Semantic Candidate Retrieval for Few-Shot Entity Linking

Jianyong Chen, Jiangming Liu(✉), Jin Wang, and Xuejie Zhang

School of Information Science and Engineering, Yunnan University, Kunming City,
Yunnan Province, China
cjy@mail.ynu.edu.cn, {jiangmingliu,wangjin,xjzhang}@ynu.edu.cn

Abstract. Entity Linking (EL) is the task of automatically linking entity mentions in texts to the corresponding entries in a knowledge base. Current EL systems exhibit the great performances on the standard datasets, but in real-world applications, they are computationally intensive and expensive in large-scale processing, and the entity entries are limited to the knowledge bases. The newly-emerging entities may hinder the generalization ability of the EL systems. To this end, we propose the semantic candidate retrieval method for the few-shot entity linking task. The semantic candidates corresponding to the mentions are selected by inverted indexing, and then, the semantic ranker is proposed to choose the top appropriate candidate to be linked. The proposed model achieves the accuracy of 53.19% in the shared task 6 of NLPCC-2023.

Keywords: entity linking · semantic candidate · inverted index

1 Introduction

Entity Linking (EL) is the task of automatically connecting entities mentioned in the text to their corresponding entries in a Knowledge Base (KB), such as Wikipedia, which is a collection of facts relating to those entities. EL is widely used in natural language processing (NLP) applications, including question answering [1], information extraction [2], and natural language understanding [3]. It is essential for connecting unstructured text with knowledge bases, allowing access to a wealth of carefully selected material.

The entity mentions that appear in the context often pose ambiguity and cannot be directly linked to the KB. Specifically, entity mentions in the text invariably involve the inherent ambiguity of natural language expressions, where the same entity has multiple mentions referring to multiple entities. As shown in Fig. 1, the mention of *England* is attached to the entity *England Football Team* instead of a country, a part of the United Kingdom. In real-world application, the newly-emerging entities do not exist in the knowledge bases in the inference stage, which is called zero-/few-shot entity linking [4].

To this end, existing EL methods apply the semantic retrieval in a Siamese network with information stored in the knowledge base, such as textual entity descriptions or fine-grained entity types. However, If these are billions of textual

F. Liu et al. (Eds.): NLPCC 2023, LNAI 14304, pp. 41–47, 2023.
https://doi.org/10.1007/978-3-031-44699-3_4

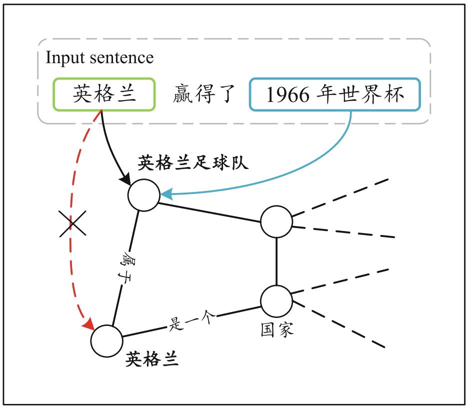

Fig. 1. An illustrated example of zero- and few-shot entity linking.

descriptions, these methods will be computationally intensive, resulting in the expensive computation.

To alleviate this problem, we propose a semantic candidate retrieval based on inverted indexing for zero- and few-shot entity linking. An inverted indexing is built to store the mappings from the mentions of entities to a set of relevant textual descriptions. Instead of all the entity candidates, we select several semantic-related entities candidates to be matched with the mentions by the inverted indexing with the highest scores of term frequency-inverse document frequency (TF-IDF). After that, the final entity linked to the mention is obtained by measuring the cosine similarity between the selected entity candidates and the context via the sentence-transformer [5].

The experiments are conducted on the development set released by the official organizers. The experimental results show that our model has a 23% improvement in Recall@10 over baseline, reaching 95.6%.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 presents the proposed semantic candidate retrieval based on an inverted index. Section 4 shows the experimental results and the ablation study. Conclusions are drawn in Sect. 5.

2 Related Work

EL usually consists of two basic stages, candidate entity generation and entity disambiguation. This section briefly summarizes the related work about the two stages.

2.1 Candidate Generation

Much prior work on candidate generation use a Dictionary-Based approach. This method is applied to almost all entity linking systems. The main idea is to make

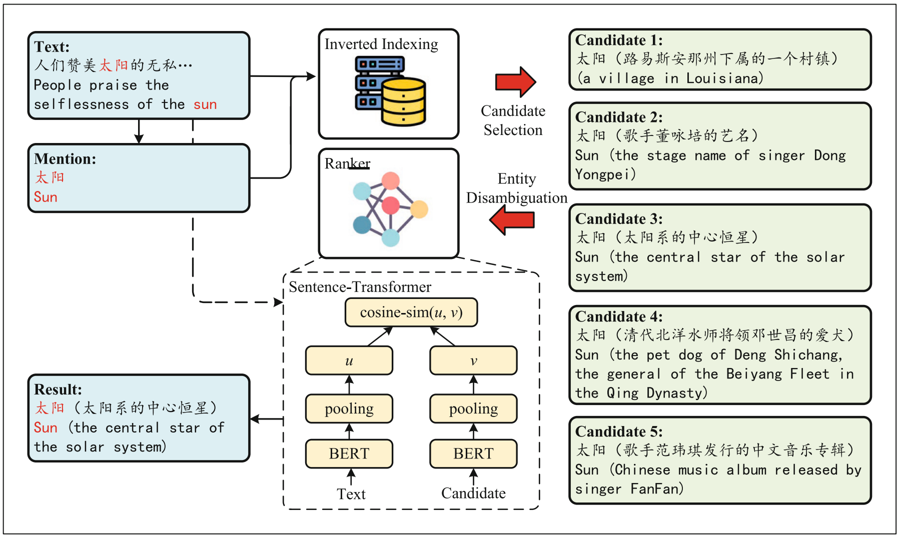
Text:
人们赞美太阳的无私…
People praise the selflessness of the sun
Mention:
太阳
Sun
Inverted Indexing
Candidate Selection
Ranker
Entity Disambiguation
Sentence-Transformer
cosine-sim(u, v)
u
v
pooling
pooling
BERT
BERT
Text
Candidate
Result:
太阳（太阳系的中心恒星）
Sun (the central star of the solar system)
Candidate 1:
太阳（路易斯安那州下属的一个村镇）
(a village in Louisiana)
Candidate 2:
太阳（歌手董咏培的艺名）
Sun (the stage name of singer Dong Yongpei)
Candidate 3:
太阳（太阳系的中心恒星）
Sun (the central star of the solar system)
Candidate 4:
太阳（清代北洋水师将领邓世昌的爱犬）
Sun (the pet dog of Deng Shichang, the general of the Beiyang Fleet in the Qing Dynasty)
Candidate 5:
太阳（歌手范玮琪发行的中文音乐专辑）
Sun (Chinese music album released by singer FanFan)

Fig. 2. Overall architecture of the proposed semantic candidate retrieval for zero- and few-shot entity linking.

full use of all kinds of information provided by Wikipedia, including redirection pages, disambiguation pages, anchor text, etc., to construct a mapping relationship dictionary between entity names and all linked entities, and then use the information in the dictionary to generate candidate entities gather.

2.2 Entity Disambiguation

Previous work focuses on designing effective artificial features and complex similarity measures to obtain better disambiguation performance. He et al. learn distributed representations of entities to measure similarity without human features by keeping words and entities in the joint semantic space and ranking candidate entities directly based on vector similarity [6]. Sun et al. propose the embedded representations of the entities and the contexts via convolutional neural networks [7]. Based on BERT [8], Chen et al. integrate the entity similarity into the local model of the latest model to capture the entity type information [9].

3 Semantic Candidate Retrieval

As shown in Fig. 2, the proposed semantic candidate retrieval consists of the inverted indexing module and the ranker. The candidate generation module is designed to select top-n candidate entities together with their descriptions in KB, and the ranker is designed to obtain final entity according to the cosine similarity between the entity descriptions and the context.

3.1 Candidate Generation

Given the entities in the KB, it is expensive to enumerate all entity descriptions in the KB for each mention. Instead, we aim to search the relevant entities by applying TF-IDF scoring method on the entity descriptions. In particular, given one mention, we select the entity descriptions in KB that contain the mention by inverted indexing. After that, we obtain the top-n relevant entity descriptions according to the TF-IDF score of the mention over the selected entity descriptions.

3.2 Entity Disambiguation

Given the context $u = [u_1, u_2, ..., u_m, ..., u_l]$, where u_i is the ith word and u_m is the mention, A set of candidates of entity descriptions, $V = \{v\}^n$, are generated, as described in Sect. 3.1. Each candidate of entity description $v = [v_1, v_2, ..., v_k]$ consist of sequence of words. The Simaese network with dual encoders is applied to project the entity descriptions v and the context u to obtain the hidden representations in a vector space, and the cosine similarity is computed,

$$\mathrm{sim}(u, v) = \frac{\phi(u)^T \psi(v)}{\phi(u)\psi(v)}, \tag{1}$$

where ϕ and ψ are BERT encoders, and representations of `[CLS]` is used to be the representation of input texts.

In addition, we adopt the TF-IDF score calculated by the inverted indexing as the prior knowledge. The final score is produced with the cosine similarity as

$$score(u, v) = P(v|u) \cdot \mathrm{sim}(u, v) \tag{2}$$

3.3 Post-processing

We define the function that returns the maximum frequency of the entity linked by the mention:

$$f(u_m) = \max_{e \in E} c(e, u_m), \tag{3}$$

where the function $c(e, u_m)$ returns the frequency of the entity e linked by the mention u_m. In the post-procerssing, we assign the mention u_m if $f(u_m) \leq 2$ with the entity $e = \arg max_{e \in E} c(e, u_m)$.[1]

4 Experiments

4.1 Dataset

Table 1 shows the data descriptions in Entity Linking, where each sample contains a *mention*, as well as a text. The *start* and *end* index indicate the starting

[1] The entity is chosen with random sampling if more than one entities satisfy the condition.

Table 1. An example in Entity Linking provided by the share task in NLPCC 2023.

Field	Description	Example
id	The id of this sample	hansel-eval-zs-1463
text	The text which contains the mention to be linked.	…吉尔莫·德尔·托罗的《匹诺曹》，在上个月…
start	Starting position of the mention in the text.	29
end	Ending position of the mention in the text.	32
mention	A word or phrase to be linked.	匹诺曹
gold_id	The id of the corresponding entity in the KB.	Q73895818
source	The source of text.	https://www.1905.com/news/20181107/1325389.shtml
domain	The field that the text is talking about.	news

and ending position of the mention in the text, respectively. We use the data provided from the shared task in NLPCC 2023 for the experiments. This task aims at testing the generalization of Chinese EL systems for infrequent and newly-emerging entities. The dataset is a human-calibrated and multi-domain Chinese EL benchmark with Wikidata as KB, consisting of 9,879,813 mentions with 541,058 entities for training and 9,674 mentions with 6,320 entities as a validation set. The evaluation metrics are recall and accuracy.

4.2 Baselines and Results

In order to investigate the modules in the proposed models, we take the two baseline models:

- FTS: search for entity candidates by matching the mention with the wiki titles. Then sort the candidates according to the number of mentions in the wiki.
- RIS: obtain the candidates using the elastic search only without additional selection strategy.

Table 2 shows the recall of the candidate generation in the second column. Since both RIS-ST and RIS use the elastic search algorithm, they have the same recall scores. The models equipped with the elastic search algorithm achieved the highest recall score of 0.94. Table 2 shows the accuracy of the entity linking. Compared to RIS, the proposed model achieves the best accuracy of 0.65 with considering the cosine similarity via the sentence-transformer.

4.3 Analysis

Indexing Option. The main difference between using *keyword* and *text* as index options in Elastic Search is that *keyword* is for the exact query on structured data, while *text* is for the full-text search on unstructured data. To investigate

Table 2. The recall of the candidate generation across models and the accuracy of the entity linking

Model	Recall(%)	Accuracy (%)
FTS	71.2	/
RIS	94.0	59.2
RIS-ST (ours)	94.0	65.4

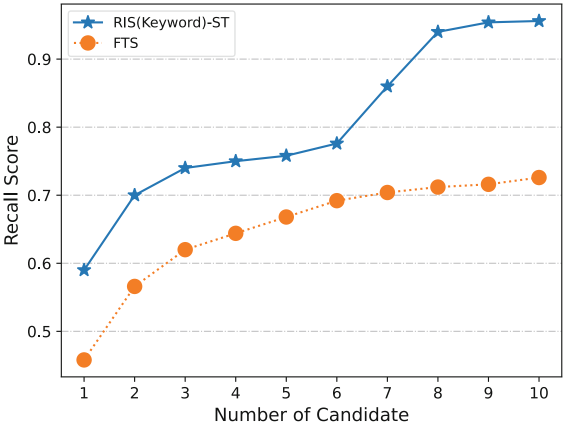

Fig. 3. The recall in choosing top-n candidates.

which indexing option is better for entity linking, we carry the comparative experiment, where one model uses the text option, and the other uses the keyword option. The experiments shows that the model using the keyword options can achieves the recall of 94.0% in the candidate generation, which is better than the model using the text option (86.6% recall).

Post-processing. In the test set, a total of about 900 entities or emerging entities were post-processed (described in Sect. 3.3), resulting that the accuracy improved by 3% points, compared to the models without post-processing.

Amount of Candidates. The top-n candidates are chosen as the results of the candidate generation. If the size of candidates is too small, the correct links will not appear in the list of candidates. In contrast, the more candidates there are, the more noisy links will be added. Therefore, it is important to make a trade-off. Figure 3 depicts the recall score of RIS-ST with different value of n. Based on the elbow approach, we optimize the n to be 8 in our final models.

Hand-off Test Results. Table 3 shows the test results provided by the all ablation models in the shared task of NLPCC 2023. We recommend the elastic search to obtain 8 high-quality candidate sets with indexing options using keywords. The sentence-transformer is used to make full use of the contextual information of wiki text to select reasonable hypotheses from the candidate sets. We also propose that post-processing for tail entities or emerging entities can also improve the overall performance of the model.

Table 3. The test results across different models.

Strategy	Accuracy (%)
RIS(Text)	48.59
RIS(Text)-ST	49.69
RIS(Keyword)-ST	50.18
RIS(Keyword)-ST-Post Processing	53.19

5 Conclusion

In this paper, the method based on inverted index and semantic analysis and sorting is proposed for Chinese entity linking. In addition, we discovered that the post-processing shows the significant improvement in the task of entity linking for the newly-emerging entities. Together with the modules and the strategies, our models achieved 53.19% in the share task 6 of NLPCC-2023 Accuracy, and the final model is ranked 3 among all the submission models.

References

1. Sorokin, D., Gurevych, I.: Mixing context granularities for improved entity linking on question answering data across entity categories, 4 (2018)
2. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: issues, techniques, and solutions. IEEE Trans. Knowl. Data Eng. **27**, 443–460 (2015)
3. Neumann, M., King, D., Beltagy, I., Ammar, W.: Scispacy: fast and robust models for biomedical natural language processing, 2 (2019)
4. Xu, Z., Shan, Z., Li, Y., Hu, B., Qin, B.: Hansel: a Chinese few-shot and zero-shot entity linking benchmark. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, pp. 832–840. Association for Computing Machinery (2023)
5. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks, 8 (2019)
6. He, Z., Liu, S., Li, M., Zhou, M., Zhang, L., Wang, H.: Learning entity representation for entity disambiguation (2013)
7. Sun, M., Guo, Z., Deng, X.: Intelligent BERT-BiLSTM-CRF based legal case entity recognition method. In Proceedings of the ACM Turing Award Celebration Conference - China, ACM TURC '21, pp. 186–191, New York, NY, USA, (2021). Association for Computing Machinery
8. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Google, and A I Language. BERT: pre-training of deep bidirectional transformers for language understanding (2018)
9. Chen, S., Wang, J., Jiang, F., Lin, C.-Y.: Improving entity linking by modeling latent entity type information, 1 (2020)