Saarland University Spoken Language Systems Group at TAC KBP 2011

Fang Xu Stefan Kazalski Grzegorz Chrupała Benjamin Roth
Xujian Zhao Michael Wiegand Dietrich Klakow
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
lsv_trec_qa@lsv.uni-saarland.de

Abstract

In this paper we describe our participation in the Knowledge Base Population (KBP) track at TAC 2011. The architecture of our slot filling system is the same as last year. We mainly focus on developing a new system for the cross-language entity linking task. We compare the performance of monolingual retrieval and cross-lingual retrieval for entity linking. For NIL entity clustering, we group retrieved documents into reference clusters and assign NIL entities to a cluster by calculating the similarity between its document and each cluster.

1 Introduction

We report on our participation in two tracks of TAC 2011: slot filling and cross-lingual entity linking. The basic set-up we used for slot filling is the same as last year, and we refer the reader to last year's report for the details (Chrupała et al., 2010). The results of the two runs we submitted to this year's slot filling evaluation are described in Section 6.

This year we focused on developing a system for the cross-lingual entity linking task. Given an entity and a background document mentioning it, the entity linking task is to find whether the entity exists within the knowledge base (KB), or set as NIL if it cannot be found. TAC 2011 KBP proposed a new cross-language entity linking (CLEL) task. The KB is a subset of English Wikipedia while the background documents are in either Chinese or English. The cross-language scenario raises more challenges than the previous monolingual task. The main problems of CLEL include:

- mining name variation of entity mentions in documents;
- disambiguating different meaning of entities;
- connecting knowledge between different languages.

We introduce several components in our system that address these problems.

The core components of our system include document retrieval and entity clustering. The retrieval module returns the most likely KB entity as a linked target, while clustering is utilized to group NIL entities with the same reference. We performed cross-language information retrieval (CLIR), which requires a translation component to translate Chinese queries into English queries to retrieve English KB. To achieve this goal we derive translation candidates from manually constructed dictionaries, phrase tables generated by machine translation methods, and results of a commercial translation system. We also combine the retrieval results generated by using multiple translated queries. However, the CLIR module performs worse than the retrieval of Chinese queries on Chinese KB, which is a collection of articles extracted from Chinese Wikipedia. Therefore our final system adopts two parallel monolingual pipelines for Chinese and English entities, respectively. The framework of our system is presented in Figure 1.

The English and Chinese systems share the same workflow:

1. **Document and query processing** with natural language processing tools;

2. **Document retrieval**, to retrieve relevant articles with query expansion from the whole Wikipedia page archive;

3. **NIL entity classification**, to determine NIL entity by judging whether the title of the top ranked Wikipedia article is mapped to a KB id [1];

4. **NIL entity clustering**, to cluster NIL entities by leveraging large amounts of relevant documents.

The rest of the article is organized as follows. We detail the main components in Sections 2, 3 and 4. Section 5 presents the experimental comparisons of cross-lingual and monolingual retrieval results for Chinese entities. In Section 6 we present our evaluation results. Finally, Section 7 concludes and describes possible areas for future work.

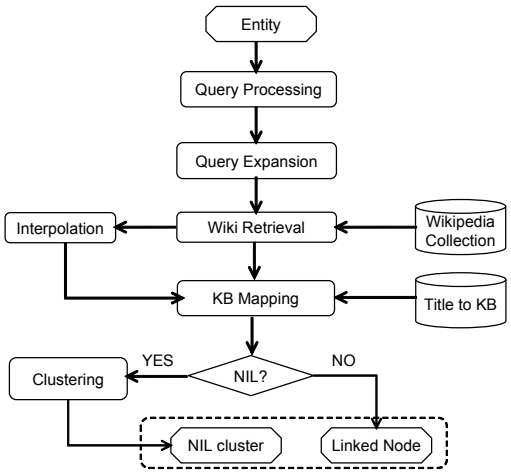

Figure 1: Entity Linking System.

2 Prepossessing

2.1 Background Knowledge Extraction

As shown in Figure 1, the retrieving document collection to be used for retrieval consists of all articles extracted from Wikipedia, so afterwards the mapping from the entities of he KB to Wikipedia titles is used to determine NIL entities. The English Wikipedia is the version released in Oct. 2008, which was used for building the TAC KB. We extract all articles from the latest version of the Chinese dump released in May 2011 and used them as the supportive KB for Chinese EL. Wikiextractor[2] is applied to generate plain articles from both Wikipedia dumps. Chinese articles are written in both traditional and simplified Chinese. As the CLEL task only provides source documents and queries in simplified Chinese, we use the mediawiki-based tool[3] to convert traditional Chinese words into simplified Chinese, accordingly.

Since the KB is partly selected from the whole Wikipedia article collection, we manage to map almost all KB entries to English Wikipedia articles by matching their titles.[4] Combining the mapping of English titles to KBs and the interlanguage links between Chinese and English pages, we generate mappings between Chinese pages and KB ids. Many KB nodes cannot be mapped to any Chinese Wikipedia title. The Chinese Wikipedia is not as sufficiently well-developed as its English counterpart. Chinese Wikipedia contains about one tenth articles of English Wikipedia. Moreover many disambiguation pages do not have Chinese counterparts.

2.2 Document and Query Processing

Initially both source documents and Wikipedia articles are preprocessed for further usage. The English processing is the same as our previous slot filling system (Chrupała et al., 2010). The Chinese preprocessing includes the following steps:

1. Replace html escape characters and remove noisy html garbage (especially for web documents).

2. Sentence segmentation for each passage.

3. Tokenization and POS tagging of Chinese sentences.

4. Named Entity (NE) recognition.

[1] We construct two dictionaries mapping Chinese and English Wikipedia titles to KB ids respectively.

[2] medialab.di.unipi.it/wiki/Wikipedia_Extractor

[3] github.com/tszming/mediawiki-zhconverter

[4] We only lose 1.4% of the KB entries, most of them being non-English titles.

For Chinese processing in steps 3 and 4, our system uses the NLPR Chinese processing tool (Wu et al., 2005).

Unlike English, Chinese sentences are written without spaces to delimit words, therefore we need to break each sentence into successive separate tokens. To relieve the bad influence of segmentation errors and the out-of-vocabulary issue, we tokenize the Chinese source documents with different segmentation criteria: POS, NE, unigram (individual Chinese character) and bigram (two consecutive characters). We also perform the same segmentation of Chinese queries and retrieve from a corresponding document index.

2.3 Acronym Expansion

In order to improve the quality of the retrieval module, we adopt a simple method to expand acronym queries. We consider the English query with all capital letters as an acronym, while the NLPR tool provides the recognition of Chinese acronym words. We use two ways to select the expansion candidates from background documents. If the acronym appears within parentheses, the previous N contiguous tokens are chosen as candidates (N is the number of English letters or Chinese characters in the acronym), otherwise we consider all recognized NEs. The candidates whose initials are identical to the letters from the acronym are chosen as the expansion. If several candidates still exist afterwards, we select the one with the largest term frequency in the document. For example, in the text "*..referring to the National Food Authority (NFA),..*" it is obvious to extract "*National Food Authority*" as the expansion of the query "*NFA*".

3 Document Retrieval

The purpose of this module is to retrieve for every query the relevant entries from KB. We adapted the retrieval component of our slot filling system (Chrupała et al., 2010) to the entity linking task. The Chinese tokens for the expansion are extracted from the POS-segmented documents. We retrieve documents from unigram-, bigram-, POS- and NE-segmented corpora. The monolingual Chinese document ranking is created by interpolating the different retrieval results with the optimal parameters validated on the Chinese development Data.

The most relevant article is considered as a reference to each query. Then if a mapping from the title of the article is found in the dictionary, the mapped KB id is set as the linked id, otherwise the query is regarded as NIL entry. In the following section, we will describe the method to cluster NIL queries into different reference clusters.

4 Entity Clustering

NIL entity clustering automatically groups entities with no KB reference into clusters so that those within a cluster refer to the same target (sense). It extends the people name disambiguation task in SemEval-2007 in (Artiles et al., 2007) by including more types of named entities, e.g. locations and organizations.

Based on our observations of the development data, we assume that all the entities in the same coarse group share identical strings, so we ignore rare cases of different references to entities (e.g. queries "Ford Motor Co." and "Ford" refer to the same company) and cross-lingual reference (e.g. the queries "Hyderabad" and "海得拉巴" in Chinese). Hence the first step of our clustering approach is to group entities with identical names into one coarse group. Entities within a coarse group represent different senses of the entity, such as *Washington* means a person, a city or an organization under different contexts. The next step is to scatter them into different sense clusters, which are considered as final NIL clusters. We utilize the bag-of-words(BOW) from surrounding passages of entity mention from background document to represent the intended sense of the entity.

An important issue in clustering is to determine the number of sense clusters. Since the number varies from entity to entity it is difficult to train a adaptable clustering model for all entities. Moreover some background documents offer sparse and insufficient information to support its mentioned entity. Rather than clustering only background documents, we cluster relevant documents retrieved from source collections using the Indri IR tool.[5] The top 1000 retrieved documents are a much larger documents collection, on which we build sense clusters

[5] www.lemurproject.org/indri

using Hierarchical Agglomerative Clustering (HAC) algorithm with a single linkage (Zhao and Karypis, 2002). Documents in each cluster refer to one sense of the entity.

HAC[6] initially assigns each document to its own cluster and iteratively merges the most similar pair of clusters to form a hierarchical tree, which provides a view of the semantic sense of the entity at different levels of abstraction. We transfer the hierarchy into disjoint clusters by cutting it using the combination similarity of merged clusters (Manning et al., 2008). After removing stop words, each document is represented as a BOW vector $\overrightarrow{x}$ of TFIDF values. Based on the tuning and validation on developing data, we examine two different parameters of hierarchy cutting which will be described in Section 6.2.

Each cluster is considered as an accumulation of relevant terms with respect to single entity sense, therefore the background document sharing more terms with the cluster is most likely to share the same sense. We define the centroid $\overrightarrow{\mu}$ of a sense cluster ω:

$$\overrightarrow{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\overrightarrow{x} \in \omega} \overrightarrow{x}. \quad (1)$$

We assign the query to the nearest sense cluster by measuring Euclidean distance between its document and each cluster centroid in vector space.

5 Comparison of Cross Lingual and Monolingual EL

For the Chinese EL system, the performance is highly influenced by the mapping from Chinese Wikipedia articles to the English KB. This is mainly because of less content of Chinese Wikipedia and insufficient cross-lingual page linking. Plenty of disambiguation information is lost during mapping, i.e. English Wikipedia provides a disambiguation page mentioning 15 article named *Denver* while there exists only one Chinese page about *Denver in Colorado*.[7] For those reasons CLEL is a much better option, which discovers the representations of a meaning in multiple languages by query translations from various MT systems and resources.

	All Entities	**PER**	**ORG**	**GPE**
Overall	0.65	0.67	0.65	0.62
in-KB	0.46	0.25	0.45	0.58
NIL	0.86	0.88	0.8	1

Table 1: Performance of cross lingual EL strategy for Chinese queries on TAC 2011 development data.

	All Entities	**PER**	**ORG**	**GPE**
Overall	0.7	0.71	0.86	0.52
in-KB	0.51	0.45	0.67	0.47
NIL	0.91	0.84	1	1

Table 2: Performance of monolingual EL strategy for Chinese queries on TAC 2011 development data.

We translate each query and its expanded terms into English. To minimize the influence of machine translation(MT) ambiguity and errors, we utilized multiple translation strategies, including a translation dictionary created from bilingual hyperlinks in Chinese Wikipedia Pages, a phrase table extracted from LDC parallel Chinese to English NE lists, N-best translation of Chinese queries and NEs from a statistical MT system (Wu et al., 2011), and online translations from Google[8] only for queries.

For each Chinese query Q_c we create a collection T of English queries using all those translations. For an English token t_e in the translated query T_i, $P(t_e|T_i)$ is a relevance language model (Lavrenko and Croft, 2001) which is estimated over all tokens from T.

$$P(t_e|T) = \sum_{T_i \in T} P(t_e|T_i) P(T_i|Q_c) \quad (2)$$

where $P(t_e|T_i)$ is calculated by maximum likelihood estimation with Dirichlet smoothing. The translation probability from Chinese to English $P(T_i|Q_c)$ is defined as the phrase-to-phrase translation probability if T_i is generated from the SMT system, otherwise it is set to 1 if T_i is the result of dictionary matching or Google translation.[9]

[6] We use the implementation in scipy `http://www.scipy.org/`

[7] See `zh.wikipedia.org/wiki/Denver` and `en.wikipedia.org/wiki/Denver_(disambiguation)`

[8] `translate.google.com`

[9] They are assumed to produce perfect translations of the query.

For weights in the inference network of retrieval model (Turtle and Croft, 1990), we use $P(t_e|T)$ instead of the dice-coefficient we used last year.

As in TAC 2010, we use micro-averaged accuracy[10] (Ji et al., 2010) to compare the effects of different retrieval strategies on the end-to-end performance of the system. The best cross-lingual EL performance is achieved by combining Google and SMT translation results (listed in Table 1). The result in Table 2 shows the evaluation of monolingual retrieval on Chinese KB. The overall performance of monolingual retrieval strategies is better than that of cross-lingual retrieval. Although we incorporate multiple translation resources, some Chinese entities like person names do not receive a proper translation. Therefore we adopt the monolingual retrieval strategy for Chinese entities.

6 Results

6.1 Slot Filling

Our slot filling system is a pipeline which progressively refines answers: first we retrieve relevant documents, then rank sentences from these documents, and finally rank entities in the relevant sentences using a remotely supervised relation classifier (Chrupała et al., 2010). For some slot types we did not have enough data to learn a classifier. We created the following two runs which differ in the strategy they adopt for these slot types with no relation classifier:

1. Precision-oriented strategy. We do not return any answers for those slot types.

2. Recall-oriented strategy. We return the answers from the top ranking sentences, provided they have the named entity type which matches the slot type.

Table 3 shows the evaluation results of the two runs. It can be seen that the precision-oriented run 1 achieved a much better overall F-score.

6.2 Cross-lingual Entity Linking

Our three runs submitted to the cross-lingual Entity Linking task adopt the retrieval method described in Section 3. The baseline run **lsv1** only employs the coarse clustering, whereas the runs **lsv2** and **lsv3** employ HAC clustering on the result of the baseline using different parameter settings. The configurations are as follows:

- **lsv1:** simply collect entities with the same literal names into one cluster
- **lsv2:** at most 2 sense clusters for top relevant documents of each identical entity
- **lsv3:** 50 documents in each sense cluster for each identical entity

	Precision	**Recall**	**F1**
Run 1	0.21	0.13	0.16
Run 2	0.09	0.14	0.11

Table 3: Evaluation results for slot filling

Run	**Prec.**	**Recall**	**F-score**
lsv1	0.514	0.581	0.545
lsv2	0.515	0.577	0.544
lsv3	0.519	0.579	0.547

Table 4: Performance of different system configurations on the 2011 cross-lingual entity linking test data

Considering the performance of different clustering configurations in Table 4, there is no (notable) difference between the runs **lsv2**, **lsv3** and **lsv1**.

	All	**PER**	**ORG**	**GPE**
English Entities				
Overall	0.51	0.46	0.58	0.46
in-KB	0.34	0.41	0.39	0.21
NIL	0.80	0.84	0.81	0.77
Chinese Entities				
Overall	0.65	0.57	0.77	0.62
in-KB	0.44	0.30	0.59	0.47
NIL	0.80	0.70	0.86	0.98

Table 5: Micro-averaged accuracy of Chinese and English entities on TAC 2011 training data.

Table 5 summarizes the overall and entity-type dependent accuracies of Chinese and English entity linking results. Unlike most teams of the for-

[10]Its motivation is similar to the MRR metric, which only considers the rank of the top ranked.

mer evaluations we do not employ a KB node candidate generation method. We retrieve all articles from the Wikipedia collection, which is much larger than the KB collection. The low linked KB accuracy of both languages indicates that it is really hard for retrieval-based methods to find the right reference from a large collection of documents even with query expansion. Regarding each entity type, the performances on English GPE entities and Chinese PER entities suggest a limited domain adaptation of the retrieval method. It is better to adopt specific modules for different entity types. The title-to-id mappings result in a fair NIL accuracy for both languages as we expected, however it is too strict to verify only top one entry and simply ignore desirable targets in the following top ranked KBs.

7 Conclusion

We have developed a cross lingual entity linking system for TAC KBP 2011, which uses a parallel monolingual architecture mainly consisting of document retrieval and entity clustering modules. We show the feasibility of using Chinese Wikipedia as the knowledge base to connect cross-lingual knowledge and avoid the propagation of translation errors to the retrieval module. To reach better solutions and deeper understandings of the CLEL problem, we propose the following future work:

- utilizing entity filtering and reranking model to improve the retrieval performance;
- further investigating cross lingual document retrieval;
- one alternative architecture is to perform entity clustering first and use a cluster-based retrieval model to link KB entries with a cluster of entities.

References

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2007. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 64–69, Prague, Czech Republic, June. Association for Computational Linguistics.

Grzegorz Chrupała, Saeedeh Momtazi, Michael Wiegand, Stefan Kazalski, Fang Xu, Benjamin Roth, Alexandra Balahur, and Dietrich Klakow. 2010. Saarland university spoken language systems at the slot filling task of tac kbp 2010. In *Proceedings of TAC 2010*.

Hengi Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the TAC 2010 Knowledge Base Population Track. In *Proceedings of Text Analytics Conference (TAC2010)*.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 120–127, New York, NY, USA. ACM.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. *Introduction to Information Retrieval*.

H. Turtle and W. B. Croft. 1990. Inference networks for document retrieval. In *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '90, pages 1–24, New York, NY, USA. ACM.

Youzheng Wu, Jun Zhao, Bo Xu, and Hao Yu. 2005. Chinese named eentity recognition based on multiple features. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 427–434, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaofeng Wu, Junhui Li, Jie Jiang, Yifan He, and Andy Way. 2011. DUC multi-Engine MT System for CWMT'2011. In *Proceedings of China Workshop on Machine Translation CWMT'2011*.

Ying Zhao and George Karypis. 2002. Evaluation of hierarchical clustering algorithms for document datasets. In *Data Mining and Knowledge Discovery*, pages 515–524. ACM Press.