Improving Entity Linking by a Novel Pipeline Method For Chinese Short Text

Feiran Zhu

Zhejiang University, College of Computer Science and Technology

Abstract

Entity linking (EL) is a quite difficult task which involves aligning mentions in the text to the corresponding knowledge base when the entity in the knowledge base has the same meaning as mention. In CCKS-EL task, there is another extra task, when the mention can not be linked to the knowledge base, it's defined as a NIL entity and should be given its category. EL is one of the basic task in the Nature Language Processing (NLP) , which can be used for knowledge base question answering and knowledge reasoning. Due to the diversity of mention and lack of semantics and context information in Chinese short text, it brings great challenge to EL task. In order to solve this problem, we propose a novel pipeline method. First, we consider EL task as a point-wise ranking problem, if the mention can not be linked to the knowledge base, then we use a classification model to classify the mention type as a nil type. In the evaluation of DuEL2.0 dataset, out model achieves 89.266 in the F1 score in the third place.

Keywords: Entity Link, Deep Learning, Bert, NLP

1. Introduction

Entity linking, which map the mention detected in the text to a given knowledge base (KB, e.g., Wikipedia), plays a vital important role in many areas, such as question answering, standard information extraction, and an accurate entity linking system is quite crucial for the upper-level tasks. The knowledge base is a semantic network which is designed to describe conceptual entities in the objective world and their relationships. One entry in KB is a series of semantic relations description and some alias with the same meaning of the entity. The richer the KB is, the greater the value of the EL task, and the difficulty of this task also increased. In recent years, English EL technology is rapidly developing and a lot of entity linking system have been created such as Dbpedia Spotlight [1] developed by Dbpedia.org.

Traditional EL task is mainly focused on long document, which provide sufficient semantic information to assist disambiguation. In contrast, there are many difficulties in Chinese short text EL task as follows: (1) Chinese is rich in semantics, the same words may have very different semantic information especially in short texts; (2) The colloquialization of short texts is more serious, making entity disambiguation more difficult; (3) The mention which can not be linked to the KB should be assigned a nil type, increasing the difficulty of the task. As is talked above, we can not simply apply EL method used in long text.

In this paper, we propose a two-stage pipeline method for the EL task in DuEL2.0 dataset. For the first EL task, we recognized it as a point wise ranking problem using a BERT model pretrained by Zhuiyi Technology called Semi-BERT. The input of first model is the concatenate of the query text and the all descriptions in the candidate KB entry and the output is a rank score for the candidate entry. Then we sort the candidates by the score, if the score of the candidate entity with the highest score is greater than the threshold, we assign the mention to this candidate entity with the highest score in the KB, else we use a classification model to classify the type of this mention. We believe that the type of mention will also have a great impact on the accuracy of entity links, so we designed an auxiliary task to predict the type of currently linked entities. Different from the general classification task, we spliced the category information with the short text, and divided the entity types through the logits in the category information, so as to obtain more semantic information.

There are three main contribution of this paper as follows:

1. The pretrained model BERT using UniLM[2] mask method is skillfully deployed in both EL task and mention classification task, which can dig out the semantic information of the short text;

2. In the process of entity linking, we introduce an auxiliary task for the short text to predict the type of the mention, which can be useful for the model to catch the interactive information of query and KB entry description;

3. Draw on the NL2SQL approach, we introduce an novel method to classify entity type, which significantly improves the entity nil type classification task.

2. RELATED WORK

EL is an important step in the population of the knowledge base. With the continue expansion of the knowledge base, EL technology has received more and more attention. At present, there are three main kinds of entity linking methods, i.e., rank based methods, binary classification based methods (such as pair wise ranking methods), and graph based methods. In a binary classification based method, the correlation features of mentions and candidate entities are typically used to train a binary classification model.

The traditional methods of EL are mostly based on rule-based judgement such as dependency tree. Cheng et al. [3] manually designed features to evaluate the local context compatibility and document-level global coherence of referent entities. However, the rule-based methods are not sufficiently generalized and not suitable for short text. As search technology improves, the demand of EL for short text has greatly increased.

Cornolti et al. [4] put the query statement into the search engine such as Wikipedia to get some short text that related to the query to add some semantic information. This idea slightly alleviates the problem of incomplete context and semantic information. Deepak et al. [5], similarly, proposes a method to put short text into Wikipedia and obtain the top k relevant sentences as candidate entities by a simple fasttext model. Then the entities are extracted from these sentences as extra feature to establish the characteristics of connection with entry in KB. Their method does work in the EL task in short text but suffer from quite a lot error propagation.

Pan et al. [6] build features such as vocabulary features, word categories, and name entities categories, and then use a SVM classifier for the binary classification task. Sun et all. [7] proposed to use a deep learning approach to obtain semantic representation of mentions, context, and entities. Ganea et al. [8] achieved entity disambiguation in virtue of entity embeddings and local context window attention mechanism. It substantially outperforms the traditional methods on standard benchmark(e.g., AIDA-CoNLL). However, all of them neglect the latent entity type information in the immediate context. It's our assume that this may sometimes cause the models link mentions to the incorrect entities with the wrong type. To verify this, we conduct error analysis and do some experiments on the DuEL data set, it turns out the information of entity type do effect the accuracy of the entity linking task.

With the emergence of pre-trained models, just like BERT or ERNIE [9], and continue to prove its effectiveness, the pre-trained model gradually dominates various NLP tasks. The model like BERT can effectively capture the context and semantic information in the text, especially in short text and medium text. Therefore, it is a reasonable solution to use a pre-trained model on short text ranking and matching task.

3. BACKGROUND

- Formally, given a short sentence consists some entity mentions which have already been tagged m_1, ..., m_n, the goal of an entity linking system is to assign each m_i an entity e_i in the given KB or predict that there is no corresponding entity in the KB(i.e., e_i=NIL) and the type of this mention should be given.

- As entity linking model integrating both local and global features can be formulated as a conditional random field. Formally, we can define a scoring function g to evaluate the entity assignment $e_1, ..., e_n$ to mentions $m_1, ..., m_n$ in a short sentence S.

$$g\left(e_1, \ldots, e_n \mid S\right)=\sum_{i=1}^n \Psi\left(e_i \mid S\right)+\sum_{j \neq i} \Phi\left(e_i, e_j \mid S\right) \tag{1}$$

4. PIPELINE MODELS

4.1. Entity Linking for short text

In a first step, we consider entity linking task as a point wise ranking task. We sampled all the entities in the knowledge base that are the same as mention, the sampling rule is that as long as the subject or alias in a certain entry of the KB is the same as the mention, it will be added to the candidate list. In the training set, we only sampled three negative examples paired with a positive example, but in the test set we sampled all the candidate examples. The reason for this is to increase the recall rate as much as possible while reducing the training cost.

Inspired by the methods of information retrieval, we concatenate the query context and the corresponding entry description in the knowledge base. The construction of description is composed of information splicing of triples. Then we calculate the similarity of the two part. The candidate knowledge base entry of the mention will be sorted by the similarity score in reverse order. If the candidate entry with the highest score is lower than the handcrafted threshold, the mention will be moved to the classification model to get a nil mention type, else, the mention will be linked to this candidate knowledge base entry.

We use Semi-BERT which is pretrained by Zhuiyi Technology, which is quite suitable for the similar sentence judgment tasks. Then, we fine-tune the model in the downstream task: DuEL2.0 dataset. Figure 1 shows the overview of our model. The model consists of three module: an input layer, a BERT encoder layer, and a classification layer. The first CLS token in BERT model contains the global semantic and contextual information, so we use the output of this token to represent the interactive information between two sentences. Also, there may be multiple different mentions in the short text, so we concatenate the

CLS feature with feature vector of the start and end position of the corresponding candidate entity, then through the fully connected layer with the sigmoid activation function, the probability of the candidate entry is obtained.

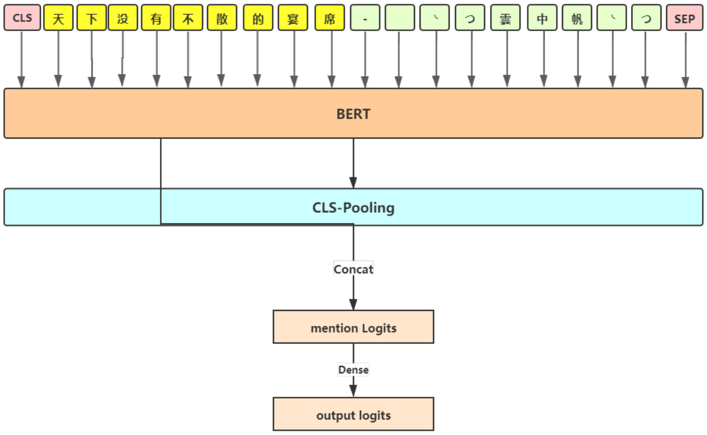

Figure 1: Entity Link model. Inputs: the tokens of two context pair. Outputs: the entity score.

4.2. Entity linking with auxiliary task

4.3. Mention type classification in short text

Jonathan Raiman[10] design a Neural Type model to guide entity disambiguation, a very important point in the article is that when we know the type of candidate entity, this disambiguation task is almost solved. Simultaneously, in the task of DuEL, if a mention can not be linked to the knowledge base, the should be given a nil type. Therefore, we trained a separate classifier for NIL mention cause nil entities and ordinary entities are very different in category distribution. We use the first and the last token in the entity span to classify it, and use cross entropy loss same as equation 2 to train the ordinary nil mention classification model.

4.4. A modified structure for mention type classification

The nil entities classification model mentioned above can effectively handle the task of nil mention classification in the medium length text. However, if in a very short text with only a few words, this model can not capture the context and sentiment information from the extremely short text. To solve this problem and draw on the NL2SQL task, we propose a new and effective model framework.

The above approach of nil type classification task has a problem: only the original text information is used, but the label information of the event type text is ignored. In the DuEL task, the tags have some semantic information. For example, the types like ”Physical geography”, ”Diagnosis and treatment methods” can provide as an auxiliary function for the classification task. If semantic information of the tag can be integrated into the model, it will improve the model’s understanding of the label. Therefore, I surround 24 nil types with an unused label in the vocab, and then splice them to the original text. The schematic diagram of the model is shown in Figure 2.

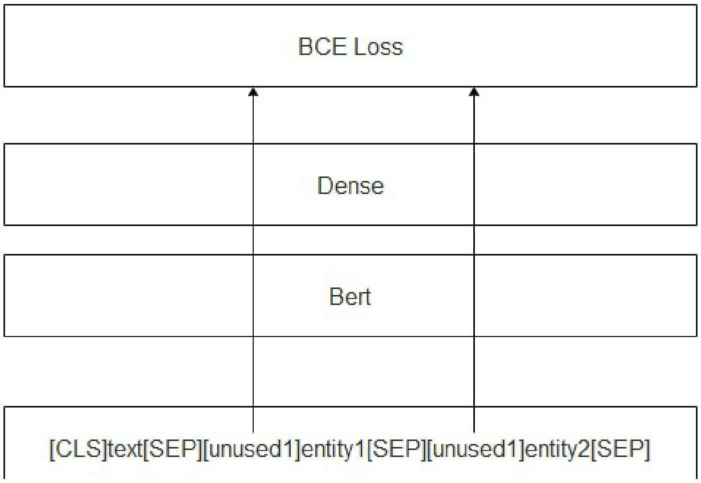

Figure 2: Modified Deep Type model for the classification task

5. EXPERIMENTS

5.1. Datasets

We conduct experiments on DuIE2.0 dataset which is provided by CCKS 2020 Chinese short text entity link task. The knowledge base for this task includes approximately 360,000 entities from the Baidu Baike knowledge base. Each entity in the knowledge base contains a KB-ID, a string name, upper type information, and a series of triples ¡subject, predicate, object¿ information form related to this entity. Each row in the knowledge base represents a record in the knowledge base, and the format of each record is a json format. The training data include 70000 short text, the dev / test1 include 10000 short text and the test2 include 30000 short text. The average length of the short text is about 35 Chinese characters, covering entities in different fields (including examples and concepts of each vertical category), Such as characters, movies, TV, novels, software, organizations, events, etc., as well as general concepts.

5.2. Results

Table 1. shows the accuracy score on DuEL2.0 datasets of our methods.

Table 1: Accuracy on DuEL2.0(test set)

Methods	**score**
baseline	87.5±0.05
Semi-BERT	87.9±0.06
Semi-BERT with auxiliary task	88.2±0.09
Semi-BERT with modified classification	88.8±0.05
all above	**89.266**±0.05

Conclusion

In this paper, we propose to improve the entity linking task by three ways: using a novel mask way to pre-train BERT model; adding an auxiliary task to help the model to utilize entity category information; putting a modified model to work on the nil type classification task. The experiments result show that our three models both significantly outperforms the baseline model and achieved good results at third place.

Acknowledgment

The work is partly funded by National Key Research and Development Program of China, and funded by China Conference on Knowledge Graph and Semantic Computing 2020 Organizing Committee.

References

[1] Mendes, P. N., Jakob, M., García-Silva, A., Bizer, C., "A. DBpedia spotlight: shedding light on the web of documents." In: Proceedings of the 7th international conference on semantic systems, pp. 1-8. ACM (2011).

[2] L. Dong. W Wang, Unified Language Model Pre-training for Natural Language Understanding and Generation, arXiv:1905.03197 [cs.CL]

[3] Cheng, X., and Roth, D. 2013." Relational inference for wikification". In EMNLP, 1787–1796.

[4] Cornolti, M., Ferragina, P., Ciaramita, M., Rüd, S., Schütze, H.: A piggyback system for joint entity mention detection and linking in web queries. In: Proceedings of the 25th International Conference on World Wide Web, pp. 567-578. International World Wide Web Conferences Steering Committee (2016).

[5] Deepak, P., Ranu, S., Banerjee, P., Mehta, S.: Entity linking for web search queries. In: European Conference on Information Retrieval, pp. 394-399. Springer, Cham (2015).

[6] Pan, X., Cassidy, T., Hermjakob, U., Ji, H., Knight, K.: Unsupervised entity linking with abstract meaning representation. In: Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies. pp. 1130–1139 (2015)

[7] Sun, Y., Lin, L., Tang, D., Yang, N., Ji, Z., Wang, X.: Modeling mention, context and entity with neural networks for entity disambiguation. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)

[8] Ganea, O.E., Hofmann, T.: Deep joint entity disambiguation with local neural attention. arXiv preprint arXiv:1704.04920 (2017)

[9] Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Wu, H.: ERNIE: Enhanced Representation through Knowledge Integration. arXiv preprint arXiv:1904.09223 (2019).

[10] Raiman JR, Raiman OM. DeepType: multilingual entity linking by neural type system evolution. InThirty-Second AAAI Conference on Artificial Intelligence 2018 Apr 27.