中国人名识别中规则抽取的一种基于实例的方法

方晓珊[1] 盛焕烨[2]

[1]（上海交通大学计算机科学与工程系，上海 200030）；

[2]（上海交通大学，上海 200030）

Email: fang-xs@cs.sjtu.edu.cn;

hysheng@mail.sjtu.edu.cn

摘要：基于规则的命名实体识别系统通常有很好的性能。但是书写规则是一项费时的工作，而且一些出现频率较低的规则容易被忽略。本文介绍一种从实例中获取词典-句法规则的学习方法。我们基于预定义的高频规则自动抽取低频规则并且手工确认新规则和识别出的人名。学到的规则可以加入基于规则的命名实体识别系统的规则库中以识别更多的人名。试验显示使用新规则后从 Chinese Treebank 中抽取到的人名数增加了 14.3%。

关键词：命名实体识别，词典-语义规则

Pattern Extraction for Chinese Person Name Recognition Using a Example Based Approach

Xiaoshan Fang[1], Huanye Sheng[2]

[1]*Computer Science & Engineering Department, Shanghai Jiao Tong University, Shanghai 200030, China,* fang-xs@cs.sjtu.edu.cn

[2]*Shanghai Jiao Tong University, Shanghai 200030, China,* hysheng@mail.sjtu.edu.cn

Abstract: Handcrafted rule-based systems for the NE tasks attain high levels of performance, but it is a time consuming work to construct rules. It is easy to find the frequently occurring patterns, whereas those low-frequency patterns are difficult to be discovered. This paper presents an example based learning method for the acquisition of lexico -syntactic patterns. These learned patterns can be added to a pattern-based rule set for named entity recognition. We automatically extracted low frequency patterns based on the predefined high-frequency patterns and manually validated the new patterns and outputs of terms. The experiments show that the number of person names extracted from the Chinese Treebank increased by 14.3% after using the new patterns.

Keywords: named entity recognition, lexico – syntactic pattern

1 Introduction

Handcrafted rule-based systems for the NE tasks attain high levels of performance, but it is a time consuming work to construct rules. It is easy to find the frequently occurring patterns, whereas those low-frequency patterns are difficult to be discovered. My work used machine-learning approaches to automatically learn a set of low-occurrence patterns for external evidence of Chinese person names from a corpus with the predefined high-occurrence patterns. The combination of a machine learning approach and a handcrafted approach improves the system's efficiency. The experiments showed that the number of person names extracted from the Chinese Treebank increased by 14.3% after the use of new patterns.

In general, the named entity recognition task involves the recognition of entity names (for people and organizations), place names, temporal expressions, and certain types of numerical expressions. In this paper we use the Chinese person name as our experiment object for the named entity recognition.

Chinese person names have no upper and lower case, which makes it more difficult than English and European languages to recognize. In addition, it is well known that there is no space between Chinese words; so it is not evident from the orthography where the word boundaries are. Therefore, the performance of Chinese extraction systems is partly affected by their word segmentation module.

Little work has been done in the automatic pattern acquisition for Chinese NE tasks. Hearst (1992, 1998) reports a method using lexico-syntactic patterns to extract lexical relations between words from English texts. Morin (1999) intended to bridge the gap between term acquisition and thesaurus construction by offering a framework for organizing multi-word candidate terms with the help of automatically acquired links between single-word terms. Landau introduced supervised and unsupervised methods for extracting semantic relations between words.

In section two of this paper, the method of pattern acquisition from corpus is presented. Then, section three deals with the result of experiments and an evaluation of the frequency patterns. Finally, briefly describe our future work.

2 Pattern Extraction Algorithm

Inspired by Hearst (1992, 1998), Our procedure of discovering new patterns through corpus exploration, is composed of the following eight steps:

(1) Collect the context relations for person names, for instance person name and verb, title and person name, person name and adjective.

(2) For each context relation, we use the high occurrence pattern to collect a list of terms. For instance, for the relation of title and person name, with a pattern NN+NR, we extract the terms of title, for example, 记者 (reporter), 选手 (team player), etc. Here NN+NR is a lexico – syntactic pattern found by a rule writer. NN and NR are POS tags in the Corpus. NR is proper noun. NN includes all nouns except for proper nouns and temporal nouns.

(3) Validate the terms manually.

(4) For each term, retrieve sentences contain this term. Transform these sentences to lexico-syntactic expression.

(5) Generalize the lexico-syntactic expressions extracted in last step by clustering the similar patterns.

(6) Validate the candidate lexico-syntactic expressions.

(7) Use new patterns to extract more person names.

(8) Validate person names and go to step 3.

Based on this method, we learned five new patterns for the relation title - person name from twenty-five texts in the Chinese Penn Treebank. Use all the six patterns we extract 120 person names form these texts. 15 of them are new. We will give a detail description in section 3, experiments. This new person names can also be used for

person name thesaurus construction.

2.1 Automatic Classification of Lexico-syntactic Patterns

Hearst introduced an algorithm that automatically acquires lexico - syntactic patterns by classifying similar patterns. We used it in the fifth step in the above algorithm.

As described in reference [3], Lexico - syntactic expressions 1 and 2 are required from the relation HYPERNYM:

1. NR and in NR such as LIST
2. NR such as LIST continue to plague NR

They can be abstracted as:[①]

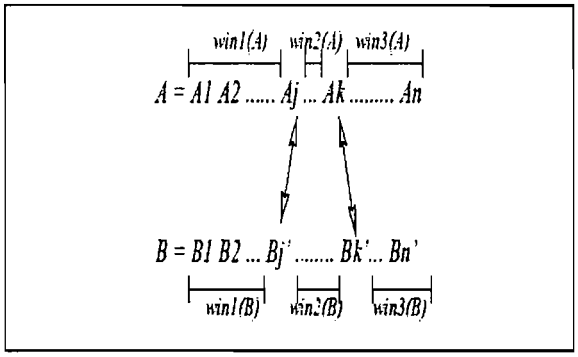
$$A = A_1 A_2 \cdots A_j \cdots A_k \cdots A_n \text{ with } \begin{cases} RELATION(A_j, A_k) \\ k > j+1 \end{cases} \quad (1)$$

And

$$B = B_1 B_2 \cdots B_{j'} \cdots B_{k'} \cdots B_{n'} \text{ with } \begin{cases} RELATION(B_{j'}, B_{k'} \\ k' > j'+1 \end{cases} \quad (2)$$

Following is a function measuring the similarity of lexico-syntactic expressions and that relies on the following hypothesis:

Fig. 1. Comparison of two lexico – syntactic expressions



Hypothesis 2.1 (Syntactic isomorphy) If two lexico-syntactic expressions A and B indicate the same pattern then, the items A_j and Bj', and the items Ak and Bk' have the same syntactic function.

Let Win1 (A) be the window built from the first through j-1 words, Win2(A) the window built from words ranking from j+1th through k-1th words, and Win3(A)be the window built from k+1th through nth words (see Figure 2). The similarity function is defined as follows:

① A_i is the ith item of the lexicon – syntactic expression A, and n is the number of items in A. An item can be a lemma, a punctuation mark, a symbol, or a tag (NR, LIST, etc.) The relation k > j+1 states that there is at least one item between A_j and A_k.

$$Sim(A,B)=\sum_{i=1}^{3} Sim(Win_i(A),Win_i(B))$$

with

$$\begin{cases} Win_1(A)=A_1A_2\cdots A_{j-1} \\ Win_2(A)=A_{j+1}\cdots A_{k-1} \\ Win_3(A)=A_{k+1}\cdots A_n \end{cases} \text{ and } \begin{cases} Win_1(B)=B_1B_2\cdots B_{j'-1} \\ Win_2(B)=B_{j+1}\cdots B_{k'-1} \\ Win_3(B)=B_{k+1}\cdots B_{n'} \end{cases}$$

The function of similarity between lexico-syntactic patterns Sim (Wini (A), Win (B)) is defined experimentally as function of the longest common string.

All lexico-syntactic expressions are compared two by two by previous similarity measure, and similar lexico-syntactic expressions are clustered. Each cluster is associated with a candidate lexico-syntactic pattern. For instance, the sentences introduced earlier generate the unique candidate lexico – syntactic pattern:

3. NR and in NR such as LIST

3 Experiments

We use Chinese Penn Treebank as experimental corpus, which published by the Linguistic Data Consortium (LDC), as training corpus.

3.1 Relations

We consider five relations. They are:

- ✧ Title and Person Name, e.g. 记者 (reporter) - 黄昌瑞 (Huang Chang-rui) (from Chinese treebank, text 325)
- ✧ Person Name and Verb, e.g. 狩野 (Shou Ye) - 强调 (emphasize) (from Chinese treebank, text 318)

Person names are often followed by verbs. These verbs are processed as the right boundary of PNs in NE recognition task. There are one character verbs and two character verbs. For example:

One character verb: 说(say), 讲(speak), 谈(talk).

Two character verbs: 报导(report), 会见(meet), 介绍(introduce).

- ✧ Person Name and Adjective, e.g. 美国的(American) - 琳达尔波特(An Ameican person name) (from Chinese treebank, text 314), or 年逾古稀的 (seventy years of age)– 张学良(Zhang Xue-liang).

Some special adjectives used before PNs are clues for NE recognition. For example, 年逾古稀的叶剑英元帅 (Marshal Ye Jian-yin who is seventy years of age).

- ✧ Person name and Conjunctions, e.g. 伏明霞(Fu Ming-xia) - 和(and) - 池彬(Chi Bing) (from Chinese treebank, text 325).

Conjunctions that found in Chinese Treebank are 、, ，, 和, 同 and 及. These punctuations connect PNs in the text.

✧ Location names and organization name used before PNs, like 太原钢铁公司(Tai Yuan Steel Company)李双良(Li Shuang - liang), or 中国女子排球队(Chinese women Volleyball team)郎平(Lang Ping), are also useful clues for person name recognition. I will consider these useful clues in my future work.

These relations describe some useful context information for named entity recognition. In this paper we only consider extracting the lexico – syntactic patterns according to the relation title and person name.

3.2 Patterns and pattern classification

Table 1 shows the patterns extracted for the Title – Person Name.

Table1. A predefined pattern and the extracted patterns

Pattern	Example
NN NR	选手 player 里查德 Richard
NN NR 和 NR	选手 player 兰卫 Lan Wei 和 and 陈晟 Chen Hao
NN 是 NR 和 NR	选手 player 是 is 李大双 Li Da-shuang 和 and 黄华东 Huang Hua-dong
NN NR、NR	选手 player 米切尔 Micheal、杜蒙德 Dumond
NN NR NR	记者 reporter 黄昌瑞 Huang Chang-reui 杨爱国 Yang Ai-guo

In table 1, the first pattern NN NR is a high occurrence pattern, it occurs 105 times in the text 301 to text 325. A normal rule writer, not a skilled one, can easily construct it. According to Chinese Penn Treebank, an NR is a name of a particular person, politically or geographically defined location (cities, countries, rivers, mountains, etc.), or organization (corporate, governmental, or other organizational entity). NT is the tag of temporal noun, which can be the object of prepositions such as 在 (at), 从 (since), 到 (until), or 等到 (until). All other nouns are tagged with NN.

We tried the procedure using just one term of "title used before person names" at a time. In the first case we tried the term "选手", and with this we found the new patterns (2) – (4). We tried pattern (1) and retrieved a list of new terms. With the new term "记者(reporter)" a new pattern - pattern (5) is acquired.

We generalized new patterns manually. For example the sentence (6) and (7) can be transform to lexico – syntactic patterns (8) and (9).

Table 2. Sample sentences and lexico-syntactic patterns

我国选手兰卫和陈晟已取得复赛权。Chinese players Lan Wei and Chen Hao have gained the…

选手刘小光和常昊均在中盘取胜。Player Liu Xiao-guang and Chang Hao both wined in the middle of the competes.

我国 (Chinese)NN NR 和 (and)NR…

NN NR 和(and) NR 均(both)在中盘(in the middle of the compete) VV

Compare these two patterns, according to hypothesis 2.1, a candidate pattern NN NR 和 NR is generated. After being validated, it was used to extract new person names.

3.3 Error analysis

There are a few words extracted are not person names. There are two kinds of mistakes.

The first word of the subtitle is the name of the news agency, which is tagged with <NR>. When the last word in the title is tagged with <NN>, the news agency is recognized as a PN. For example, the title of text 324 is

世界游泳锦标赛兰卫、陈晟获男子 1 米跳板复赛资格<NN>
(The world swimming competition, Lan wei and Chen Hao enter the second competition.)

The subtitle is

新华社<NR>罗马 9 月 1 日电杨爱国黄昌瑞 (Xin Hua new agency, Rome, September 1st, Yang Ai-guo, Huang Chang-rui)

新华社(Xin Hua new agency) is extracted as a PN. It is easy to correct this kind of mistake.

A few location names have same lexico – syntactic patterns with the person names'. For example, in text 317 the is a sentence:

其中 <NN>包括 <VV>下 <DT>一 <CD>届 <M>远南 <NR>运动会 <NN>东道主 <NN>泰国 <NR>曼谷 <NR>考察团 <NN>,<PU> (It includes Bangkok, Thailand, which is the amphitryon of next Yuan Nan athletic meeting.)

Here 泰国(Thailand) and 曼谷(Bangkok) are extracted as PNs. We used a dynamic location name dictionary to remove them.

3.4 Statistical results

Chart 1. Person names extracted by original observed pattern and by with new patterns

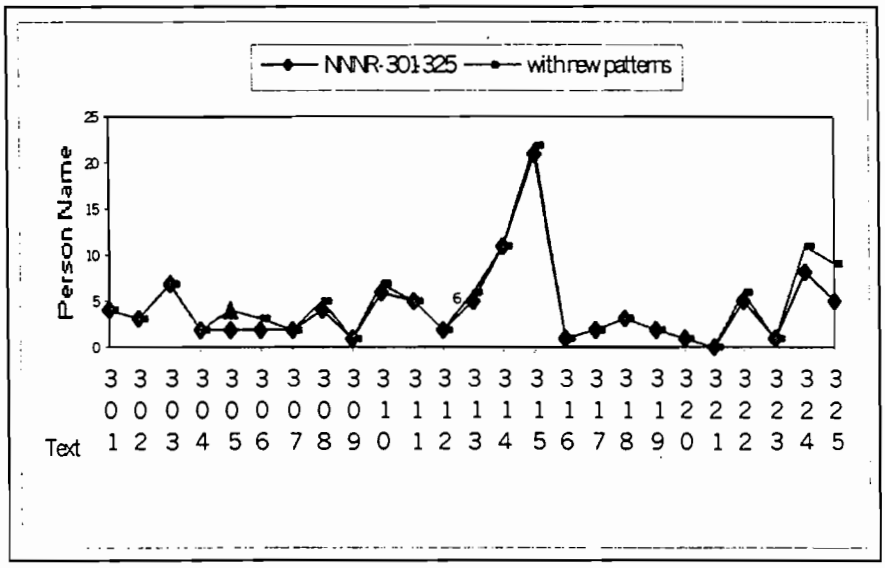


Chart 1 compares the number of person names extracted by pattern 1 and the number of person names extracted by all the four patterns.

From the text 301 to text 305 we totally have 105 sentences that contain these patterns. Totally there are 120 person names. We use the pattern NN NR 105 times. The pattern NN NR 和 NR occurs four times. Pattern NN NR NR 7 times, pattern NNNR`NR 4 times. The frequencies of each pattern are described in above chart. By using the new patterns the number of person names extracted from the Chinese Treebank increased about 14.3%.

$$\text{The increase of PNs extracted with new patterns} = \frac{\text{The number of PNs extracted with new patterns} * 100\%}{\text{The number of PNs extracted with NN NR}} \quad (3)$$

It is interesting that the frequencies of the new patterns are low. These patterns are easy to be ignored when a rule writer look through a large volume of text to find the patterns. Chart 2 is a pattern – frequency curve for these four patterns.

With all the patterns we extracted for the four relations we get a smooth pattern – frequency curve for Chinese person name recognition.

The extracted person names and title terms will be used to thesaurus construction in my nearly future research.

4 Conclusion and Future work

Supervised approach can be used for Chinese information extraction task. We present a bootstrapping algorithm that extracts patterns and generates semantic lexicons simultaneously.

We can see that accurate recognition and categorization of names in unrestricted text is a difficult task. Certain types of NEs, such as person names, not only name a person but can also serve as a source of many other kinds of names. We will try some statistical methods, such as Hidden Markov Models, to recognize person names by analysing the probability in which a word can be a person name in a certain situation.

The Chinese treebank is not so large as we wish. Since Chinese annotated corpus is very rare, our next work includes looking for larger corpus or tries unsupervised training approach or combined both methods.

Acknowledgement:

My work is supported by project COLLATE in DFKI (German Artificial Intelligent Center) and Computational Linguistic Department and project "Research on Information Extraction and Template Generation based Multilingual Information Retrieval", funded by the National Natural Science Foundation of China

5 References:

[1] Dimitrios Kokkinakis, Martin Gellerstam, Yvonne Cederholm, Torgny Rasmark. Sparkdata Presentation / Discussion Paper for the Fefor NE Recognition Workshop, January 2001, http://spraakdata.gu.se/nn/fefor.html

[2] Fei Xia, The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0) October 17, 2000.

[3] Borthwick, A Maximum Entropy Approach to Named Entity Recognition, Ph.D. (1999) New York University. Department of Computer Science, Courant Institute.

[4] Finkelstein-Landau, Michal and Morin, Emmanuel (1999), "Extracting Semantic Relationships between Terms: Supervised vs. Unsupervised Methods", In proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure, Dagstuhl Castle, Germany, May 99, pp. 71-80.

[5] Emmanual Morin, Christian Jacquemin, Project Corpus-Based Semantic Links on a Thesaurus, (ACL99), Pages 389-390, University of Maryland. June 20-26, 1999

[6] Marti A. Hearst, Automated Discovery of WordNet Relations, in WordNet: An Electronic Lexical Database, Christiane Fellbaum (ed.), and MIT Press, 1998.

[7] Marti A. Hearst, 1992. Automatic acquisition of hyponyms from large text corpora. In COLING' 92, pages 539-545, Nantes.

[8] Kaiyin Liu, Chinese Text Segmentation and Part of Speech Tagging, Chinese Business Publishing company, 2000

[9] Douglas Appelt: Introduction to Information Extraction Technology

[10] http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html