Finding Names in Chinese Text using a Hybrid Rule Induction Model*

Jimin Liu, Jing Xiao and Tat-Seng Chua
School of Computing, National University of Singapore
E-mail: {liujm, xiaojing, chuats} @comp.nus.edu.sg

Abstract: This paper presents a hybrid rule induction approach to extract named-entities (NE) from Chinese text. The method employs two main strategies. First, it adopts a greedy approach to extract all possible combinations of words and names in order to overcome the word segmentation problem. Second, it employs the generalized induced rules, supplemented by the default-exception trees and decision tree, to resolve the ambiguities in the extracted names. The method has been tested on the MET2 test set and it is able to achieve an overall F_1 measure of over 91%.

Keywords: Rule induction learning, lexical chaining, decision tree, Chinese named-entity extraction

1. Introduction

The task of named entity (NE) extraction is to identify all meaningful entities in a document and to classify them into the categories of person, organization, place, time and date. NE recognition is one of the most important tasks in information extraction, and a lot of works has been done to tackle this problem. Earlier approaches used mostly handcrafted heuristic rules [11], while recent methods focused on machine-learning approaches [1,6].

Most reported methods tackled the NE extraction problem in English, in which recent methods have reported a high accuracy of over 95% [8]. However, there is less corresponding research in Chinese. Moreover there are several fundamental problems that are either unique to Chinese or are not critical in English, such as word segmentation. Furthermore, there are few openly available language resources that can be used to build and evaluate Chinese language systems. Examples of the available resources include: PKU-Corpus [15], Hownet [5], Chinese Treebank [12], and MET2 [8]. These resources, however, are relatively small and are in less widespread use as compared to the English counterparts.

There are several recent works on Chinese named entity recognition [2, 7, 14]. These approaches used mostly handcrafted rules, supplemented by word or character frequency statistics. These methods also require a lot of resources to model the names.

This paper proposes a template-based rule induction model that incorporates machine learning techniques to tackle the Chinese NE extraction problems. (a) In order to address the uncertain word segmentation problem, it adopts a greedy approach that extracts all possible combinations of word tokens and names. (b) In order to overcome the problems of flexibility in language usage, it adopts the default exception (DE) trees to model all exceptions in the usage of key words in inducing names. (c) It employs a combination of generalized rules together with decision tree to resolve the ambiguities in the extracted names. (d) In order to tackle the problem of sparse training resources, it develops a quasi probability model based on bigram statistics obtained from both the tagged corpus and known name listed to extract P-Names. This paper describes the overall system design and

* This research is supported by A*STAR and the Ministry of Education of Singapore under the research grant RP3989903.

implementation.

2. The Strategy for Chinese NE Extraction

In Chinese, different types of names have different structures and requires different context and cue-words (CWs) for their recognition. Examples of cue-words for names include: surnames such as "黄" (Huang) or titles such as "经理" (Manager) that induce a person name; place suffices such as "山" (mountain) that introduce a mountain name; and organization suffices such as "公司" (company) that induce an organization name, etc.

While different lists of CWs can be extracted to help induce different types of names, there are many ambiguous cases in which the presence of a CW does not induce a name. Examples of CW ambiguity include "该公司" (that company) that does not introduce a company name, or "爬山" (climb mountain) that does not introduce a mountain name, etc. Also a proper name in Chinese may consist of common words. For example, for the string "黄金富有一本书" (Huang Jin Fu has a book), if we treat "黄" as a surname, then it is possible that "黄金富" may be correctly induced as a name. However, it is equally possible that both "黄金" (gold) and "富有" (rich) are segmented as common words since they are both high frequency words. Thus a purely probabilistic approach would either miss or wrongly segment the names.

To overcome these problems, we adopt a hybrid machine learning strategy to extract the NEs in the text as follows.

a) The first essential step is to perform word segmentation, including phonetically translated name (p-name) extraction, using a greedy approach. We make extensive use of corpus statistics, word dictionary, and known name lists to perform the preliminary segmentation by: (i) extracting the numbers, dates and times using reliable techniques; (ii) segmenting words by using a simple dictionary-based forward longest matching technique; (iii) identifying all possible P-Names using the method discussed in [13]. The result is a list of possibly overlapping words and P-Names, and positions of all CWs that could indicate the presence of possible names and their types.

b) For each CW for a specific name type, we employ the DE (default-exception) tree to identify all possible exceptions on the use of this particular CW to induce a name. The DE trees derived from the training corpus are then used to filter out the instances where the presence of CWs does not induce a name. As a result of this step, some segmented words may be removed, re-aligned or new word segments introduced.

c) At each CW and P-Name position, we employ the appropriate name and context pattern rules to extract all possible names. The details of how these rules may be found will be discussed in Section 4. The result is a list of possibly overlapping or conflicting names.

d) In case of ambiguities in the list of names recognized, we employ the decision tree to find the best possible name in that specific context.

e) The approach is targeting at finding names that appear for the first time in which they are expressed in full with the necessary context. Subsequent occurrences of the same name may have less context or are abbreviated. Thus, we employ sub-string matching to locate subsequent occurrences of the extracted names in all likely name positions. In addition, we utilize heuristic rules to extract names that appear in constructs such as the enumerated name list.

The following Sections describe only the induction of generalized pattern rules and the use of decision tree to resolve ambiguity in the extracted names. Other related issues can be found in [3,4].

3. The Training Language Resources

In order to support the above NE extraction process, we use the PKU-Corpus [15] as the basic language training resource. The corpus contains one-month of news report from China's People Daily. The corpus is manually segmented into words with appropriate POS tags. There are 37,419 sentences containing 1,121,787 words.

Using the PKU-Corpus as the base, we derived the following language resources: (a) We build a common word dictionary by removing all words that are tagged as number, time, and name. The resulting common word dictionary contains 37,025 words. (b) We use the PKU corpus, supplemented by MET2 training set, to extract a list of CWs for different types of names. (c) We obtain a list of Chinese place names from MET2 training set. (d) Finally, we collected about 8,000 organization names, and 100,000 P-Names from the web by using a bootstrapping approach [13].

4. The Induction and Generalization of Pattern Rules

Given a training corpus with proper tagging of names, we employ the rule induction method to extract all instances of names together with their context. Because of the uncertainty in word segmentation, we do not utilize the syntactic and semantic constructs to guide the NE extraction process. Instead, we make heavy use of context of names to derive the generalized pattern rules for both the name.

Each pattern rule is represented as: $<p_1, \ldots, p_i><p_{i+1}, \ldots, p_j>$ $<p_{j+1}, \ldots, p_k>$, where p_x (x=1,2,..k) denotes an atomic pattern such as a word, word set, number, data, time, location and possible name. $<p_{i+1}, \ldots, p_j>$ constitutes the name; and $<p_1, \ldots, p_i>$ and $<p_{j+1}, \ldots, p_k>$ respectively represents the left and right context pattern list.

By applying the above pattern rules to each instance of the name in the training corpus, we can derive a large number of rules. Although it is easy to extract all instances of names from the training corpus, the key, however, is how to generalize them to handle the more general cases.

Here, we adopt a lexical chaining approach to generate semantic word groups [3], and use these word groups to generalize the rules. Our lexical chaining method utilizes HowNet as the basic linguistic resource.

For example, given an instance of an organization name "山东大成农药股份有限公司", it is first segmented into [山东][大] [成][农药][股份有限公司]. Since "山东" is a place name, and "股份有限公司" can be easily detected as a suffix of an organization name, the instance can be generalized to <place> [大][成] [农药]<股份有限公司>. Next, through lexical chaining, we can generalize "农药" to <chemical>, and thus the instance is deduced to <place> [大][成] <material> <股份有限公司>. Finally, we can replace [大][成] by <kernel-name> by applying some heuristic rules, and the instance is generalized to <place><kernel-name><material> <股份有限公司>.

5. Disambiguation of Extracted Names using Decision Tree

As the induced rules are designed to capture all possible names, there will be many instances of overlapping possible names being detected. Hence there is a need to select the best rule or the best possible name in a given context. In other word, given two conflicting rules $R^{(i)}$ and $R^{(j)}$, we should be able to resolve which one is better.

Here we employ the decision tree to resolve multiple ambiguous names. In order to learn such a decision tree, we compare the support and word length of both rules by computing a difference vector D_{ij} containing 6 discrete features:

$$D_{ij}(R^{(i)},R^{(j)}) = \{f_1, f_2, f_3, f_4, f_5, f_6\} \quad (1)$$

where f_1 measures the difference between the length (the number of atoms) of the name part of rules $R^{(i)}$ and $R^{(j)}$; f_2 measures the difference between the length of the whole rule pattern; f_3 and f_4 respectively compute the differences in

length of their left and right context; f_5 measures their relative name occurrence frequency; and f_6 gives the relative support of the CW in both rules.

From the training corpus, we identify all positions with conflicting names. At each position, if there are u+v rules that are applicable, in which, u rules $\{R^{(p1)},...R^{(pu)}\}$ give the correct name, and v rules $\{R^{(n1)},...R^{(nv)}\}$ give the wrong name. We generate u*v positive training examples by using the name differences $D_{(pi)(nj)}(R^{(pi)},R^{(nj)})$, for i=1,..u, j=1..v. Similarly, we can generate u*v negative training examples by using the differences $-D_{(pi)(nj)}((R^{(pi)},R^{(nj)})$.

After we have setup the D_{ij} values for all conflicting cases, we employ C5.0 algorithm to learn the decision tree.

6. Experimental Results and Discussions

One serious problem in Chinese NLP is the lack of openly available datasets, making it difficult to evaluate and compare systems. Bearing this in mind, we use only openly available datasets for our training and testing. Here, we use a combination of PKU-corpus, Hownet, MET2 Chinese resources, and two name lists (for foreign and organization names) collected from the web by a bootstrapping approach. We use these resources to build the necessary dictionaries as described in Section 3, and use the PKU corpus to derive the generalized pattern rules. In order to compare the performance of our system with other reported works, we test our system using the MET2 test set.

Table 1 tabulates the results of our system in terms of recall (Rc), precision (Pr) and F_1 measures on the MET2 test set. For comparison purpose, we also include the corresponding results reported in [2, 14] for the formal MUC-7 tests. The detailed analysis of our results can be found in [3].

Table 1 shows that our system could achieve a high performance of over 91% in F_1 values for all the name types. Its performance is significantly better than both the reported systems, especially for the person and organization names. The results demonstrate that our rule induction learning model that is tolerant to word segmentation and name extraction errors is very effective. In particular, we found that the use of DE-trees and decision tree to resolve multiple ambiguous names significantly improved the performance of the system.

The main errors in our system come from five sources. (a) We miss out many Japanese names as our system is not tuned to recognizing Japanese names, which is neither a Chinese name nor a P-Name. (b) We miss some person names because of the unexpected format (like an instance with a blank between the surname and the first name) and missing surnames in the cue-word list. (c) Some place and organization names are wrongly tagged because of the inconsistency in the manual tagging of these two entities within and between the training and test sets. Examples are the names with suffix "中心", which are mostly tagged as organization name in the PKU-corpus, and mostly as place name in the MET2 corpus. (d) Some common nouns, such as 月球 (moon), 太阳 (sun) and 土星 (Jupiter), are tagged as names in the test corpus. (e) There are missing concepts in Hownet.

Table 1:Results of MET2 under different configurations

	Name Type	N_C	N_P	N_W	N_M	N_S	Rc	Pr	F_1
Our Chinese NE Finder	**Org[n]**	**347**		**2**	**14**		**92**	**91**	**91.5**
	Person	**171**		**1**	**0**		**98**	**91**	**94.4**
	Place	**691**	**0**	**17**	**42**	**61**	**92**	**90**	**91.0**
Results of Chen et al. 98 [2]	Org[n]	293	0	7	77	44	78	85	81.3
	Person	159		0	0		91	74	81.6
	Place	583	0	65	102	194	78	69	73.2
Results of Yu et al. 98 [14]	Org[n]	331		0	14		88	89	88.5
	Person	160		0	7		92	66	76.7
	Place	682		0	1		91	89	90.0

7. Conclusion

Chinese NE is a difficult problem because of the uncertainty in word segmentation and ambiguities in NE identification. Many existing techniques that require knowledge of word segmentation, and syntactic or semantic tagging

of text cannot be applied. In this paper, we propose a template-based rule induction model to tackle this problem. The main contributions of our approach are two-fold. First we induce rules for names and their context, and generalize these rules using lexical chaining. Second, we adopt a greedy approach in generating multiple overlapping word tokens and possible names, and employ a combination of generalized induction rules, DE-trees and decision tree to resolve the ambiguities. We tested the system on the MET2 test set and the results have been found to be superior to all reported systems.

We plan to further test our system on a large-scale test corpus. We will refine our techniques on a wide variety of text corpuses, and in applying the bootstrapping techniques to overcome the problem of data sparseness. Finally, we will extend our work to perform relation and information extraction in multilingual text.

References:

[1] Bikel, D.M., Schwartz, R. and Weischedel, R.M. 1999. An Algorithm that Learns What's in a Name. Machine Learning 34(1-3), 211-231

[2] Chen, H.H., Ding, Y.W., Tsai, S.C. and Bian, G.W. 1998. Description of the NTU System used for MET-2. *Proc. of the 7th Message Understanding Conference.*

[3] Chua, T.S. and Liu, J.M. 2002. Learning Pattern Rules for Chinese Named Entity Extraction. In Proc. of AAAI'02.

[4] Chua, T.S. and Liu, J.M. 2002. A Hybrid Rule Induction Approach for Chinese Named Entity Extraction. Submitted for publication.

[5] Dong, Z.D. and Dong, Q. 2000. HowNet, available at: http://www.keenage.com/zhiwang/e_zhiwang.html.

[6] Isozaki, H. 2001. Japanese Named Entity Recognition Based on a Simple Rule Generator and Decision Tree Learning, Proc. of Association for Computational Linguistics, 306-313.

[7] Luo, Z.-Y. and Song, R. 2001. An Integrated and Fast Approach to Chinese Proper Name Recognition in Chinese Word Segmentation, Proc. of the Int'l Chinese Computing Conference, Singapore. 323-328.

[8] Marsh, E. & Perzanowski, D. 1998. MUC-7 Evaluation of IE Technology: Overview of Results, *Proc. of 7th MUC.* At: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

[9] Riloff, E. 1996. Automatically Generating Extraction Patterns from Untagged Text. Proc. of AAAI'96, 1044-1049

[10] Soderland, S., Fisher, D. Aseltine, J. and Lehnert, W. 1995. Crystal: Inducing a Concept Dictionary, IJCAI'95

[11] Weischedel, R. 1995. BBN: Description of the PLUM System as Used for MUC-6. Proc. of the 6th Message Understanding Conference, 55-69.

[12] Xia, F., Palmer, M., Xue N.W., Okurowski, M.E., Kovarik, J., Chiou, F.D., Huang, S-Z., Kroch, T. and Marcus, M. 2000. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. *Proc of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, 2000.

[13] Xiao, J., Liu, J.M. and Chua, T.S. 2002. Extracting Pronunciation-Translated Named Entities from Chinese Text using Bootstrapping Approach. To appear in *"1st SIGHAN Workshop on Chinese Language Processing", COLING '2002.* Sep. Taipei, Taiwan.

[14] Yu, S.H., Bai, S.H. and Wu, P. 1998. Description of the Kent Ridge Digital Labs System Used For MUC-7, 1998, *Proc. of the 7th Message Understanding Conference*

[15] Yu, S.W. 1999. The Specification and Manual of Chinese Word Segmentation and Part of Speech Tagging. At: http://www.icl.pku.edu.cn/Introduction/corpustagging.htm

About the authors:

Jimin Liu received his PhD degree from Institute of Computing Technology, Chinese Academy of Science. He is currently pursuing research work at the National University of Singapore (NUS). His research interests include natural language processing, image processing and computer vision.

Jing Xiao is presently a PhD. Candidate in the School of Computing, NUS. Her research interests include information extraction and machine learning.

Tat-Seng Chua was the Founding Dean of the school of Computing, NUS, from 1998-2000. His research interests are in multimedia information retrieval and information extraction. He serves in the editorial boards of: IEEE Transaction of Multimedia, The Visual Computer, and Multimedia Tools and Application.