Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 108C (2017) 2287–2291

Procedia Computer Science

International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland

Recognizing Compound Entity Phrases in Hybrid Academic Domains in View of Community Division

Yang Yan[1,2], Tingwen Liu[1], Quangang Li[1], Jinqiao Shi[1], Li Guo[1], and Yubin Wang[1]

[1] Institute of Information Engineering ,Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
{yanyang9021,liutingwen,liquangang,shijinqiao,guoli,wangyubin}@iie.ac.cn

Abstract

Classifying compound named entities in academic domains, such as the name of papers, patents and projects, plays an important role in enhancing many applications such as knowledge discovering and intelligence property protection. However, there are very little work on this novel and hard problem. Prior mainstream approaches mainly focus on classifying basic named entities (*e.g.* person names, organization names, twitter named entities, and simple entities in specific sci-tech domain etc). We use context templates to extract the possible candidate compound entities roughly, which is used for reducing searching space of text splitting. We reduce the text splitting problem to the community division problem, which is addressed based on the dynamic programming strategy. The construction of indicative words set used in segment validating is reduced to the classical minimum set cover problem, which is also addressed based on dynamic programming. Experimental results on classifying real-world science-technology compound entities show that GenericSegVal achieves a sharp increase in both precision rate and recall rate by comparing with the supervised bidirectional LSTM approach.

Peer-review under responsibility of the scientific committee of the International Conference on Computational Science

Keywords: Named Entity Recognition, Template Extraction, Minimum Set Cover Problem , Community Division

1 Introduction and Relate Work

As an classical information extraction task, Named Entity Recognition (NER) aims at locating and classifying elements in text into predefined basic categories of entities such as the names of persons, organizations, locations, quantities, twitter entities *etc.* However, much less work have been done on classifying a novel and common type of named entity phrases, referred to as compound named entity in this paper. It is a fundamental task for knowledge acquisition in various domains such as science, technology, laws, business, etc. In these domains, analysts are most interested in those ever-increasing compound entities, including titles of sci-tech projects papers, patents, university foundations, all of which plays an important role in application such as intellectual property protection, open source intelligence analysis [11], scholar data discovery, machine translation *etc.*

In this paper, compound entities in hybrid academic domains are informally defined as complicated and embedded professional phrases that consist of multiple continuous words or basic entities such as

1877-0509
Peer-review under responsibility of the scientific committee of the International Conference on Computational Science
10.1016/j.procs.2017.05.014

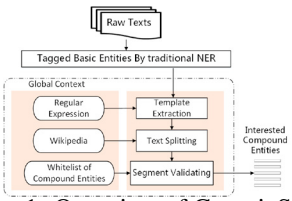

Figure 1: Overview of GenericSegVal

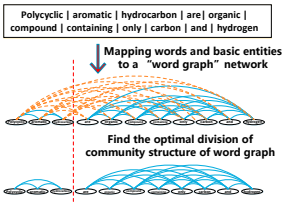

Figure 2: Text Splitting Example

title of papers, patents, university foundations, sci-tech projects etc. We note that traditional multi-word entity recognition is not applicable to our compound NER problem duo to two causes: First, compound named entities often span multiple words or entities in long distances and have huge structural differences. Second, compound entity is very sparse in Internet, as there are very little texts containing the same compound entity. Thus, it is impossible to locate compound entities through statistical learning algorithms. This paper tries to give the first and practicable approach on this problem. Figure 1 gives our weak supervised framework from community division perspective to achieve the recognition of compound entities. First,we use traditional POS tagging to label the basic entities among the raw texts. Second, we run rough extraction by regular expression produced by context pattern which collocate with compound entities. Third, we map the matched ordered words and basic entities to a undirected weighted "word-graph" network like figure 2. Then, we find the optimal dividing of network and verify wether each divided community constitute a compound entities. Finally, we verify each community of ordered words and preliminary entities wether consists of a compound entities.

Research on NER starts at 1990s. Prior NER work mainly focus on identifying the names of persons, organizations, locations *etc.* There exists burgeoning specific NER tasks such as query log NER [3], biomedical NER [14, 2], chemical NER [9], social network events NER [5, 4, 7] *etc.* Our work is also related to NLP task of quality phrase (muti-word semantic unit) mining [6] and word sequence segmentation. Earlier works of mutiword phrase segmentation use predefined POS patterns or annotations to learn rules on POS-tagged dataset [8, 12]. Those types of methodology is not generic in domains and dependant on expensive manual notation, which make it challenging to their widespread application. Despite great efforts have been made on multiwords named entities recognition and segmentation, most of them focus on simple entities or sci-tech terms on specific domains and heuristic rules. The problem of recognizing compound entities in hybrid domains consisting of multi-entities, noun phrases, conjunctions, which has complicated structure and longer length, is still far from being solved.

2 Our Approach

2.1 Template Extraction

We use Gibbs sampling to compute each word's topic distribution $p(z = k|w)$ in Equation 1 iteratively. Its each step is updated by the new $n_k^{(t)}$ and $n_m^{(k)}$, which is sampled from multinomial distributions generated by $Dir(\vec{\alpha} + \vec{n}_k)$ and $Dir(\vec{\beta} + \vec{n}_m)$. We take the highest probability of words $p(w|z = k)$ in each topic k as the template words when Gibbs sampling algorithm converge. We use these template words to scan the left context reversely from compound entities and right context from forward direction. The matched template words constitute the regular expression for context templates.

$$p(z = k|w) \propto \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_m^{(k)} + \alpha_k)} \cdot \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V}(n_k^{(t)} + \beta_t)} \tag{1}$$