of entity is independent, so the above formula can be expressed as:

$$p(e \mid c) = p(t_1 \mid c)\, p(t_2 \mid c) ... p(t_k \mid c) \quad (5)$$

The key of calculating an entity`s contextual knowledge can be transferred to calculating the frequency of each term, i.e to calculate $p(t_i|c)$. Using a method based on context similarity, $p(t_i|c)$ can be expressed as below:

$$p(t_i \mid c) = \frac{count_e(t_i)}{\sum_t count_e(t)} \quad (6)$$

$count_e(t_i)$ is the frequency at which the entity`s contextual term t_i appears in a corpus.

4. Experiment and evaluation

4.1. Experimental data

The experimental knowledge base is DBpedia[13] dataset which is structured information extracted from Wikipedia by the DBpedia community. DBpedia is a multi-lingual knowledge base among which the English knowledge base has the most abundant sources. The English dataset of DBpedia in 2015 consists of more than 500 million facts (RDF triad) and 4.5 million affair. It has a wide coverage area, including people, creatures, medicine and so on. Therefore we choose the English dataset of DBpedia as our knowledge base and 500 abstracts of biomedical literature from PubMed as a literature data set.

4.2. Results and discussion

4.2.1. A discussion about the results of candidate entity generation

We compare the two methods of generating candidate entities here, namely, the mentioned generating method(Heuristic)based on heuristic rules and the direct matching method (DMatch). The following table shows the different recall rate of the candidate generation method and the average size of the candidate entity set.

Table 2 Comparison of different candidate generation methods

name	Recall rate	Average number of candidates
Heuristic	0.935	12.6
DMatch	0.807	125.2

As we can see from the table above, the recall rate of our method is higher than the direct matching method (DMatch). On the other hand, the scale of the candidate entity set generated by our method is smaller than the direct matching method. At present, few studies have compared recall rates and average number of candidates at the same time as we do. The methods proposed in this paper can be divided into the following strategies:

- DMatch: Match the name mention directly in the knowledge base
- fByPop: screen the entity by its popularity
- fBySim: screen the entity by the similarity between candidate entities and name mention

The above strategies are added to the candidate entity generation module in turn in this experiment to evaluate the role of each part. As a result, the recall rates and the average number of candidates are as follows:

Table 3 Results of different strategies

	Recall rate	Average number of candidates
DMatch	0.807	125.2
+fByPop	0.945	15.6
+fBySim	0.935	12.6

As we can see above, the result of DMatch consist of large quantities of candidate entities. Using these strategies together, we get a collection of candidate entities with high recall rates and small average numbers.

In recent years, some researchers have referred to the methods of candidate entity generation in the process of studying entity disambiguation, among which the three main methods are: the supervised machine learning method proposed by Zhang, expressed as ML. The method adopted by Nguyen who generates candidate entities with a series of heuristic rules and selects an entity by its frequency in the knowledge base, recorded as HL. And Zhou`s method called as RL which means to calculate and rank the similarity between the source document and the Wikipedia text of the entity with the implicit semantic index, thus selecting the top 50 as the final candidate entities. In the data

set of this paper, we compare the above methods with our own method. The comparison results are as follows:

Table4 Results of different methods

name	Recall rate	Average number of candidates
ML	0.965	95.8
HL	0.903	57.2
RL	0.912	32.5
Our own method	0.935	12.6

By contrast, we found that the average number of candidates of ML is large though the recall rate is high. The collection of candidate entities is too large to meet the requirement of minimizing the set of candidate entities. On the contrary, HL and RL can provide average number of candidates much smaller, but the recall rate is low. However, the average number of candidates of our own method has been significantly reduced, and the recall rate is higher than 0.93, which reflect a better application effect.

4.2.2. Discussion on the results of entity disambiguation

We compare our method with the BOW algorithm, the algorithm based on semantic similarity (C &R), and the method based on graph (Graph). The results are as shown below. The accuracy of our method is much higher than that of others, indicating that our method is more suitable for the disambiguation tasks of dataset.

Table4 Accuracy of different algorithms

name	accuracy
BOW	0.35
C&R	0.60
Graph	0.74
Our	0.83

5. Conclusion

In the light of the features of entity disambiguation of biomedical literature, we have proposed an algorithm for biomedical field which can be divided into candidate entity generation and entity disambiguation. We aim to generate a small set of candidate entities with a high recall rate during candidate entity generation, and have adopted a method based on probabilistic model to complete entity disambiguation of literature in the biomedical field. Our experimental results show that our method, compared with the traditional method of entity disambiguation, can achieve better results in the entity disambiguation of literature. We intend to study the disambiguation of the acronyms in the literature next to make our research more complete.

Acknowledgements

The work is supported by National Key Basic Research Program of China(2014CB744605), National Natural Science Foundation of China (61272345),Research Supported by the CAS/SAFEA International Partnership Program for Creative Research Teams, the Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (25330270).

References

[1]Zhang ,W., Sim ,Y.C., Su ,J., Tan,C.L. Entity linking with effective acronym expansion, instance selection and topic modeling[C]//Proceedings of the Twenty-Second international joint conference on Artificial Intelligence, 2011: 1909-1914.

[2]Nguyen,H.T., Cao,T.H. Named entity disambiguation: a hybrid approach[J]. International Journal of Computational Intelligence Systems, 2012, 5(6): 1052-1067.

[3] Zhou,S., Inui,C.K.K. Exploring Linguistic Features for Cross -document Named Entity Disambiguation[J]. 2015.

[4]Miller,G.A.,Charles,W.G. Contextual correlates of semantic similarity[J]. Language and cognitive processes, 1991, 6(1): 1-28.

[5]Csomai,A..,Mihalcea,R. Linking documents to encyclopedic knowledge[J]. IEEE Intelligent Systems, 2008, 23(5).

[6]Cucerzan,S.Large-Scale Named Entity Disambiguation Based on Wikipedia Data[C]//EMNLP-CoNLL. 2007, 7: 708-716.

[7]Bunescu,R.C., Pasca,M. Using Encyclopedic Knowledge for Named entity Disambiguation[C]//EACL. 2006, 6: 9-16.

[8]Zhang,W.,Su,J., Tan,C.L. A Wikipedia-LDA Model for Entity Linking with Batch Size Changing Instance Selection[C]//IJCNLP. 2011: 562-570.

[9]Han,X.,Sun,L. An entity-topic model for entity linking[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012: 105-115.

[10]Han,X.,Sun,L. A generative entity-mention model for linking entities with knowledge base[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 945-954.

[11] Tang ,B.R..Named Entity Disambiguation Based on Wikipedia [D]. Beijing University of Science and Technology, 2011.

[12]Yang,G., Liu,B.Q., Liu,M.Graph-based Method for Named Entity Disambiguation[J] .Journal of Intelligent Computer and Application, 2015 (2015 05): 52-55.

[13]Li,X.M,Zhang,J. Z.Study on DBpedia in Knowledge Community Environment [J]. Library, 2013 (4): 27-30.

[14]Guo,Y.H.Research on context-based entity linking technique.[D].School of Computer Science and Technology,2014