

BUPTTeam Participation at TAC 2015 Knowledge Base Population

Yongmei Tan, Di Zheng, Maolin Li and Xiaojie Wang

Center for Intelligence Science and Technology and Technology
Beijing University of Posts and Telecommunications
Beijing, China

{ymtan, zhdi, mlli, xjwang}@bupt.edu.cn

Abstract

This paper overviews BUPTTeam’s system participated in the Trilingual Entity Discovery and Linking (TEDL) task at TAC 2015. In this year TEDL is a new Trilingual entity discovery and linking task, and then there are more challenges. In this paper, we propose a novel method to recognize name mentions in raw texts and link them to a knowledge base (KB) entries based on the following four steps: 1) preprocessing, 2) Named Entity recognition, 3) mention expansion, 4) candidates generation, 5) candidates clustering, 6) candidates ranking. The evaluation results show that our method significantly outperforms state-of-the-art TEDL task.

1 Introduction

The goal of EDL track at Text Analysis Conference (TAC) 2015 is to automatically discover entity mentions from three languages (English, Chinese and Spanish) raw texts and link them to a knowledge base, and cluster NIL mentions across languages. More entity types and mention types were also added into some languages (Ji et al., 2015).

Compare to the KBP2014 EDL task, the main differences in KBP2015 TEDL (Tri-lingual Entity discovery and linking) are concluded as the followed.

- EDL is extended from mono-lingual to tri-lingual (English, Chinese and Spanish).
- Previous Entity Linking and EDL tasks mainly focused on three main types: Person (PER),

Organization (ORG) and Geo-political entities (GPE). Two new entity types – natural locations (LOC) and facilities (FAC) for all three languages are added in TEDL this year.

- Titles (TTL) for English are added. A mention of a title that refers to the position itself will be tagged as a title, whereas a title being used as a reference to a specific, real-world person will be tagged as a nominal PER.

- A new KB based on Freebase snapshot is prepared.

In this paper, we propose a novel method for TEDL.

Our contributions are summarized as follows:

- We apply Elasticsearch to index Freebase. In this way, it is very easy and fast to find the related information from Freebase. (Section 3)
- We propose a novel entity linking method based on topic-sensitive random walk with restart to find the mapping entity for a mention or mentions. (Section 3)

The results show that our system gets the highest score and take the first place.

2 Related Work

The first KBP track held in 2009 and then the research in the area of entity linking has greatly developed (McNamee et al., 2009). The problem of entity linking is recast as one of cross-document entity co-reference (Monahan et al., 2011). The task of entity linking has attracted a lot of attention, and many shared tasks have been hosted to promote entity linking research (Ji et al., 2010; Ji and Grishman, 2011; Cano and others, 2014; Carmel et al., 2014; Ji et al., 2015).

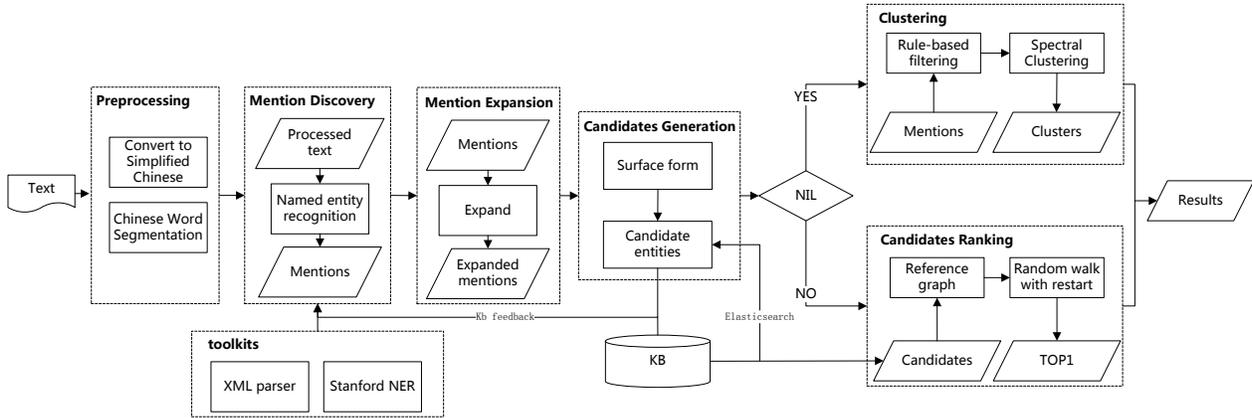


Figure 1: System Architecture

3 System Architecture

The architecture of our TEDL system is described as Figure 1. It includes the following six components.

- 1) Preprocessing
- 2) Named Entity recognition
- 3) Mention expansion
- 4) Candidates generation
- 5) Candidates clustering
- 6) Candidates ranking,

3.1 Preprocessing

There are many traditional Chinese, typos and nicknames in raw texts. We convert traditional Chinese to simplified Chinese.

In Knowledge base, we abbreviate and normalize URIs of Freebase, so that they can be easily represented and handled, and ignore unwanted triples.

We build the Elasticsearch Index. Elasticsearch is a distributed scalable real-time search and analytics engine.

3.2 Named Entity Recognition

We use Stanford NER¹ to recognize most mentions. In addition, mentions representing author whose type is person and linking result is always NIL can be extracted in Discussion Forum documents.

After getting the linking result of a mention, we can obtain extra useful information from the knowledge base which would help us recognize more mentions. For example, after identifying a

person with an entity in freebase, the person’s full name, partial names, aliases, abbreviations, and alternate spellings can be got. In this way some mentions that are not be recognized by Stanford NER could be retrieved. We also prepare an alias dictionary which could help us find more mentions.

3.3 Mention Expansion

Sometimes mentions are nickname, alias, acronyms or part of their full names. We use some heuristic rules to expand their surface form from the associated document where the mention appears.

In addition, the expansion that is adjacent to an acronym is also leveraged such as British Petroleum (BP), TAC (Text Analysis Conference) et al.

3.4 Candidates Generation

This stage attempts to identify potentially correct Freebase entries for mentions. We generate the possible candidates set E_m for each mention m by Elasticsearch where m ’s type is corresponding to candidates’ type in Freebase described as the following table. Elasticsearch assigns each candidate a score according to some constraints.

Table 1: The corresponding mapping relation

Mention’s type	Entity’s type in Freebase
ORG	organization.organization
LOC	location.location
GPE	geography location.country location.administrative division location.statistical region

¹ <http://nlp.stanford.edu/ner/>

PER	people.person
FAC	architecture.structure

3.5 Candidate Ranking

In most cases, the size of E_m is larger than one. Therefore, we rank the candidates and selection the best one using the following method.

1) Building Referent Graph

Referent graph is a strongly connected graph represented by $G=(V, E)$, where V is set of all mentions in a document and all candidates, E is set of all edges. Each edge is either between a mention and an entity or between an entity and an entity.

● Local Mention-to-Entity Compatibility

The compatibility between a mention m and a specific entity e is calculated based on the Bag of Words model:

$$CP(m, e) = \frac{m \cdot e}{|m||e|} \quad (1)$$

Where the mention m is represented as a vector of its context words, and the entity e is represented as a vector of its information text in Freebase. All words are weighted using the TF-IDF schema.

● Semantic similarity between Entities

The semantic similarity between entities can be computed as:

$$NGD(e_1, e_2) = \frac{\log(\max(|S(e_1)|, |S(e_2)|) - \log(|S(e_1) \cap S(e_2)|))}{\log(|W|) - \log(\min(|S(e_1)|, |S(e_2)|))} \quad (2)$$

$$SR(e_1, e_2) = \frac{1}{kNGD(e_1, e_2) + 1} \quad (3)$$

where $e_1, e_2 \in E_m$, $S(e_1)$ and $S(e_2)$ are the sets of entities that are related to e_1 and e_2 , W is the set of all entities in Freebase. Formula (2) is the formal Google distance. Formula (3) is the semantic similarity between entities. k is a parameter.

● Weight of Edge

There are three types of edges in Referent Graph and the weight of each type can be computed respectively as follows:

$$P(m \rightarrow e) = \frac{CP(m, e)}{\sum_{e \in N_m} CP(m, e)} \quad (4)$$

$$P(e_i \rightarrow m) = \frac{CP(m, e_i)}{\sum_{m \in N_{e_i}} CP(m, e_i) + \sum_{e \in N_{e_i}} SR(e_i, e)} \quad (5)$$

$$P(e_i \rightarrow e_j) = \frac{SR(e_i, e_j)}{\sum_{m \in N_{e_i}} CP(m, e_i) + \sum_{e \in N_{e_i}} SR(e_i, e)} \quad (6)$$

where N_m refers to a set of adjacent candidate entities e in the graph. N_{e_i} refers to the set of mentions and entities which are adjacent with the candidate entity e_i . The transition probability matrix T on the graph G can be calculated by formula (4), (5) and (6).

2) Based on topic-sensitive random walk with restart

The random walk original vector α on G is the vector of $|V| \times 1$. The value vector α is composed of two parts, respectively as initial value of mention m and candidate entity e . The value of mention m is as shown in the following formula. α_{m_i} refers to the corresponding value of mention m_i in the vector α .

$$\alpha_{m_i} = \frac{TF-IDF(m_i)}{\sum_{m \in V_m} TF-IDF(m)} \quad (7)$$

Where m_i refers to the i th mention in the document and V_m refer to the set of all mentions in G ; $TF-IDF(m_i)$ refers to TF-IDF weight of m_i .

The initialization of values in the vector α are shown in formula (8) as follows. $e.c$ and $m.c$ are the topics of entity and mention respectively.

$$\begin{cases} \alpha_{e_i} = Indegree(e_i) + Outdegree(e_i), e \in \{e|e.c = m.c, e \in V\} \\ \alpha_{e_i} = 0, e \in \{e|e.c \neq m.c, e \in V\} \end{cases} \quad (8)$$

After completion of initialization of vector α , the sum of all the items of vector α after standardization disposal is 1 thus to make sure that vector α is a correct initialization vector.

Formula (9) and (10) illustrate the process of random walk with restart:

$$\mathbf{r}^0 = \alpha \quad (9)$$

$$\mathbf{r}^{t+1} = (1 - \lambda) \times T \times \mathbf{r}^t + \lambda \times \alpha \quad (10)$$

where \mathbf{r}^t refers to the intermediate result of random walk with restart, t refers to times of iteration, and λ refers to a parameter.

Making $\mathbf{r}^{t+1} = \mathbf{r}^t$, eventual stationary distribution can be calculated as shown in formula (11):

$$\mathbf{r} = \lambda(\mathbf{I} - cT)^{-1}\alpha, c = 1 - \lambda \quad (11)$$

For a mention m , the optimal target entity can be calculated as shown in formula (12).

$$m.e = \arg \max_{\varepsilon} (CP(m, e) \times \mathbf{r}(e)) \quad (12)$$

where $m.e$ refers to the optimal target entity of mention m . $\mathbf{r}(e)$ refers to corresponding value of candidate entity e in the stationary distribution vector.

3.6 Clustering

If the candidates set E_m for the mention m generated by the Candidate Generation is empty, the linking result of mention m is NIL. We cluster the NIL mentions as the following two steps.

Firstly NIL mentions are clustered by the strict rules. These rules can be divided into five types.

Secondly we cluster the NIL mentions based on Spectral Clustering algorithm (make use of the spectrum of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions).

4 Results and Discussion

We submitted five runs for our system. Table 2 lists the performance of NER and NER classification of our best run. We get the first place in Chinese (F1=76.9%) and overall (F1=66.1%).

Table 2: Measures NER and Classification of Entity/Mention Type

	strong_typed_mention_match			rank
	P	R	F1	
English	0.800	0.647	0.715	2
Chinese	0.792	0.748	0.769	1
Spanish	0.637	0.707	0.670	3
All	0.759	0.691	0.724	1

Table 3 describes the linking performance and NIL detection without clustering. We get the highest results in Chinese and the first place in all.

Table 3: Measures NER and either Linking to the Reference KB or Detecting an Entity as NIL (not in the Reference KB)

	strong_all_match			rank
	P	R	F1	
English	0.709	0.574	0.634	2
Chinese	0.759	0.717	0.737	1
Spanish	0.560	0.622	0.590	3
All	0.692	0.632	0.661	1

The performance of NER and NIL clustering are shown in the Table 4. We achieve the highest F1 in English and Chinese, but get the second place in all.

Table 4: Measures NER and Clustering

	mention_ceaf			rank
	P	R	F1	
English	0.765	0.619	0.684	1
Chinese	0.782	0.739	0.760	1
Spanish	0.584	0.648	0.614	3
All	0.646	0.589	0.616	2

Table 5 describes all kinds of evaluation measures on five mention's types. The best result is GPE type. FAC achieves a much lower accuracy than other types.

All kinds of evaluation measures on two different text genres are shown in Table 6. Typically DF is harder to handled, because text got from Discussion Forums is often irregular and shorter, so we couldn't acquire enough context information.

Table 5: Measures NER, NER Classification, Linking and Clustering on the Pre-defined Five Types

	strong_typed_mention_match			strong_all_match			mention_ceaf		
	P	R	F1	P	R	F1	P	R	F1
PER	0.799	0.693	0.742	0.702	0.609	0.652	0.542	0.472	0.505
ORG	0.569	0.642	0.603	0.485	0.551	0.516	0.532	0.605	0.566
LOC	0.708	0.454	0.553	0.670	0.429	0.523	0.694	0.445	0.542
GPE	0.842	0.770	0.804	0.773	0.707	0.739	0.789	0.721	0.754
FAC	0.241	0.048	0.081	0.230	0.046	0.077	0.218	0.044	0.073

Table 6: Measures NER, NER Classification, Linking and Clustering on the Different Text Genres

	strong_typed_mention_match			strong_all_match			mention_ceaf		
	P	R	F1	P	R	F1	P	R	F1
NW	0.710	0.613	0.658	0.619	0.535	0.574	0.668	0.577	0.620
DF	0.796	0.759	0.778	0.748	0.714	0.731	0.645	0.615	0.629

5 Conclusions

We describe our system on TEDL task of TAC 2015. The evaluation results show that our method significantly outperforms state-of-the-art TEDL task.

We built a complete and robust system including named entity recognition, mention expansion, candidate entity generation, candidate entity ranking and NIL clustering that can be applied to different languages and mention's types. We use a graph-based method to do entity linking. Elasticsearch is introduced to index Freebase making it effective to retrieve information and improve the out system's efficiency and scalability.

Many useful features have been used in the candidates ranking and clustering such as Surface Features, Contextual Features and Topic Features.

There are several aspects to be improved. First, the performance on Spanish entity discovery need to be improved. Second, the mention's type of FAC get lower score, because we can't recognize most mention, the type of which is FAC. In the future we will explore more methods to solve this problem.

Acknowledgments

We gratefully acknowledge the support of the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation of China (NSFC) (90920006). Any opinions, findings, and conclusion or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the Ministry of Education of the People's Republic of China.

References

- Wei Zhang, Yan Chuan Sim, Jian Su, Chew Lim Tan. Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling. Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, 2011
- K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in SIGMOD, 2008, pp. 1247–1250.
- J. Guo, G. Xu, X. Cheng, and H. Li, "Named entity recognition in query," in SIGIR, 2009, pp. 267–274.

- Wei Chan, Jianyong Wang, Ping Luo, Min Wang. LINDEN: Linking Named Entities with Knowledge Base via Semantic Knowledge. Proceeding of the 21st international conference on World Wide Web, 2012
- W. Shen, J. Wang, P. Luo, and M. Wang, "A graph-based approach for ontology population with named entities," in CIKM, 2012, pp. 345–354.
- J. Hoffart, M. A. Yosef, I. Bordino, H. Furstenuau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, "Robust disambiguation of named entities in text," in EMNLP, 2011, pp. 782–792.
- W. Zhang, Y. C. Sim, J. Su, and C. L. Tan, "Entity linking with effective acronym expansion, instance selection and topic modeling," in IJCAI, 2011, pp. 1909–1914.
- Han X, Sun L, Zhao J. Collective entity linking in web text: a graph-based method.[J]. Proceedings of International Conference on Research & Development in Information Retrieval, 2011:765-774.
- Alhelbawy, A., Gaizauskas, r.. Graph Ranking for Collective Named Entity Disambiguation, The 52nd Annual Meeting of the Association for Computational Linguistics (ACL2014), Maryland, USA, 2014.
- Zheng Z, Li F, Huang M, et al. Learning to link entities with knowledge base[J]. Proceedings of the Annual Conference of the North American Chapter of the ACL, 2010:483-491.ngs of Text Analysis Conference, 2011.
- Silviu Cucerzan. TAC Entity Linking by Performing Full-document Entity Extraction and Disambiguation. In Proceedings of Text Analysis Conference, 2011.
- Wei Zhang, Jian Su, Bin Chen. I2R-NUS-MSRA at TAC 2011: Entity Linking. In Proceedings of Text Analysis Conference, 2011.