FRDC's Cross-lingual Entity Linking System at TAC 2013

Qingliang Miao, Ruiyu Fang, Yao Meng, Shu Zhang
Fujitsu R&D Center CO., LTD.
No. 56 Dong Si Huan Zhong Rd, Chaoyang District, Beijing, 100025 P. R. China
{qingliang.miao| fangruiyu| mengyao| zhangshu}@cn.fujitsu.com

Abstract

In this paper, we present FRDC's system at participating in the cross-lingual entity linking (CLEL) tasks for the NIST Text Analysis Conference (TAC) Knowledge Base Population (KBP2013) track. We propose a joint approach for mention expansion, disambiguation, and clustering. In particular, we adopt a lexicon and rule based method for entity classification, a collaborative acronym expansion method and a heuristic combination ranking method that merged ListNet, SVM ranking with web search engine ranking. The results achieved in the TAC cross-lingual entity linking tasks show that our approach is competitive. Our best run achieves 0.655 in B^3+ F1 measure.

1 Introduction

The goal of Knowledge Base Population (KBP) track at Text Analysis Conference (TAC) 2013 is to automatically discover information about named entities and to incorporate this information in a Knowledge Base (KB). The cross-lingual entity linking (CLEL) task we are addressing is part of the TAC KBP2013 evaluations. In the CLEL task, given a Chinese or English query ID, the name of the query, the source document containing the query, and the position of the query in the source document. The type of query can be a person (PER), organization (ORG) or geo-political entity (GPE). The system is required to identify the ID of an English Knowledge Base (KB) entry to which the name refers; or NIL if there is no such KB entry. In addition, a CLEL system is required to cluster together all the NIL queries and provide a unique ID for each cluster.

Entity linking task, however, can be no-trivial due to the mention ambiguity and variation issues. The mention ambiguity issue means that a mention could refer to multiple entities in different context. In contrast to mention ambiguity, mention variation indicates that an entity may be mentioned in different ways such as official name, nickname, aliases, abbreviation or even misspellings (Xianpei Han and Le Sun., 2011; Yu Zhao et al., 2011). CLEL is more complicated due to the cross-lingual ambiguity.

According to (Heng Ji et al., 2011) there are two kinds of methods for cross-lingual entity linking. One is based on machine translation and monolingual entity linking techniques. This kind of systems first translate a Chinese query and its associated document into English, and then run English mono-lingual entity linking to link the translated query and document to English KB. The other one is based on Chinese mono-lingual entity linking and cross-lingual knowledge base linkages. Systems belonging to this schema first apply Chinese entity linking system to link a Chinese query with Chinese KB, and then use cross-lingual KB linkages to map the Chinese KB nodes with English ones.

Our system consists of three main modules: entity mention expansion, entity resolution and NILs clustering. The entity mention expansion is crucial in CLEL systems. Effective entity mention expansion can find most true candidates in the KB and keep the candidate size controllable. In FRDC's system, we discover the entity candidates through three resources: 1) contextual information in the source document; 2) knowledge repository build using Wikipedia dumps and other Chinese encyclopedia; 3) web search engine. For the entity resolution, we develop rich and extensible set of features based on string and semantic similarity and combine multi-rankers to decide the final answers. Finally, we implement a clustering approach for the NILs clustering. The clustering approach utilizes different contextual information for different entity type, and we tune parameters for each entity type separately.

The rest of the paper is structured as follows. In the following section we review the existing literature. We introduce the proposed approach in section 3. We conduct comparative experiments and present the experiment results in section 4. At last, we conclude the paper with a summary of our work and give our future working directions.

2 Related Work

Most previous systems conduct entity disambiguation and NILs clustering in a cascaded way (Angela Fahrni et al., 2012). According to (James Mayfield et al., 2012), two main strategies are adopted in CLEL. The first one is translating source document to the target KB language and then reuse existing English EL systems (Paul McNamee et al., 2011; Taylor Cassidy et al., 2011). The other one is mapping articles in the source collection language Wikipedia to entries in the English Wikipedia by leveraging inter-language links (Sean Monahan et al., 2011). In overview report (Heng Ji et al., 2011) summarizes several useful approaches for entity linking such as query classification, acronyms expansion, entity context modeling and join inference. Since significant differences may exist between entities, several systems have utilized query-dependent ranking models, specific to the type of query. Experimental results indicate that query-dependent ranking outperform baseline methods of using a unified ranking method (Ivo Anastácio et al., 2011; Taylor Cassidy et al., 2011). Rule or pattern based acronym expansion can effectively reduce the ambiguity of the acronym mentions but fail in expanding more complicated acronyms, Zhang Wei et al. 2011 proposed supervised learning algorithm to expand complicated acronyms, which leads to 15.1% accuracy improvement over state-of-the-art acronym expansion methods. Entity context modeling utilizes "collaborators" and "supporters" to joint translation and disambiguating entities (Heng Ji et al., 2011). Recently, Angela Fahrni et al. proposed a joint system for entity disambiguation, recognition of NILs and clustering using Markov Logic (Angela Fahrni et al., 2012).

3 FRDC's System

In this section, we first introduce the overview of our CLEL system, and then present the detailed approach in each step.

3.1 System Overview

Figure 1 illustrates the overview of our CLEL system. We first classify the queries into PER, ORG and GPE, and then we expand the queries to generate candidates in the KB. After that, Chinese documents and candidates are translated into English ones, and mono-lingual English entity linking approach is adopted to re-rank the candidates. Finally, NILs are clustered.

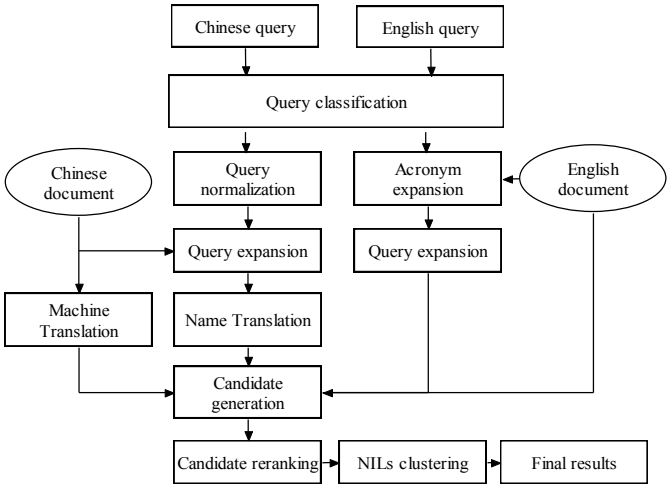

Figure 1. Overview of Cross-lingual Entity Linking System

3.2 Query Classification

We build a hybrid approach to classify entities into different types (PER, ORG, GPE). Specifically, the classification model leverages lexicons derived from our address database, person name database, web encyclopedia and some heuristic rules. FRDC's address database contains Chinese geographical addresses from national-level to township-level and the world's major cities. FRDC's person name database contains China's common names, surnames and the world's common names. We also use the lexicons to extract frequent patterns to classify entity type. For example, GPE entities usually end with "state", "province", "city", "region", "county". ORG entities usually end with "corporation", "government", "university", "commission". In our query classification module, English abbreviations

are classified as ORG. We first use mention as query to search against lexicons, if the mention is in lexicon, we assign the type according to corresponding lexicon. Otherwise, we use patterns and rules to classify the mentions, if the mentions match patterns or comply with rules, we assign the corresponding type. For those mentions whose type cannot be determined by lexicons and rules, we use Stanford NER to estimate the query type.

3.3 Query Expansion

Before query expansion, we considered two query reformulation mechanisms. The first one reformulates acronyms for the named entity references according to textual patterns e.g. finding expressions like "China National Petroleum Corporation (CNPC)". In some cases, there is a long distance between acronyms and their full names, therefore we adopt a collaborative expansion strategy to expand them. For instance, in document d_1, we have text "China National Petroleum Corporation (CNPC)", so we can find full name using patterns directly. In document d_2, acronym "CNPC" is in the title of d_2, while the full name "China National Petroleum Corporation" is located in last paragraph of d_2. In this case, we query full name derived from d_1 against d_2, if the full name is in d2, we expand the acronyms use the same full name. Moreover, we notice single token mentions are more ambiguous, therefore we adopt the method in (Suzanne Tamang et al., 2012) to reformulate mentions only have one token. After that, we use a traditional query expansion method to expand mentions. Note that, we expand Chinese mentions at Chinese side using a Chinese repository. After expansion, we translate the candidates into English using a cascade translation method. We first index the cross-lingual lexicon (Valentin I. Spitkovsky and Angel X. Chang., 2012). And then, we use candidate name as query to retrieve corresponding English translations. If the lexicon does not contain the candidate, we use machine translation systems as discussed in next section.

3.4 Candidate Generation

The aim of candidate generation is to get KB entries that mention may refer to. To handle the Chinese mentions and source documents, we use machine translation techniques. In the run, which does not access the web, we use an off-line Chinese to English statistical machine translation system developed by FRDC (Zhongguang Zheng et al., 2011) to translate Chinese mention and corresponding source document into English. In other runs, we translate the mentions by using online translation systems and combine the results by a voting mechanism. After translation, we index the translated Chinese documents and KB documents. In candidate generation module, we first obtain all the KB entries that might refer to mention, and then we filter the KB entries based on string similarity. In particular, we first token the English names in the candidate set and search the index of the KB's "title" field to get candidates. After that, we rank the retrieved candidates according to similarity score, and candidate's similarity score lower than a threshold is filtered out from the candidate set.

3.5 Disambiguation

In this section, we mainly introduce the training features and ranking algorithms we used in our system. In addition, we introduce the heuristic combination ranking method merged ListNet, SVM ranking with web search engine ranking.

Learning Features

Previous work has introduced several learning features (Zhicheng Zheng et al., 2010; Ivo Anastácio et al. 2011). In our system, the learning features we selected are described as follows.

(1): Surface features focuses on the mention's string similarity with reference entry independent of context.

Exact name Match: set to be one if the mention's name exactly match the candidate's name, zero otherwise.

Start-with or End-with match: If the mention name or candidate name is the prefix or suffix of each other, the feature is set to one, zero otherwise.

Tokens in common: The number of overlap tokens after tokenization of the mention name and candidate name.

Sub-string: If the mention name or candidate name is the sub-string of the other, the feature is set to be one, zero otherwise.

Levenshtein similarity: The Levenshtein distance of the mention name and candidate name.

Query expanding method code: The code indicates which method the expanding algorithm has used to

generate the candidate. This code reflects the confident of candidates.

Type match: If the type of the mention is consistent with the candidate, the feature is set to be one, zero otherwise. The mention type is identified by query type classification module discussed in section 3.2.

(2) Contextual Features: These features model the contextual information of the mention and corresponding candidates.

Document text similarity: TF-IDF value between candidate's document and mention's source document are scored and generate corresponding weight vector, then cosine similarity between the two weight vectors is scored as feature value.

Context containing: If the candidate name co-occurrences with the mention in source text, the value is set to be one, zero otherwise.

Entities context similarity: Boolean value vector of entities that co-occurrence with the mention and candidate in the whole document, is calculated respectively for each entity type (PER, ORG, GPE). Entities are derived using Stanford Named Entity Recognizer (Jenny Rose et al., 2005).

Ranking model

Learning to rank algorithms are mainly created for ranking the retrieved documents in IR system. These ranking models can be adopted in entity linking scenario as well and they turned out to be effective in achieving good performance (Heng Ji et al., 2011). Our system adopted two ranking models to rank candidates: Pointwise SVM and ListNet (Zhe Cao et al., 2007). Each candidate is assigned a ranking score by the ranking model, and the top ranked candidate is considered to be the correct answer of the mention, or discard it and generate NIL by the system. We use an empirical threshold to determine whether to discard the top ranked candidate or not. For SVM ranking model, the threshold score is 0.9 and for ListNet model the threshold is 0.7.

Heuristic combination ranking

We have noticed that many candidates' ranking score by SVM and ListNet ranker are close to the threshold, which indicates these ranking models are uncertain about these candidates, therefore these candidates are more ambiguous to the ranking models. We also notice that the results of web search engine are usually correct for GPE and PER, ORG with longer full names. In order to get more confident results, we adopt heuristic combination ranking method merged web search engine ranking (System A) with ListNet, SVM ranking (System B). The voting mechanism is as follow：

i. If the linked result of System A is same as the result of System B，then we adopt the results with high confidence.

ii. If the output of the two systems are different. For all candidates generated by system B, if the candidates' similarity code has value 1 or 2, which indicates the candidate name matches query in the expansion string exactly, we use the result of system B.

iii. If the similarity code of candidates in system B is larger than 2, or system B produces no candidates, we use the results of system A.

iv. Otherwise, we trust the results in system B.

3.6 NIL Cluster

As Heng Ji et al. 2011 point out that due to the different characteristic of entity types, it's beneficial to use different methods for each entity type separately. Following this idea, we adopted different cluster strategy for different entity types.

The main idea of cluster method is that entities have same profile tend to belong to same entity cluster. Profile can be expressed as either context features or attribute from slot filling. Profile is usually viewed as the most important criterion for entity clustering, such as which country the organization is located in or which company the person is employed in. Inspired by this idea, we first extract context around the mention in the document, then we extract neighbor entities in the context. Due to the different characteristic of each entity type，the features we used to calculate each type's clustering criteria is different. For ORG entities, we mainly consider the geographical affiliation around it, therefore the neighbor entity with GPE type is assigned more weight, and for PER and GPE type we view all neighbor entities as equal.

The NILs clustering module is based on hierarchical clustering algorithm, the clustering procedure fall into two steps: division and agglomeration. First we use the top down approach by view the NIL observations that belongs to same entity type and have the identical string surface as

one cluster, and then we use splitting rules to split clusters recursively. For the NIL queries of ORG type, the split rule is that if the geographical affiliation of two query entities is different, they are divided into two clusters. Secondly we adopt a bottom up approach which treat each observation starts in its own cluster and merge pairs of cluster with context similarity rules (merging rules). The merging rules are different according to entity type. PER and ORG mentions are more ambiguous than GPE ones according to prior knowledge, so the rules for PER and ORG contain not only string similarity but also context window similarity and co-occurrence entity similarity. The splitting and merging features are listed in table 1.

Entity Type	Features for splitting	Features for merging
ORG	string match score geographic location	contextual similarity entity similarity string match score
GPE	string match score	string match score
PER	string match score	contextual similarity entity similarity string match score

Table 1: Clustering features

4 Experiment

In this section, we first introduce the data we used for evaluation, and then we present the experiment results.

4.1 Data

The English reference Knowledge Base consists of 818,741nodes derived from an October 2008 dump of English Wikipedia. We use KBP2011 cross-lingual entity linking training data sets to develop our systems, and then conduct blind test on KBP2013 cross-lingual entity linking evaluation data sets. The detailed data statistics are summarized in Table2.

Corpus		#Queries		
		PER	ORG	GPE
Cross-lingual	Training	817	660	685
	Testing	706	735	714

Table 2: Data sets

4.2 Cross-lingual Entity Linking Results

Entity Type	System	B^3+ F1
PER	Baseline	0.491
	SVM	0.489
	ListNet	0.502
	SVM+WSE	0.519
	ListNet +WSE	0.528
GPE	Baseline	0.750
	SVM	0.773
	ListNet	0.775
	SVM+WSE	0.788
	ListNet +WSE	0.790
ORG	Baseline	0.622
	SVM	0.626
	ListNet	0.625
	SVM+WSE	0.645
	ListNet +WSE	0.645
ALL	Baseline	0.622
	SVM	0.630
	ListNet	0.635
	SVM+WSE	0.652
	ListNet +WSE	0.655

Table 3: Cross-lingual Entity Linking Evaluation Results

Table 3 summarizes the results of our cross-lingual entity linking system.

5 Conclusions and Future Work

FRDC participated with a joined approach in the Chinese cross-lingual entity linking subtasks. The system first classifies the entity according to their types, e.g. ORG, GPE and PER, and then queries are expanded by local context, Wikipedia dumps and web data. We adopt a heuristic combination ranking method in the candidate resolution module. A type-dependent clustering module is developed for NILs clustering. Our system performs well on TAC 2013 data. The experiment results show the heuristic combination ranking model performs better than SVM and ListNet ranking model.

References

Angela Fahrni, Thierry Göckel, Michael Strube. 2012. HITS' Monolingual and Cross-lingual Entity Linking System at TAC 2012: A Joint Approach. In Proceedings of the Fifth Text Analysis Conference.

Heng Ji, Ralph Grishman and Hoa Trang Dang. 2011. Overview of the TAC2011 Knowledge Base

Population Track. In Proceedings of the Fourth Text Analysis Conference.

Ivo Anastácio, Bruno Martins, Pável Calado. 2011. Supervised Learning for Linking Named Entities to Knowledge Base Entries. In Proceedings of the Fourth Text Analysis Conference.

James Mayfield, Javier Artiles and Hoa Trang Dang. 2012. Overview of the TAC2012 Knowledge Base Population Track. In Proceedings of the Fifth Text Analysis Conference.

Jenny Rose Finkel, Trond Grenager, Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In Proceeding of the 43rd Annual Meeting on Association for Computational Linguistics.

Paul McNamee, James Mayfield, Veselin Stoyanov, Douglas W. Oard, Tan Xu, Wu Ke, David Doermann. 2011. Cross-Language Entity Linking in Maryland during a Hurricane. In Proceedings of the Fourth Text Analysis Conference.

Sean Monahan, John Lehmann, Timothy Nyberg, Jesse Plymale, Arnold Jung. 2011. Cross-Lingual Cross-Document Coreference with Entity Linking. In Proceedings of the Fourth Text Analysis Conference.

Suzanne Tamang, Zheng Chen and Heng Ji. 2012. CUNY-BLENDER TAC-KBP2012 Entity Linking System and Slot Filling Validation System. In Proceedings of the Fifth Text Analysis Conference.

Taylor Cassidy, Zheng Chen, Javier Artiles, Heng Ji, Hongbo Deng, Lev-Arie Ratinov, Jiawei Han, Dan Roth, Jing Zheng. 2011. CUNY-UIUC-SRI TAC-KBP2011 Entity Linking System Description. In Proceedings of the Fourth Text Analysis Conference.

Valentin I. Spitkovsky and Angel X. Chang. 2012. A cross-lingual dictionary for English Wikipedia concepts. In LREC, 2012.

Wei. Zhang, Yan Chuan Sim, Jian Su and Chew Lim Tan. 2011. Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling. International Joint Conferences on Artificial Intelligence 2011.

Xianpei Han and Le Sun. 2011. A Generative Entity-Mention Model for Linking Entities with Knowledge Base. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.

Yu Zhao, Weipeng He, Zhiyuan Liu, Maosong Sun. 2011. THUNLP at TAC KBP 2011 in Entity Linking. In Proceedings of the Fourth Text Analysis Conference.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to Rank: From Pairwise Approach to Listwise Approach. In Proceedings of the 24th international conference on Machine learning.

Zhicheng Zheng, Fangtao Li, Minlie Huang, Xiaoyan Zhu. 2010. Learning to Link Entities with Knowledge Base. The 2010 Annual Conference of the North American Chapter of the ACL.

Zhongguang Zheng, Naisheng Ge, Yao Meng, Hao Yu. 2011. HPB SMT of FRDC Assisted by Paraphrasing for the NTCIR-9 PatentMT. In Proceedings of NTCIR-9 Workshop Meeting.