A Hybrid Method for Chinese Entity Relation Extraction

Hao Wang

Joint work with Z.Y. Qi, H.W. Hao, B. Xu

Computational-Brain Research Center
Institute of Automation, Chinese Academy of Sciences

Outline

- **Introduction**
- **Our Framework**
  - Chinese Semantic Knowledge Base Construction
  - High-frequency： the improved selecting candidate sentences method
  - Low-frequency： the heuristic rules based method
- **Experiments**
- **Conclusions**

Introduction

- Task Definition
  - The input of this problem is multi-structured data, including structured data (infobox form), semi-structured data(tables and lists) and non-structured data (**free text**).
  - The output is <entity1, relation, entity2>, we call it entity relation which is represented in triples.

- An example:
  - given the sentence “姚明出生于上海” (Yao Ming was born in Shanghai) as input, the relation extraction algorithm should extract “<姚明, 出生地, 上海>”(Yao Ming, birthplace, Shanghai)from it.

Introduction

- Significance:
  - These fact triples can be used to build a large, high-quality knowledge base, which can benefit to a wide range of NLP tasks, such as **question answering**, **ontology learning**, **knowledge graph** and **summarization**.

- Challenge of Chinese language:
  - Current research mainly focuses on the processing of English resource and the study conducted on Chinese corpus is less.
  - Chinese language need word segmentation, and the proper nouns don't have the first letter capitalized. The Chinese entity relation extraction is more difficult and more challenging.

Our Framework

- Our framework:
  - We first build a Chinese semantic knowledge base, using the corpus of Douban web pages, Baidu encyclopedia and Hudong encyclopedia.
  - An improved selecting candidate sentences method trained by CRF model is used to extract high frequency relation words of the knowledge base.
  - The method based on some simple rules and knowledge base is used to extract low-frequency relation words.

Specifically, our contributions are:

- We propose candidate sentences selecting method, which can reduce the mistakes introduced by automatic tagging training data and improve the extraction performance.
- It's hard to get enough training data for some rare relations. Here, we propose the method based on some simple rules and knowledge base to extract these low-frequency relation words.

Our Framework

- Chinese Semantic Knowledge Base Construction
  - We extract the infobox knowledge from these corpus and represent them in triples format <arg1 , rel, arg2> , like< 中国，首都，北京>, and then store these triples in our knowledge base.
  - We should first traverse our knowledge base to get the frequency of this relation word.
  - If the frequency number is greater than 500, the corresponding relation is high-frequency; otherwise we regard it as low-frequency relation.

中文名	卧虎藏龙	主　演	周润发，杨紫琼，章子怡，张震
外文名	Crouching Tiger, Hidden Dragon	片　长	120 min
制片地区	中国，美国	上映时间	法国：2000年5月16日
导　演	李安	分　级	USA:PG-13
编　剧	王度庐，王蕙玲	对白语言	汉语普通话
类　型	爱情，动作，冒险，剧情	色　彩	彩色
		奖　项	奥斯卡最佳外语奖

Fig. 1. An Infobox from Baidu Encyclopedia

Our Framework

- Candidate sentences selecting method
  - Traverse the knowledge base to get the corresponding <arg1, rel, arg2> triples.
  - These triples are used to locate candidate sentences in wiki-page by a scoring model.
  - To generate testing data
    - do word segmentation and pos tagging for the candidate sentences which are chose as training data, and then choose the nouns and verbs
    - choose the top-n highest frequency words as key words;
    - these key words are used to determine the candidate sentences for extracting
  - For a triple <arg1, rel, arg2>, we train conditional random field model to label the arg2 in the testing data.
  - Finally convert the annotation results to entity relation triples.

Our Framework

❖ Candidate sentences selecting method

In preparing for training data step, there are two methods to score a sentence based on the given triple <arg1, rel, arg2>:

(a) Score method 1: $score1 = bArg1 * (b\,\mathrm{Re}\,l + 1) * bArg2$

(b) Score method 2: $score2 = (bArg1 + 1) * (b\,\mathrm{Re}\,l + 2) * bArg2$

- If arg1 appears in this sentence, then bArg1 = 1, otherwise bArg1 = 0.
- If arg2 appears in this sentence, then bArg2 = 1, otherwise bArg2 = 0.
- If rel appears in this sentence, then bRel = 1, otherwise bRel = 0.

In preparing for training data step, two methods to get the final candidate sentences from these highest score sentences:

(a) Selecting the highest score sentence first appeared in an article;

(b) Selecting all highest score sentences.

In preparing the data for extracting step, two methods to extract triples from the wiki-page content (testing data):

(a) Choosing all the sentences in the wiki-page content;

(b) Selecting some sentences from the wiki-page content based on keyword matching.

Our Framework

- The heuristic rules based method

Algorithm1: The Heuristic Rules based Entity Relation Extraction Algorithm

Input: The target relations, some entities, corresponding categories and unstructured content

Output: Entity relation triples <arg1,rel,arg2>, arg1 and arg2 are entities and rel is the relation

1 Begin
2 Confirm the template <class1, rel, class2> of a target relation; here class1 and class2 are the categories of unknown entities. For example, a given relation "director", we can confirm class1 is movie or teleplay and class2 is people.
3 Produce an entity library, which contain entities and corresponding categories.
4 Get the keywords of target relation by using domain knowledge.
5 Select candidate sentences, which should contain keywords, and entities of class1 and class2 in our entity library.
6 Generate some simple rules to extract the entity relation.
7 End

Experiments

- *The Comparison of Various Methods of Select Candidate Sentences*

(1) Choosing all the sentences in the wiki-page content, part of the experiment results:

Table 3. The extraction results of choosing all the sentences in the wiki-page content

Relation Words	Precision	Recall	F1	Relation Words	Precision	Recall	F1
Geo_area1	0.1570	0.1371	0.1463	Movie_Director1	0.1443	0.0833	0.1057
Geo_area2	0.1600	0.1421	0.1505	Movie_Director2	0.1633	0.0952	0.1203
Geo_area3	0.1486	0.1320	0.1398	Movie_Director3	0.1782	0.1071	0.1338
Geo_area4	0.1534	0.1371	0.1448	Movie_Director4	0.1458	0.0833	0.1061
Geo_district1	0.3118	0.2944	0.3209	EDU_Start_time1	0.3097	0.2909	0.3000
Geo_district2	0.3059	0.2640	0.2834	EDU_Start_time2	0.3117	0.2909	0.3009
Geo_district3	0.3333	0.3147	0.3238	EDU_Start_time3	0.3397	0.3212	0.3302
Ge_district4	0.3086	0.2741	0.2903	EDU_Start_time4	0.3333	0.3152	0.3240

Experiments

- *The Comparison of Various Methods of Select Candidate Sentences*

(2) Selecting some sentences from the wiki-page content based on keyword matching, part of the experiment results:

Table 4. The extraction results of selecting some sentences from the wiki-page content based on keyword matching

Relation Words	Precision	Recall	F1	Relation Words	Precision	Recall	F1
Geo_area1	0.3119	0.1726	0.2222	Movie_Director1	0.4833	0.1726	0.2544
Geo_area2	0.3736	0.1726	0.2361	Movie_Director2	0.5439	0.1848	0.2756
Geo_area3	0.2661	0.1675	0.2056	Movie_Director3	0.4110	0.1786	0.2490
Geo_area4	0.2623	0.1624	0.2006	Movie_Director4	0.5469	0.2083	0.3017
Geo_district1	0.4294	0.3706	0.3978	EDU_Start_time1	0.6993	0.6061	0.6494
Geo_district2	0.4000	0.2538	0.3106	EDU_Start_time2	0.7252	0.5758	0.6419
Geo_district3	0.4535	0.3959	0.4228	EDU_Start_time3	0.7329	0.6485	0.6881
Ge_district4	0.4031	0.2640	0.3190	EDU_Start_time4	0.7211	0.6424	0.6795

Experiments

- *The Comparison of Various Methods of Select Candidate Sentences*

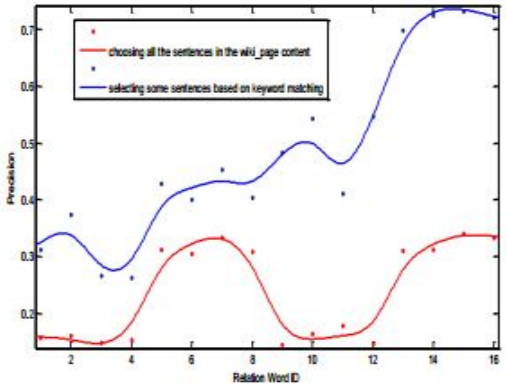

Fig. 2. The precision of different candidate sentences selecting methods

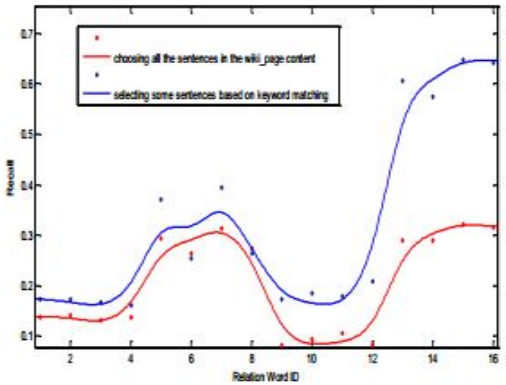

Fig. 3. The recall of different candidate sentences selecting methods

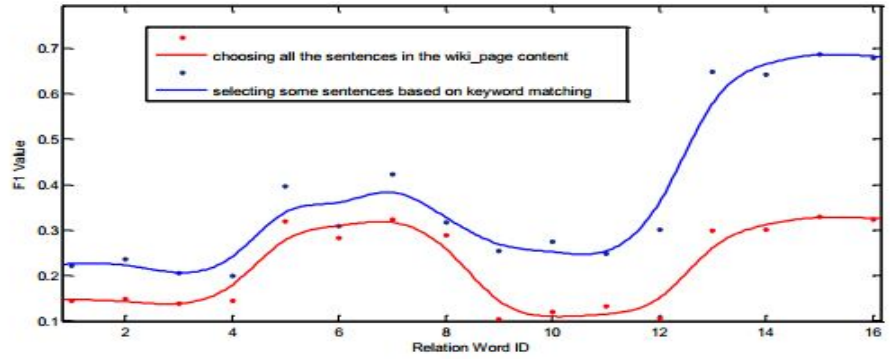

Fig. 4. The F1 Value of different candidate sentences selecting methods

Experiments

- *The Comparison of Various Methods of Select Candidate Sentences*
- 1, the real results should be better than shown above
- 2, the extraction results of different relation words vary a lot, some results are very good, but some are not.
- 3, the method of label 2 can get the highest precision and the method of label 3 can get the highest recall, and it is hard to conclude which method can get the highest F1 value.
- 4,The method of label 2 can get accurate and related training data, so this method can achieve the highest precision. The method of label 3 can get abundant training data to achieve the highest recall.

Experiments

- *The Competition of Sougou Web-based Entity Relation Extraction*

Table 5. Some example of Sougou Web-based Entity Relation Extraction Competition

Category	Relation	Sentence	Triples
人物	父母	冉甲男与父亲冉平一起担任编剧的电影《画皮2》备受期待。	<冉甲男，父母，冉平>
	夫妻	林妲怡与蒋家第四代蒋友柏结婚，婚后息影。	<林妲怡，夫妻，蒋友柏>
	兄弟姐妹	曾维信的奶奶胡菊花，是胡耀邦的亲姐姐。	<胡菊花，兄弟姐妹，胡耀邦>
书籍	作者	《沙床》当代高校生活的青春忏情录 作者：葛红兵。	<沙床，作者，葛红兵>
歌曲	作词	《幻想爱》是陈伟作词作曲，张韶涵演唱的一首歌曲。	<幻想爱，作词，陈伟>
	作曲		<幻想爱，作曲，陈伟>
	演唱者		<幻想爱，演唱者，张韶涵>
电影/电视剧	导演	李安导演的《卧虎藏龙》诠释了中国武侠的魅力。	<卧虎藏龙，导演，李安>
	编剧	电影海上烟云由柯枫自编自导。	<海上烟云，编剧，柯枫>
	原著	根据琼瑶原著《含羞草》改编的台湾电视连续剧《含羞草》。	<含羞草，原著，含羞草>
	演员	电视剧 《龙堂》由著名演员张丰毅、陈小春主演。	<龙堂，演员，张丰毅> <龙堂，演员，陈小春>
	原声音乐	电影《大兵金宝历险记》主题曲是刘佳演唱的美丽国。	<大兵金宝历险记，原声音乐，美丽国>

Experiments

❖ *The Competition of Sougou Web-based Entity Relation Extraction*

- We adopt different methods to extract different frequency relation words.
- An improved selecting candidate sentences method trained by conditional random field model is used to extract ***high-frequency*** relation words of the knowledge base.
- And the method based on some si mple rules and knowledge base is used to extract ***low-frequency*** relation words.

- Finally we submitted a total of 364944 triples. The precision is 46.3% and we rank the fourth place.

CASIA

Thanks For Your Time !