An Entity Linking Method Based on Entity Category and Word Embedding

HaiBo Ma, Yue Li and Yang Li
Beijing University of Technology, 100124, Beijing
email: bjut@foxmail.com

Abstract. Named entity linking is a process of linking a given reference in a document to a knowledge base. In natural language processing, entity linking can enhance the computer's understanding of unstructured text data. Applying traditional entity linking methods, especially entity linking methods for person names and organization names, has its limitations. Similar vocabulary as an entity to be linked is difficult to make full use of its contextual semantic information for ambiguity elimination. This paper makes full use of the entity's category attribute and the semantic information contained in the context to design an entity linking method based on entity category and semantic word embedding. First, training text classification model based on corpus to obtain entity attributes. Then the semantic feature is extracted by the word vector template to perform entity disambiguation through the semantic classification model. Finally, the results of the entity linking are predicted by means of model ensemble. Experiments show that the accuracy of the method after fusion on the entity linking dataset has improved.

1. Introduction
In the era of rapid development of the Internet, the rapid growth of text data has caused the problem of information explosion. Strengthening the analysis and understanding of unstructured text data by computer through the relevant methods of natural language processing is of great value to the use of this data. In natural language, nouns often have a polysemy. In this case, entity linking can eliminate ambiguity and improve computer text comprehension. An entity linking task refers to extracting named entities in a document, including names of people, organizations, etc., and linking them to unambiguous entities in the existing knowledge base. The main task of the entity linking is to take two steps, firstly identify the reference item to be linked in the text, obtain the candidate list in the corresponding knowledge base, and then link the reference to the unambiguous entity in the knowledge base based on the entity disambiguation algorithm.

At present, the methods of entity linking mainly include a machine learning classifier based method, a topic model based method, a graph model based method, and a vector space model based method. As stated in the reference [1], the classifier-based method is to use the multiple entities corresponding to the reference item as the result of classification, use the classifier to train and predict the result, but this method generally uses noun co-occurrence as a feature, lacking contextual semantic information dataset noise impact big. As stated in the reference [2], the topic model performs entity disambiguation based on the document theme, but it is difficult to extract deeper contextual semantic information of higher dimensions by linking only through the topic. As stated in the reference [3], the graph model-based method constructs the semantic network to extract the relevant features to link the
entities. However, if the amount of data is large, the computational cost will be doubled. As stated in the reference [4], the vector space model converts the context of the referential terms into vectors, and the entity disambiguation is performed by the similarity calculation. This method does not fully apply the important information in the external corpus and knowledge base.

The main contribution of this paper is to propose a named entity linking method based on entity attributes and word embedding, mapping context nouns in the text and corresponding entities to be linked to the same semantic topic space. In this way, the classification model is used to semantically disambiguate the entity. The innovation is that from the perspective of deep learning, the entity disambiguation method is proposed in the way of classifier classification prediction. Not only can the context semantic information be fully utilized, but also the semantic classification information of the entity can be used for disambiguation, and the accuracy rate can be improved. The structure of this paper is as follows: Section 1 is introduction. Section 2 introduces the method proposed in this paper; Section 3 is the entity linking part; Section 4 is the experimental part; the last part is the conclusion and outlook.

2. Ensemble classification model

2.1. Semantic word embedding

Semantic classification model is in Figure 1. The neural network part of the semantic classification model adopts the CBOW model [5]. CBOW is a three-layer neural network model, from left to right, the input layer, the hidden layer and the output layer. The basic idea is to train each words are mapped into a K-dimensional real number vector containing semantic and grammatical information, K is an optional parameter, and the distance between the vectors such as Euclidean distance and cosine similarity is used to judge the semantic similarity between them. The model is to model the language model and obtain the representation of words in the distributed vector space while modelling. By solving the objective function by stochastic gradient descent, the V words in the corpus can be represented as distributed vectors with deep semantic features. Its corresponding objective function is shown below.

$$l(\theta) = \arg \max \frac{1}{v} \sum_{j=1}^{v} \left( \sum_{i=1}^{n} \log p(w_{ij} | context_{ij}) \right)$$

2.2. Extract features

Feature selection for a given training data set, \(D = (d_1, d_2, \cdots, d_n)\) is used to represent each piece of text in the training data set, and \(S = (S_1, S_2, \cdots, S_n)\) indicates that the text is a corresponding collection of unambiguous entities that have been linked to the knowledge base. Based on the assumption that all nouns in a text, including entity references, are located in similar semantic spaces, extract the nouns in the training set, and obtain the word vector template by the method in Section 2.1, and represent the
extracted nouns as distributed vector, get the noun vector set N. The distributed vector contains deep
semantic information, and the set N is clustered by k-means [6], and k center points P = (P₁, P₂, ⋯, Pₖ)
are obtained as k features where k is k -means clusters the number of cores. At the same time, the
category label of each word in the set N is obtained by calculating the distance of each word to k
center points.

The training data characterizes the label of each noun obtained in the previous section. The text d_i
in the set D can be represented as a k-dimensional vector. The frequency at which each noun category
in d_i is present is used as the weight on the dimension feature. Select k=10 to select the 10-dimensional
feature. "Apple, Google and Microsoft are the world's largest technology companies, often compared
by people." In this sentence, the clustering label corresponding to Apple and Google is 5, and the
corresponding clustering label of the technology company is 1, people correspond the clustering label
is 9. So this text can be represented as (0,1,0,0,1,0,0,0,1) corresponding to this text, the unambiguous
entity that has been linked to the knowledge base is "Apple" Find the corresponding vector from the
vector template, calculate the closest category to Apple by cosine similarity, and find that Apple
belongs to category 5, so Apple can be represented as a vector (0,0,0,0,1,0,0,0,0). Through the above
process, we can represent the text in the training set and the corresponding entity to be linked as a k-
dimensional vector pair, and k is the number of selected features.

2.3. Semantic word embedding classification model
Related work shows that in practical applications, logistic regression classifiers are similar to SVM
and random forest classification models [7]. The logistic regression classifier algorithm has the lowest
complexity. Therefore, in this part we use a logistic regression classifier to build a classification
model. Use D(d₁*, d₂*, ⋯, dₙ*) to represent the characterization of the text set, and use S(S₁*, S₂*, ⋯,
Sₙ*) to represent the characterization that a collection of unambiguous entities that correspond to the
text and are linked to the knowledge base. Where d_i* and S_i* are both k-dimensional vectors. After
characterizing the training data and presenting it in a vector, the problem can be transformed into a
multi-class problem. The significance of this is that it not only utilizes the relationship between the
text to be linked entity and the nouns in the text, but also utilizes the semantic features of the entity's
word vector.

Let x= d_i* and y= S_i*. Then (x, y) is an observation sample, and the feature selected in Section 2.2
can be obtained, which ranges from [0,1,2,⋅⋅⋅, k−1]. The conditional probability can be expressed as
follows.

\[ P(y = c_i|x) = \frac{e^{\theta_{c_i}(x)}}{1+\sum_{j=1}^{k-1} e^{\theta_{c_j}(x)}} \]  \hspace{1cm} (2)

Solving the model by constructing a likelihood function, and the sample above, yields the
likelihood function as follows. The likelihood function is solved by the gradient descent method,
and the trained semantic model of word vector is obtained.

\[ L(\beta) = \sum_{j=0}^{k-1} \sum_{l=1}^{n} y_{ij} \log(\pi_j(X_i)) \]  \hspace{1cm} (3)

2.4. Entity attribute model
Based on the assumption that entities often recur in the same domain, this paper uses the classification
properties of entities to disambiguate entities. The text categorization task is mainly to let the
computer automatically divide the text into pre-defined categories. Because the Text CNN [8] model
has high classification accuracy and high training efficiency, the text classification model adopts the
Text CNN model. The CNN model mainly includes a convolutional layer, a pooling layer, an
activation function, and a fully connected layer. The external corpus required for training comes from
THUCNews. Training through backpropagation and annotated data classification to obtain an entity
cost context classification model. The entity is linked based on the results of the entity context
classification and the category attributes of the entities in the knowledge base.
3. Entity linking

3.1. Task details

An entity linking is the process of linking a given entity's reference in the text to an unambiguous entity in the knowledge base. In this paper, three features are selected for entity disambiguation, word embedding classification model (WECM), entity category feature and feature of entity popularity [9]. The entity linking process is expressed as follows. Where \( E \) represents the final score of the candidate entity, \( e_{ij} \) represents the candidate entity corresponding to the \( m_i \), \( f_{tc} \) represents entity classification model, \( f_{ef}(e_{ij}) \) represents the candidate entity popularity score, and \( A \) \( B \) represents the weight of the semantic classification feature of the word vector.

\[
E = \arg \max \left[ A \cos \left( \text{WECM}_{m_i}, e_{ij} \right) + B f_{tc}(m_i, e_{ij}) + (1 - A - B)f_{ef}(e_{ij}) \right] 
\]  
(4)

The whole process can be divided into three parts: entity referral standardization, candidate entity expansion and entity disambiguation. Many entities in the text have a number of different names, some are aliases such as magicians, nicknames such as Dayao, some are part of the full name or abbreviations such as PKU. Therefore, it is first necessary to map the representations that appear in the text to a standard expression. Specifically, a synonym vocabulary [10] is constructed to solve this problem. Among them, the Key value represents the irregular reference of the entity, and the Value represents the standard entity. After standardizing the entity references, a list of candidate entities needs to be built for the named entities to be disambiguated. This paper constructs an ambiguous words table that stores the standard form of the entity and its corresponding list of unambiguous entities. The entity popularity feature is measured by the number of times the entity appears, and the weight of the entity popularity is set according to experience. An example are shown in Table 1.

| Unambiguous entity | Entity frequency | Weights |
|--------------------|------------------|---------|
| Apple(enterprise)  | 168              | 1.0     |
| Apple(fruit)       | 15               | 0.8     |
| Apple(figure)      | 2                | 0.6     |

3.2. Entity linking

A text entity linking method based on word vector semantic classification, the input time text and its corresponding entity to be linked, corresponding to the unambiguous entity linked to the knowledge base. In the first step, according to the synonym table in the knowledge base, the description of the reference in \( M \) is standardized, and the standardized reference set is obtained. In the second step, \( E = (e_1, e_2, \cdots, e_n) \), and the candidate entities in \( e \) are sorted by popularity. When \( |e| = 0 \), it means that there is no corresponding entity in the knowledge base, that is, the reference \( m_i \) is unlinkable, and the label NIL is returned; when \( |e| = 1 \), the unique candidate entity is directly returned as the final link entity; when \( |e| > 1 \), perform steps 3 and 4. In the third step, the semantic classification features of each candidate entity are obtained by calculating the cosine distance of each candidate entity to the tag. In the fourth step, the weighted sum of the two features of each candidate entity is calculated according to the equation, and the candidate entity with the highest result is output as the final link entity. If the weighted sum is less than the threshold, the flag NIL is returned.

4. Experiment

4.1. The data set

The data used in this paper to establish synonym tables, ambiguous vocabularies, entity popularity tables, and trained word vector templates are all Wikipedia [11], and the data version used is Chinese
Wikipedia in 2015. The data is extracted from the knowledge base through rules and statistics are obtained, and the data size is as shown in Table 2.

4.2. Comparison experiment
The comparison experiment is carried out to verify whether the algorithm is effective. This part selects the entity link method based only on the entity classification attribute, and the entity link method based only on the semantic vector classification model, and compares the current method after the model fusion. During the experiment, the same resource template and preprocessing were used for all methods. Firstly, the multi-class regression model is trained by training set data, and then the model set is used to obtain the optimal weight $\alpha=1.58$, $A=0.52$, $B=0.27$. Finally, the method is verified in the test set. Select Precision, Recall, and F1 as the evaluation indicators. Among them, the in-KB part indicates the accuracy rate of the included entities in the knowledge base, and NIL indicates the entity link accuracy rate not included in the knowledge base. The results of the comparative experiments are shown in Table 3.

### Table 2. Experimental data

|                | Unambiguous entity | Data amount |
|----------------|--------------------|-------------|
| Synonym table key number | | 531526 |
| Synonym table value number | | 183727 |
| Ambiguous words table key number | | 19284 |
| Ambiguous words table value number | | 241467 |

### Table 3. Experiment result

|               | in KB | NIL |
|---------------|-------|-----|
| Model         | Precision | Recall | F1 | Precision | Recall | F1 |
| WECM model    | 0.7927 | 0.8488 | 0.8198 | 0.8871 | 0.8648 | 0.8758 |
| Entity attribute model | 0.7341 | 0.7942 | 0.7884 | 0.8143 | 0.8194 | 0.8361 |
| Ensemble model | 0.8137 | 0.8593 | 0.8358 | 0.9134 | 0.8763 | 0.8949 |

In the experimental results, the bold portion indicates the method herein. Experiments show that the fusion method is superior to the other two algorithms in terms of accuracy and recall rate, especially the accuracy rate is significantly improved. The method of this paper is mainly a method of combining classification attributes and semantic word vectors to fully apply context semantic information. The relationship between the semantic classification feature of word vector and the number k of clustering features. Only the WECM model is used for entity linking, and different k values are selected to observe the relationship between the model and k. When k=10, the model obtains the highest F1 value. Between 5~15, the change of k value has little effect on the prediction efficiency of the model. However, when k=20, the F1 value of WECM drops significantly. By counting the number of nouns contained in each text in the evaluation data, it is found that there are an average of 8 nouns in each text. The analysis considers that when k=20, the decrease of F1 value is due to excessive selection of features and sparse training data.

5. Conclusion and outlook
Based on the assumptions that entities appear in the same field, and the nouns and contexts in the text have similar semantics, this paper proposes the idea of semantic disambiguation of text entities by using word vector semantic classification, and designs a complete entity linking method and performs on the data. verification. The experimental results show that using the entity link method based on classification attribute and word vector proposed in this paper, the link effect uses the effect of word
vector and entity attribute alone, and the link accuracy rate is significantly improved. The follow-up work is mainly focused on two points. One is to combine the word vector and the graph model for physical linking.

Reference lists
[1] Gattani A, Lamba D S, Garera N, et al. Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach[J]. Proceedings of the VLDB Endowment, 2013, 6(11): 1126-1137.
[2] 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.
[3] Han X, Sun L, Zhao J. Collective entity linking in web text: a graph-based method[C]//Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011: 765-774.
[4] Monteiro L B, Li W, Saleh A A. An Approach of Vector Space Model to Link Concrete Concepts with Wiki Entities[C]// IEEE International Conference on Computer & Information Technology; Ubiquitous Computing & Communications; Dependable. 2015.
[5] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv: 1301.3781, 2013.
[6] Hartigan J A, Wong M A. Algorithm AS 136: a k-means clustering algorithm. Journal of the Royal Statistical Soci-ety — Series C (Applied Statistics), 1979, 28(1): 100–108
[7] Fernandez-Delgado M, Cernadas E, Barro S, Amorim D. Dow need hundreds of classifiers to solve real world classification problems Journal of Machine Learning Research, 2014,15: 3133−3181
[8] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. 2012.
[9] Shen W, Wang J Y, Han J W. Entity linking with a knowl-edge base: issues, techniques, and solutions.IEEE Trans-actions on Knowledge and Data Engineer,ing, 2015,27(2):443–460
[10] Guo Y H, Qin B, Li Y Q, Liu T, Lin S. Improving can-didate generation for entity linking. In: Proceedings of the18th International Conference on Applications of Natural Language to Information Systems. Salford, UK: Springer, 2013. 225–236
[11] Wikipedia[Online], available:http://download.wikipedia.comzhwikilate-stzhwiki-latest-pages articles.xml.bz2. Octo-ber 31, 2015