CMU System for Entity Discovery and Linking at TAC-KBP 2015

Nicolas Fauceglia, Yiu-Chang Lin, Xuezhe Ma, and **Eduard Hovy**
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213, USA
{fauceglia, yiuchanl, xuezhem, hovy}@cs.cmu.edu

Abstract

This paper describes CMU's system for the Tri-lingual Entity Discovery and Linking (TEDL) task at TAC-KBP 2015. Our system is a unified graph-based approach which is able to do concept disambiguation and entity linking simultaneously, leveraging the ontology built on Freebase. The results show that our system achieves competitive results for Chinese and Spanish.

1 Introduction

Typically, a EDL system is required to tackle three sub-tasks: (i) Entity Discovery – detecting mentions of entities appearing in a document; (ii) Entity Linking – linking each entity to the most suitable entry in a reference Knowledge Base (KB), and (iii) NIL Entity Clustering – clustering NIL mentions, which do not have corresponding KB entries.

The Tri-lingual Entity Discovery and Linking (TEDL) task at TAC-KBP 2015 extends the EDL task of 2014 from two perspectives. From the data perspective, TEDL adds to the monolingual English EDL two new languages, Chinese and Spanish. It also introduces a new, much larger, Knowledge Base (KB) – Freebase. From the perspective of task design, TEDL adds two new entity types, natural locations (LOC) and facilities (FAC), and introduces person nominal mentions.

Our system for TEDL task consists of two main steps. First, we process the whole Freebase, representing it as a directed weighted graph, then computing semantic signature for each vertex (Section 2). We also need to do preprocessing for input data. Second, we build an end-to-end system for entity discovery and linking across three languages (Section 3). We use Babelfy[1] as the backbone of our system and extend it to be suited for the TEDL task. Briefly, our system is different from Babelfy in the following points:

- Our system uses the Freebase's Ontology directly, instead of merging WordNet into KB.
- For the construction of semantic signature, we use the algorithm of Personalized PageRank with node-dependent restart (Avrachenkov et al., 2014), instead of Random Walk with Restart (Tong et al., 2006) (see Section 2.1.3 for details).
- We modify the candidate extraction method and extend it to Chinese and Spanish.
- We introduce edge weights to semantic interpretation graph (Section 3.2).
- We propose a new rule-based entity type inference method (Section 3.3).

Our results show that our system is competitive for Chinese and Spanish, comparing with other systems in TEDL task at TAC-KBP 2015 (Section 4).

2 Data Preparation

In this section, we describe the preparation of data, including constructing the Freebase graph, computing Semantic Signatures, and preprocessing the input documents to transform them into Fragments.

2.1 Freebase

The reference knowledge base used in TEDL is a January 2015 snapshot of English Freebase, which includes about 81M nodes (mids) and 290M relations. Freebase is a semantic Knowledge Base with an ontology built on it. Thus, Freebase not only includes named entities such as people, places or organizations, but also concepts such as *Person, Location* and *Time*. This motivated us

[1] http://babelfy.org

to build a system based on Babelfy (Moro et al., 2014), which is an entity linking system leveraging the ontology in BabelNet, which is WordNet (Leacock and Chodorow, 1998).

2.1.1 Preprocessing

We observed that a significant part of the dump contained information about the Entertainment Business, e.g. Music (artists, recordings, etc.), TV Shows (series, directors, actors, etc.), Video Games, or Books, which are not closely related to the domain of TEDL task. To focus on the domain and deal with the size, we remove all Freebase entities naming music, books, films, TV programs, and video games. This yields a smaller KB with around 37M nodes and 123M relations.

The reason why we shrink the KB is that we are not supposed to build a open-domain system for the TEDL task. The entities of music or books (and some others) are not related to the domain of our task and would introduce noise for both candidate extraction (Section 3.1) and entity linking (Section 3.2). Another reason is that removing irrelevant nodes results a much smaller-scale KB, making our system much more efficient.

After the implementation of our system, we found that there are some nodes that are used for maintaining freebase only, such as nodes for managing Freebase users and permissions. We did not get a chance to remove these "meaningless" nodes from our KB and construct a new graph. This will be left for future work.

2.1.2 Graph of Freebase

We first represent Freebase as a directed weighted graph, where the vertices in the graph are the entities and concepts in Freebase, and the edges are the relations between them. For relations, the only information we use is that of the vertices that these relations connect, while ignoring the relation predicates. Following Moro et al. (2014), the weight of each edge is calculated as the number of triangles (cycles of length 3) that this edge belongs to. To implement the graph, we used the WebGraph framework (Boldi and Vigna, 2004).

2.1.3 Semantic Signature

A *semantic signature* is a set of highly related vertices for each concept and entity in Freebase graph. To calculate semantic signatures, we first compute the transition probability $P(v'|v)$ as the normalized weight of the edge:

$$P(v'|v) = \frac{w(v, v')}{\sum\limits_{v'' \in V} w(v, v'')}$$

where $w(v', v)$ is the weight of the edge $(v \rightarrow v')$. With the transition probabilities, Semantic Signatures are computed using the algorithm of Personalized PageRank with node-dependent restart (Avrachenkov et al., 2014). It should be noted that the algorithm performed by Moro et al. (2014) to create semantic signatures is Random Walk with Restart (Tong et al., 2006), which is simulation of the Personalized PageRank algorithm used in our system. Finally, vertices with pagerank score higher than a threshold (η) are kept to build the semantic signature. In our system we set $\eta = 10^{-4}$.

2.2 Input File

For each language, two kinds of data, Newswire and Discussion Forum are given in *xml* format. As described in the task definition, every document is represented as a UTF-8 character array and begins with the <DOC> tag. The "<" character has index 0 and offsets are counted before XML tags are removed. Therefore, to preserve the offset for each sentence, a line-by-line file reader is implemented instead of using an *xml* file parser.

In Newswire data, the tags are relatively simple and clean compared to Discussion Forum. The news' headline and paragraphs are extracted between "<HEADLINE>, </HEADLINE>" and "<P>, </P>" tags, respectively. In discussion forum data, similarly, the headline and posts are obtained between "<headline>, </headline>" and "<post>, </post>" tags. The author whose linking result is always NIL of each post is detected at the same time. However, in each post, there might exist more than one *quote*, which are repetitive text from previous posts. Quote removal is therefore a followed-up step after post extraction. Moreover, any text that are between "&lt" and "&gt"tags or in *URL* format are removed from the post as well.

3 System Architecture

Our end-to-end EDL system includes candidate extraction (Section 3.1, entity linking (Section 3.2), type inference (Section 3.3), and NIL entity clustering (Section 3.4). We use the Stanford CoreNLP pipeline (Manning et al., 2014) for preliminary steps, and adapt and extend (Moro et al., 2014) for entity extraction and linking.

Similar to the Babelfy system, our system does concept disambiguation and entity linking at the same time. This is because, as mentioned above, our system is able to exploit the ontology of Freebase.

3.1 Candidate Extraction

The task of the Candidate Extractor (CE) is, given an input string, return all the possible entities in the graph that could be associated with a substring of the input string. When processing the Freebase Dump, we keep an additional parallel data structure holding information about the names of the graph entities. For each Freebase entity we keep string labels provided by 3 predicates: name, label and alias. In the original Freebase Dump, string values have an associated language, so we only kept the values in our three languages. We implemented this name map as a Lucene index, that given a string returns all the nodes in our graph that have a label (name, label, alias) that contains the given string.

It is worth mentioning that for our multilingual task, this is the only part that deals with languages: we have different implementations of this component, one for each language. For English and Spanish, the approach is similar to the Babelfy implementation: perform POS tagging for each input sentence, and choose n-grams of length 1 to N (we used N = 5), that contain at least one NOUN, and which do not end or start in prepositions, conjunctions, punctuation, among others. For each one of these candidate fragments, we query the name index to retrieve all possible entities. For Chinese, the approach is completely different: we work at a character level, and we start with strings of N characters (we used N = 10) and search in the name index, and if there is no match, we search with N-1, and so on, until we have a match, and return each match as a Candidate Meaning.

Once the candidates have been extracted from the input document, the rest of the pipeline works in graph-space and does not depend on the input language. This makes it relatively easy to add a new language, provided Freebase has names for the new language.

	Type in Freebase
PER	people.person
GPE	location.country location.administrative_division location.statistical_region
ORG	organization.organization
LOC	location.location
FAC	architecture.structure

Table 1: Rules applied to distinguish between the 5 entity types.

3.2 Entity Linking

3.2.1 Semantic Interpretation Graph Construction

The semantic interpretation graph is constructed using a procedure similar to Moro et al. (2014). The difference is that we introduce edge weights to this graph – the weight of the edge between two vertices (v_1, f_1) and (v_2, f_2) is defined as the pagerank score between v_1 and v_2 in the Freebase graph.

3.2.2 Graph Densification

We implemented the graph densification algorithm presented in Moro et al. (2014), too. Basically, at each step of graph densification, we first find the most ambiguous mention, the one has the most number of candidate entities. Then we remove the least possible candidate entity from the most ambiguous mention, the one has smallest score. In our system, the score of a vertex (v, f) in the semantic interpretation graph is slightly different from the one in Babelfy – we use the sum of the incoming and outgoing edge weights instead of the sum of incoming and outgoing degree. Formally, the score of the vertex (v, f) is:

$$score((v, f)) = \frac{w(v, f) \cdot sum((v, f))}{\sum\limits_{(v', f)} w(v', f) \cdot sum((v', f))}$$

where $sum((v, f))$ is the sum of the incoming and outgoing edge weights of (v, f) and $w((v, f))$ is the number of fragments the candidate entity v connects to.

The above steps are repeated until every mention has less than a certain number (μ) of candidate entities. Finally, we link each mention f to the highest ranking candidate entity v^* if $score((v^*, f)) > \theta$, where θ is a fixed threshold.

	NER			Linking			Clustering			rank
	P	R	F1	P	R	F1	P	R	F1	
Eng	46.8	51.7	49.1	32.7	36.1	34.3	42.0	46.3	44.1	8th
Cmn	50.0	61.4	55.1	44.5	54.7	49.1	48.9	60.1	53.9	4th
Spa	60.2	60.8	60.5	47.3	47.7	47.5	54.1	54.5	54.3	4th
All	50.2	56.7	53.2	48.2	36.7	41.7	43.5	49.1	46.1	6th

Table 2: The offical results precision, recall and F1 measures over all three languages for our best run for three key metrics: strong typed mention match (NER), strong all match (Linking) and mention ceaf (Clustering), together with the relative ranking of our system and the best results in TEDL at TAC-KBP 2015.

3.3 Entity Type Inference

Entity type is obtained from each entity's *Types* in Freebase. We define different rules to determine such entity types. If a candidate entity has the predefined types (2nd column in Table 1), its entity type is assigned as the corresponding value (1st column). Else, it is not treated as an entity and it is discarded.

3.4 NIL Entity Clustering

The final step in our system is clustering NIL entities. In our system, we simply merge candidates with exactly the same name spelling.

4 Experiments

We submitted two runs for TEDL task. For both of the two runs, we extract top 100 candidate entities for each mention ($K = 100$), and the ambiguous parameter $\eta = 10$. The different between the two runs is the threshold parameter (θ) for entity link. The first run set $\theta = 4.0$ and the second one set $\theta = 2.5$. According to the official results, the first run is slightly better than the second one.

Table 2 shows the results precision, recall and F1 measures over all three languages for our best run for three key metrics: strong typed mention match (NER), strong all match (Linking) and mention CEAF (Clustering), together with the relative ranking of our system. According to Table 2, our system achieves competitive results (ranking 4th) for Chinese and Spanish. For English, however, our results are surprisingly low. We believe there are bugs in our submitted English results. We will update our results in the future after we fix them. Because of the low results on English, our systems ranked the 6th place for the overall results on the three languages.

5 Conclusion and Future Work

We build a unified graph-based system for the TEDL task at TAC-KBP 2015, inspired by Babelfy. Our system obtains competitive results for Chinese and Spanish. Unfortunately, our system got unreasonablely low results on English. We will fix the bugs for English and update our results in the future.

There are a few possible extensions to our approach that we can explore in the future. First, our system is not able to discover and linking nominal mentions, which make up around 5% of the test data. One possible way to solve this problem is to utilize entity coreference system. Second, the candidate extraction method, particularly for Chinese, needs to be improved. Another possible direction to improve our approach is to collaborate with cross-document coreference systems.

References

Konstantin Avrachenkov, Remco Van Der Hofstad, and Marina Sokol. 2014. Personalized pagerank with node-dependent restart. In *Algorithms and Models for the Web Graph*, pages 23–33. Springer.

Paolo Boldi and Sebastiano Vigna. 2004. The webgraph framework i: compression techniques. In *Proceedings of the 13th international conference on World Wide Web*, pages 595–602. ACM.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. 2006. Fast random walk with restart and its applications. In *Proceedings of ICDM 2006*. IEEE.