BUPTTeam Participation at TAC 2017 Knowledge Base Population

Yongmei Tan, Hu Yang and Xiaoguang Li

Center for Intelligence Science and Technology and Technology
Beijing University of Posts and Telecommunications
Beijing, China
{ymtan, hyang, xgli }@bupt.edu.cn

Abstract

This paper describes the BUPTTeam system submitted to the Trilingual Entity Discovery and Linking (TEDL) task in 2017 TAC Knowledge Base Population (KBP) contests. The architecture of our EDL system is described as Figure 1. It includes the following five components: 1) preprocessing; 2) mention recognition; 3) candidates generation; 4) candidates selecting; 5) clustering.

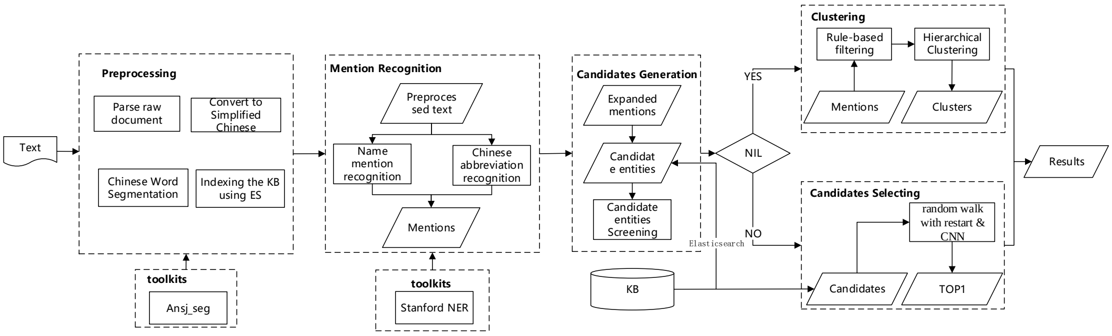

Figure 1: System Architecture

1) Preprocessing

There are many xml tags in raw text, which influence mention recognition and the parts between “<quote>” and “</quote>” are also redundancy. So we remove these tags and parts. There are many traditional Chinese words in raw texts and knowledge base. Text processing tools is good at processing simplified Chinese so that we convert traditional Chinese into simplified Chinese. We use Ansj seg for Chinese word segmentation and Elasticsearch for indexing the KB described in (Tan et al., 2016).

2) Mention recognition

We use Stanford NER to recognize most mentions. In addition, mentions representing authors can be directly extracted from the raw texts. Their type is PER and linking results are always NIL. In Chinese, two or more abbreviations representing states or provinces are often wrote as a whole, such as: "中美", where "中" refers to “China”, "美" refers to “the United States”. This phenomenon influences the performance of mention recognition, and so we collect the word list of provincial and national abbreviations to recognize those mentions.

3) Candidates generation

This step attempts to search potentially correct entities for mentions from Freebase. We generate a candidate set E_m for each mention m by Elasticsearch. It’s very hard to choose the right one from too many candidates. In order to scale the candidate set as small as possible, we filter the candidates according to some constraints.

4) Candidates selecting

We combine convolution neural network and random walk with restart algorithm for entity linking. Firstly we use convolution neural network to extract features of mention's context and entity's description text in Freebase and use random walk with restart algorithm to obtain the semantic features of mention and entity. Secondly we merge these features as the expression of the mention and entity. Finally, we using a full connection layer to calculate the distance between the mention and entity.

5) Clustering

If the candidate set E_m 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:

1) All NIL mentions are divided into five types (PER, ORG, GPE, LOC and FAC);

2) If mention m_i and mention m_j meet any of the following conditions, we merge them into the same cluster:

- Mention m_i and mention m_j have the same surface string;
- Mention m_i is the prefixes or suffixes of mention m_j;
- Mention m_j is the prefixes or suffixes of mention m_i;

Secondly, according to Harris's distributed hypothesis, if two words have similar context, their semantics are similar. We convert the mention's context into vector representation and use hierarchical clustering algorithm for clustering.