Pattern Discovery in Named Organization Corpus

Hsin-Hsi Chen and Yi-Lin Chu

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
E-mail: hh_chen@csie.ntu.edu.tw

Abstract

This paper presents how to mine formulation rules from a named organization corpus. The TEIRESIAS algorithm, which is widely used in bioinformatics domain, is adopted. The experimental results based on MET2 test bed show that the approach of regarding the morpheme of a keyword as a cluster is the best, the approach of regarding all the keywords as the same cluster is the next, and the approach of regarding each keyword as a cluster is the worse. The performance using morpheme-based approach is a little better than that of hand-crafted approach. The methodology can be easily extended to other types of named entities.

1. Introduction

Named organizations denote enterprises, companies, schools, hospitals, institutes, government offices, etc. This special kind of named entities is an open set. New organizations are set up continuously, and old organizations may be renamed or even dismissed. A lexicon cannot capture all the organization names, and that becomes an important research issue in human language processing (Chen and Lee, 1996). Because organization names are usually composed of more than two content words, the boundary identification is challenging.

Compared with the gazetteer approach, the rule-based approach employs the formulation rules to recognize named organization. The formulation rules may be hand-crafted or learned automatically (Chen, Yang, and Lin, 2003). The major drawback of the hand-crafted approach is the coverage problem. This paper adopts TEIRESIAS algorithm (Rigoutsos and Floratos, 1998) used in pattern discovery in biological sequence to mine the formulation rules from a named organization corpus.

This paper is organized as follows. Section 2 shows hand-crafted rules to recognize organization names. Section 3 introduces the TEIRESIAS algorithm and shows how to extract the rules. Section 4 employs MET2 (1998) data to evaluate the performance of pattern discovery. Section 5 concludes the remarks.

2. Hand-Crafted Rules

The structure of organization names is more complex than that of person names. Basically, a complete organization name can be divided into two parts, i.e., name and keyword. The following specifies the rules we adopted to formulate its structure (Chen, Ding, Tsai, and Bian, 1998). D is any content words.

OrganizationName →

(1) OrganizationName   OrganizationNameKeyword
   e.g., 聯合國   部隊
       lian he guo   bu dui
       United Nations   Force

(2) CountryName   OrganizationNameKeyword
   e.g., 美國   大使館

       mei guo   da shi guan
       United States   Embassy

(3) PersonName   OrganizationNameKeyword
   e.g., 羅慧夫   基金會
       luo hui fu   ji jin hui
           Foundation

(4) CountryName {D|DD} OrganizationNameKeyword
   e.g., 中國   國際   廣播電台
       zhong guo guo ji guang bo dian tai
       China   International Broadcasting Stations

(5) PersonName {D|D}   OrganizationNameKeyword
   e.g., 羅慧夫 文教   基金會
       luo hui fu wen jiao ji jin hui
           Culture and Education Foundation

(6) LocationName {D|D} OrganizationNameKeyword
   e.g., 台北   國際   廣播電台
       tai bei   guo ji guang bo dian tai
       Taipei   International Broadcasting Stations

(7) CountryName   OrganizationName
   e.g., 美國   國防部
       mei guo   guo fang bu
       United States Department of Defense

(8) LocationName   OrgnizationName
   e.g., 伊利諾州   州府
       yi li nuo zhou zhou fu
       Illinois   State Government

In this version, we collect 776 organization names and 1,059 organization name keywords.

In MET2 (1998) evaluation, the system (Chen, Ding, Tsai, and Bian, 1998) achieved the recall rate 77.72%, the precision rate 85.17% and the F-measure 80.78% for the extraction of organization names using the above rules.

3. Pattern Discovery with TEIRESIAS

TEIRESIAS Algorithm (Rigoutsos and Floratos, 1998) developed by IBM bioinformatics group is available on the web site (http://cbcsrv.watson.ibm.com/Tspd.html) and has been employed to a large number of computational biology applications. The following shows how to use it in organization name extraction.

Here, each Chinese word is in both Han character and Pinyin.

Assume we have two organization names:

"台北金融中心"

("tai bei jin rong zhong xin", Taipei Financial Center) and

"台北商業中心"

("tai bei shang ye zhong xin", Taipei Business Center). TEIRESIAS Algorithm will derive a rule shown as follows.

"台北 * 中心"

("tai bei * zhong xin", Taipei * Center),

where * denotes a wild-card.

When an unknown string "台北新聞中心" ("tai bei xin wen zhong xin", Taipei News Center) is input, it will be recognized as a named organization according to this rule.

Three parameters, including the minimum number of non-wild-card literals, the maximum extent spanned by any consecutive non-wild-card literals, and the desired minimum support, are considered in TEIRESIAS Algorithm. The pattern discovery procedure for organization names is proposed as follows.

(1) Collect keywords

An organization is usually composed of name part and keyword part. The variation of name part is quite large. It may be a combination of a person name, an organization name, or common content words. Comparatively, the keyword part is fixed. We input training organization names into TEIRESIAS Algorithm, and set the number of pattern support to be 5. In other words, the resulting rule must support at least 5 organization names. The rules that ends with wild-card, i.e., not a keyword, are removed. The last strings in the remaining rules are considered as keywords. In the following examples, "公司" ("gong si", company), "陣線" ("zhen xian", front) and "協會" ("xie hui", association) are regarded as keywords

(* * 公司) (* * Company),
(* * 陣線) (* * Front), and
(* 協會) (* Association).

(2) Cluster organization names

The organization names of different keywords may have different formulation rules. Besides, the occurrences of some types of organization names may be only a few in a training corpus. Grouping increases the possibility to discover a pattern. We cluster organization names with the following three alternatives.

(a) Each keyword denotes a cluster: the finest

(b) All the keywords denote a cluster: the coarsest

(c) The morpheme of a keyword denotes a cluster: in-between. Table 1 shows some examples.

(3) Assign features

A word in name part is assigned a feature. The feature set includes person, location, date, country, number, direction, and a word itself.

(4) Employ TEIRESIAS Algorithm

TEIRESIAS Algorithm extracts formulation rules for named organizations under the parameter settings.

Table 1. Morpheme as a Cluster

Morpheme	Examples
刊	月刊 (yue kan, monthly), 週刊 (zhou kan, weekly), 季刊 (ji kan, quarterly), 雙週刊 (shuang zhou kan, biweekly)
台	天文台 (tian wen tai, astronomical observatory), 氣象台 (qi xiang tai, meteorological observatory), 無線電台 (wu xian dian tai, radio station), 電視台 (dian shi tai, TV station)
行	花旗銀行 (hua qi yin xing, Citibank), 銀行 (yin xing, bank), 洋行 (yang xing, foreign firm), 商業銀行 (shang ye yin xing, commercial bank)

4. Experiments and Discussion

A corpus of 13,665 organization names is used for training. Some of rules mined are listed below.

(a) Each keyword denotes a cluster

(1) 第 Number Person 大學
(di Number Person da xue,
Ordinal Number Person University)

(2) 國家 * 聯合黨
(guo jia * lian he dang,
National * United Party)

(3) 全 Location * * 中心
(kuan Location * * zhong xin,
Whole Location * * Center)

(b) All the keywords denote a cluster

(1) Country 國內 航空 OrgKeyword
(Country guo nei hangkong OrgKeyword,
Country domestic airline OrgKeyword)

(2) 全 Country * 安全 OrgKeyword
(quan Country * an quan OrgKeyword,
Whole Country * Security OrgKeyword)

(3) Location Location * * OrgKeyword

(c) The morpheme of a keyword denotes a cluster

(1) Location Person 司
(Location Person si,
Location Person Department)

(2) 世界 * 同志 會
(shi jie * tong hi hui,
International * Comrade Association)

(3) 駐 * * 經濟文化 辦事處
(zhu * * jing ji wen hua ban shi chu,
foreign * * economic and culture office)

MET2 (1998) test bed is used to evaluate the quality of rules mined in Section 2. There are 377 organization names in the test data. The following summarized the experimental results for the three alternatives mentioned at step (2).

(a) Each keyword denotes a cluster
Total number of rules mined: 500
Recall: 62%, Precision: 86%,
F-measure: 72.05%

(b) All the keywords denote a cluster
Total number of rules mined: 700

Recall: 78%, Precision: 82%,
F-measure: 79.95%
(c) The morpheme of a keyword denotes a cluster
Total number of rules mined: 372
Recall: 81%, Precision: 82%,
F-measure: 81.50%

The finest approach has the best precision, but the worst recall. If a keyword of a test data is not seen in the training corpus, this approach fails to recognize this string. The rarest approach can capture all the keywords, so that the recall increases at expense of precision. The in-between approach has the best performance. We postulate that the keywords with the same morpheme have the same formulation rules. In this way, we employ the corresponding morpheme rule when the keyword is not seen before.

The patterns discovered by the second approaches are similar to the hand-crafted rules. All the rules except

OrganizationName →
CountryName   OrganizationName |
LocationName   OrgnizationName

are learned by our algorithm. Although the patterns mined by third approach are not exactly the same as the hand-crafted rules, they cover most organization names and recognize more organization names that hand-crafted rules could not. Some of them are shown as follows. The patterns matched are enclosed by parentheses.

(1) 國際新聞中心   [國際   *   心]
    guo ji xin wen zhong xin [guo ji   *   xin]
    International News Center [International   *   Center]
(2) 新西蘭賽航空公司   [新 Person   司]
    xin xi lan sai hang kong gong si [xin Person si]
(3) 西北航空公司   [Direction *   司]
    xi bei hang kong gong si   [Direction *   si]
    Northwest Airlines
(4) 香港寶華公司   [Location *   *   司]
    Hong Kong BOISE
    xiang gang bao hua gong si[Location *   *   si]
(5) 第四人民醫院   [第 Number *   院]
    di si ren min yi yuan   [di Number *   yuan]
(6) 國際通信衛星組織   [國際 *   織]
    guo ji tong xin wei xing zu zhi   [guo ji *   zhi]
    INTELSAT

Compared with performance using hand-crafted rules, i.e., 80.78% shown in Section 2, the third approach gains a little better performance, i.e., 81.50%. It shows that the automatic learning approach competes with the hand-crafted approach.

5. Concluding Remarks

The TEIRESIAS Algorithm, which is widely applied to biological domain, is used to extract the formulation rules of organization names. This methodology can be extended easily to other types of keyword-based named entities. The mined rules for organization names have F-measure 81.50% in MET2 test set, where documents are selected from newspapers in China. In contrast, the named organization corpus which is used for training is developed in Taiwan. The formulation of organization names is something different between China and Taiwan. In the future, we will try to find a training corpus coming from the same area with test set, and investigate the effects on recognition. Besides, the incomplete organization names still cannot be recognized by using the mined rules. Contextual information, which is indispensable for boundary identification, should also be investigated further.

References

Chen, Hsin-Hsi; Ding, Yung-Wei; Tsai, Shih-Chung and Bian, Guo-Wei (1998). "Description of the NTU System Used for MET2." *Proceedings of 7th Message Understanding Conference*, Fairfax, VA, 29 April - 1 May, 1998, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html.

Chen, Hsin-Hsi and Lee, Jen-Chang (1996) "Identification and Classification of Proper Nouns in Chinese Texts." *Proceedings of 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, August 5-9, 1996, 222-229.

Chen, Hsin-Hsi; Yang, Changhua and Lin, Ying (2003) "Learning Formulation and Transformation Rules for Multilingual Named Entities." *Proceedings of ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models*, July 12, Sapporo, Japan, 2003, 1-8.

MET2 (1998) *Proceedings of 7th Message Understanding Conference*, 1998, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html.

Rigoutsos, I. and Floratos, A. (1998) "Combinatorial Pattern Discovery in Biological Sequences: the TEIRESIAS Algorithm." *Bioinformatics,* 14(1), January 1998.