Learning Pattern Rules for Chinese Named Entity Extraction

Tat-Seng Chua and Jimin Liu

School of Computing,
National University of Singapore,
3 Science Drive 2, Singapore 117543
{chuats, liujm}@comp.nus.edu.sg

Abstract

Named entity (NE) extraction in Chinese is very difficult task because of the flexibility in the language structure and uncertainty in word segmentation. It is equivalent to relation and information extraction problems in English. This paper presents a hybrid rule induction approach to extract NEs in Chinese. The method induces rules and names and their context, and generalizes these rules using linguistic lexical chaining. In order to handle the ambiguities and other contextual problems peculiar to Chinese, we supplement the basic method with other approaches such as the default-exception tree and decision tree. We tested our method on the MET2 test set and the method has been found to out-perform all reported methods with an overall F_1 measure of over 91%.

1. Introduction

Named entity (NE) recognition is a task in which meaningful entities in a document are detected and classified into categories such as person, organization, location, time and date. NE recognition is one of the most important tasks in information extraction. Many approaches have been proposed to tackle this problem. Earlier approaches used mostly handcrafted heuristic rules (Appelt et al. 1995, Weischedel 1995), while recent methods focused on using machine learning approaches, such as the hidden Markov model (Cucerzan & Yarowsky 1999, Bikel et al. 1999), decision tree (Sekine et al. 1998) and maximum entropy (Borthwick 1999, Isozaki 2001) etc. Most of these methods utilize some form of syntactic and semantic knowledge of text in detecting NEs.

NE extraction is not a serious problem in English, in which recent methods have reported high accuracy of over 97% (Marsh & Perzanowski 1998). However, there are four fundamental problems that are either unique to Chinese or are not serious in English.

One of the most serious problems is the word segmentation problem (Sproat et al. 1996, Palmer 1997, Yu et al 1998). In Chinese, there is no implicit blank between words, where a word is defined as consisting of one or more characters representing a linguistic token. Word is a vague concept in Chinese, and linguists often do not have generally accepted guidelines on what constitutes a word. Without well-defined word boundaries, most syntactical and semantic analysis methods that assume words as basic entities cannot be performed effectively.

Second, there are three main types of names – the person, place and organization names. Each type of names has its own structure and cue-words. Although the presence of specific context or cue-words is needed to induce a name, there are many ambiguous cases that this is not true, especially when word segmentation is also uncertain. Also a proper name in Chinese may contain commonly used names, which further complicates NE recognition using probabilistic approaches.

Third, there are two types of names in Chinese. The first is called the S-Name, which is constructed according to Chinese semantics. The second is termed P-Name, which is a translated foreign name based on its phonetics such as the name"贝克利" (Ber-ke-ley). Each type of name includes all categories of proper names. The rules for the construction of S-Names and P-Names are different and need to be handled differently.

Fourth, there are few openly available language resources that can be used to build and evaluate Chinese language systems. Examples of such resources include: Chinese Treebank (Xia et al. 2000), MET2 (Marsh & Perzanowski 1998), PKU-Corpus (Yu 1999) and Hownet (Dong & Dong 2000), These resources are relatively small and are in less widespread use as compared to those in English. Thus one major issue is how to make the best use of existing limited resources with minimum labor cost.

Because of these problems, the usual techniques that require proper word segmentation and utilize syntactic and semantic knowledge of words cannot be applied effectively to Chinese. Also, because of the flexibility in the language in which common words often appear as part of the names, the method developed must be able to handle ambiguities and exceptions well. In fact, the problem of NE recognition should not be isolated from

the word segmentation and part-of-speech tagging (Yu et al. 1998).

In many aspects, the problems of Chinese NE recognition are similar to that of relation and information extraction (IE) in English, which is a much harder problem than English NE extraction. In English, the template-based rule induction approach has been popular in extracting structured data from the input documents such as the newswire articles or web pages (Muslea 1999). For free text, most such methods use syntactic and semantic constructs as the basis to identify and extract items of interests (Riloff 1993, Soderland et al. 1995). For on-line and web-based documents, the current methods rely on a combination of domain knowledge and known structure of documents. All these techniques cannot be applied directly to Chinese free-text as explained earlier.

In this paper, we propose a template-based rule induction approach that incorporates other machine learning techniques. The method first adopts a greedy method to extract word tokens and possible names that may be overlapping. It then uses the generalized rules induced together with other techniques, such as the DE-(default exception) trees and decision tree, to resolve the ambiguities in the extracted names. The rules are induced from the training corpus and generalized using linguistic techniques such as the semantic clustering. Our initial results demonstrate that the approach is very effective in Chinese NE extraction.

The rest of the paper discusses the use of template for Chinese NE extraction, describes the details of the induction and generalization of rules, and discusses the results of applying the system to a standard test corpus.

2. The Structures and Problems in Chinese NE Extraction

Different types of names have different structures and require different context and cue words (CWs) for their recognition. The simplest type of name is the Chinese person S-name, which has a very regular structure of:

S-Person-Name :== <surname> <first name> (1)

Most surnames (several hundreds) have only one character, with only very few (about 8) having two characters. The first name is usually composed of one or two words, and the number of characters in each word is less than or equal to 2. The surnames come from a small list of names such as "张" (Zhang), "刘" (Liu), and can be used as CWs to induce a name. We can also use the person titles such as "主席" (chair-person), "经理" (manager), as the context to confirm the presence and boundary of person names.

A typical full Chinese place name is more complicated and takes the form of:

Place-Name :== $<p_1><p_2>\ldots.<p_n>$ (2)

where $n \geq 1$; and p_i :== <pl-name $_i$>{<pl-suffix $_i$>}. Each p_i is itself a place name, and from left to right, p_{i-1} represents a bigger place than p_i. This is in reverse order to that in English. Each name p_i has an optional <pl-suffix> such as the province (省), river (河) and road ("路") etc. <pl-suffix> provides the CW to trigger the extraction of a place name.

A typical full Chinese organization name is even more complicated and typically takes the form of:

Org-Name :== { [<Place-Name>]* [<person-name> | <org-name>] }$^{\times}$ [<org-generic-name>] <org-suffix> (3)

where []* means repeat zero or more times, and {}$^{\times}$ means select at least one of the enclosed items. An example of a full organization name is "山东大成农药股份有限公司", where [山东] is a place name [大成] is the organization name, [农药] is a generic name for drug which indicates that this is a drug company, and [股份有限公司] (limited company) is the organization suffix. These suffices can be used as CWs to induce the names.

To facilitate the detection of different types of names, respective lists of CWs can be extracted from the training corpus and from known lists. Although in most cases, the presence of context or CWs at the beginning or end of the string helps to induce a name, there are many cases that are ambiguous. 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. These problems are aggravated by the uncertainty in word segmentation.

Finally, a name can be an S-Name or a P-Name. As with S-Names, a P-Name may includes any character and in any combination, and can often be confused with other common names. Fortunately, the combination of characters is often unique to P-Names and can thus be recognized using a probabilistic approach. Here we used a quasi-probabilistic model to locate all possible P-Names. The method uses bigram statistics obtained from both the training corpus and the name list. The details of P-Name finder can be found in Xiao, Liu & Chua (2002).

3. Template-based Rule Induction Approach

The above discussions suggest that a Chinese NE extraction system must meet at least three major requirements. First, it must be able to work with uncertain word segmentation and has the ability to deal with different and conflicting word combinations. Second, it must be able to handle different types of names, often with radically different format, and

triggered by the presence of different CWs. Third it must consider the context of possible names in order to resolve ambiguities in NE extraction. Employing a purely probabilistic approach will not be adequate as the system will likely confuse a name from common words or vice versa, and thus either miss the name completely or segment the name wrongly. To tackle this problem, we adopt the template-based rule induction approach commonly employed in IE tasks in English (Muslea 1999). We combine rule induction with other machine learning techniques to tackle the ambiguity problems.

We employ the template similar to that used in Riloff (1996) for Chinese NE extraction. The template composes of four parts (see Figure 1): trigger, pattern list, constraint, and output. The trigger pattern <p> can be a CW or a P-Name. As CW for different types of names are different, it also provides indication on the type of possible names to be extracted.

Trigger:	<p>
Pattern Rules:	$\underline{R}_{name}$, $\underline{R}_{context}$
Constraint set:	$\underline{D}$
Output:	$<t_1, b_1, e_1>\ldots.<t_m, b_m, e_m>$

Figure 1: Template for NE extraction

At the triggering position, we fire the appropriate rules to induce the name and its context. As the key to NE extraction is the detection of the name and its boundaries, we utilize the CWs and context to derive generalized pattern rules for name (R_{name}) and context ($R_{context}$) as:

R_{name} := S-Person-Name | Place-Name | Organization-Name (4)

$R_{context}$:= [left context] [left CW] <possible name> [right CW] [right context] (5)

We use the appropriate name pattern rules to identify all possible names, and then employ the context pattern rules to narrow down the list of possible names with more precise boundaries. The application of rules must satisfy the corresponding constraints given in the set $\underline{D}$.

The output is a list of possibly overlapping names of the form $<t_i, b_i, e_i>$, where i=1, .. m, the number of names in the output list. Here t_i indicates the category (of type person, location or organization) of the i^{th} name that starts at atomic pattern p_{bi} and ends at p_{ei}. When m>1, we employ a decision tree to select the best possible names.

One main design consideration of our approach is that the system is able to tolerate uncertainty in word segmentation and possible name extractions at all stages of NE processing. Thus a greedy approach is employed at the initial stages to identify all possible names.

4. The induction of Generalized Pattern Rules

This Section discusses the induction of pattern rules for NE extraction. Given a training corpus with proper tagging of names and common words, we apply the rule templates given in Equations (4-5) to extract all instances of names together with their context. Although it is easy to extract all instances of rules from the corpus, the key, however, is how to generalize the context to handle more general cases.

An example of a specific name instance is the tagged name segment "[经理]<张实>[在][接见][记者][时][表示]" or translated as [Manager] < 张 实 > [when] [meeting] [the reporters] [indicates]). Although by applying template (5) to this example, we can extract a context pattern rule of the form: "[经理] <X>[在>][接见]" that will introduce a person name X. However, this rule is too specific, since its left and right contexts are words. To generalize the rules, we need to replace the words in the rules with more general patterns.

One possible approach to generalize the context is to use the syntactic and semantic tags of words (Riloff 1993, 1996, & Soderland et al. 1995). Unfortunately, for Chinese language, before the words are correctly segmented, it is difficult to obtain such tagging information. Comparatively, by using synsets to replace specific words seems to be a more feasible approach. In the example above, we can generalize the [接见] by its synset {[接见, 会见….]}. However, in some cases, simply considering the synset is still too specific.

Based on these observations, we adopt a lexical chaining approach to generate the semantic grouping for the context words, and use these semantic groups to replace specific words in the rule patterns to generalize rules.

4.1 Lexical Semantic Clustering

Lexical chaining is a commonly used approach to cluster semantically related words (Green 1999, Liu & Chua 2001). Here we utilize Hownet (Dong & Dong 2000) to find lexical chains. Hownet is an important lexical semantic dictionary in Chinese NLP and can be used as the counterpart to WordNet. Each word w_j in Hownet has a set of definitions, defining different "points of view" of this word. Examples of different "points" include the definitions of its basic meaning $D_B(w_j)$ and its parent $D_P(w_j)$ etc. In this research, we consider only these two "points" of view. That is, if $D_B(w_j) = w_k$, then w_k has the same meaning as w_j, and if $D_P(w_j) = w_P$, then w_P is the parent of w_j.

By treating words related to by D_B and D_P as semantically related, we develop an algorithm to

generate semantic groups in a specific domain via lexical chaining.

1) Initialize:
   a) $\underline{W_S} \leftarrow \{(w_1,f_1), (w_2,f_2), \ldots, (w_n,f_n)\}$,
      where f_i is the frequency of occurrences of word w_i in a given domain containing n unique words.
   b) Set the output group as empty: $\underline{G_{out}} \leftarrow \Phi$
2) Generate all possible semantic groups:
   a) For each word w_i in $\underline{W_S}$, use Hownet to find out its two selected "points" of view, ie $D_B(w_i)$ and $D_P(w_i)$.
   b) Generate all possible groupings of words as:
      $\underline{G_{all}} \leftarrow \{(\underline{G_1}, c_1), (\underline{G_2}, c_2), \ldots, (\underline{G_n}, c_n)\}$
      where $\underline{G_i}$ contains words that have the same meanings as that of w_i in the sense of $D_B(w_i)$ and $D_P(w_i)$, i.e.:
      $\underline{G_i} = \{w_k \mid D_B(w_k)=\{D_B(w_i), D_P(w_i)\}$ or $D_P(w_k)=\{D_B(w_i), D_P(w_i)\}\ \}$
      and c_i is the prominence measure of $\underline{G_i}$, which is simply the aggregate frequency of all words in $\underline{G_i}$:
      $$ci = \sum_{w_k \in \underline{G_i}} f_k$$
3) Select the prominent groups as the semantic groups:
   a) From $\underline{G_{all}}$, select the group with maximum c_i as:
      $$(\underline{G_{max}}, c_{max}) \leftarrow \arg\max_{\underline{G_i} \in \underline{G_{all}}} \{c_i\};$$
      Terminate the process if $c_{max} < \sigma$
   b) Move $\underline{G_{max}}$ to $\underline{G_{out}}$, ie
      $\underline{G_{out}} \leftarrow \underline{G_{out}} \cup (\underline{G_{max}}, c_{max})$, $\underline{G_{all}} \leftarrow \underline{G_{all}} - \underline{G_{max}}$
   c) For each remaining group $\underline{G_j}$ in $\underline{G_{all}}$, do the followings:
      $\forall \underline{G_j} \in \underline{G_{all}}$,
      set $\underline{G_j} \leftarrow \underline{G_j} - \underline{G_{max}}$, $c_j \leftarrow c_j - \sum_{w \in \underline{G_j} \cap \underline{G_{max}}} c$,
      and if $\underline{G_j} = \Phi$, then $\underline{G_{all}} \leftarrow \underline{G_{all}} - \underline{G_j}$
   d) Repeat the process from step (3a) if $\underline{G_{all}} \neq$ null.

At the end of applying the above lexical chaining algorithm, we obtain a set of semantic groups, each containing a cluster of related words. These semantic groups can be used as the basis to generalize both the name and context pattern rules.

4.2 The Generalization of Name Pattern Rules

The rules for person and place names are relatively simple and can be easily generalized by replacing the surnames in S-person-names by the general pattern <surname>, and the place suffices in the place names by a general pattern <pl-suffix>.

The structure of the organization names as defined in Equation (3) is more complex and there exists many possible rule combinations. We first generalize the organization suffices using CWs for organization, and names of person and location using Equations (1) and (2) respectively. For the remaining characters within the organization instances, we identify high frequency words as possible generic name of the organization. We then use lexical chaining to generalize these likely generic names into semantic group representations, which are used as generalized <org-generic-name>. Finally, we extract the set of generalized organization rule patterns from the training examples.

4.3 The Generalization of Context Pattern Rules

Given the training corpus, we extract all instances of names together with their context as:

$$\underline{RS} = \{R^{(r)} \mid R^{(r)}=<w_{-2}><w_{-1}><name><w_1><w_2>\} \quad (6)$$

where r = 1, .. number of name instances.

Here we consider the preceding and succeeding two words as the context to <name>. We consider only two-word context because we want to capture context information of the form: <verb>+<CW>, <CW>+<verb>, and <adverb>+<verb> etc. While one-word context is insufficient, more than two-word context would be too costly.

Given *n* different words $\{w_1,w_2,\ldots w_n\}$ appear in the context of the rules, together with their occurrence frequencies as the context words. We can apply the proposed lexical chaining algorithm to generate m semantic groups from these context words as: $\{(\underline{G_1},c_1)\ldots(\underline{G_m},c_m)\}$, where m<<n. We generalize each rule $R^{(k)}$ by replacing each context word w_c in $R^{(k)}$ by its corresponding semantic word group $\underline{G_c}$.

4.4. Selection of Best Rules using Decision Tree

When more than one pattern rule can be applied at a certain situation, it will result in multiple overlapping possible names being detected. Thus there is a need to select the best rule or the best possible name in a given context. What we want is to train a decision function, $f(\underline{RS} \otimes \underline{RS} \rightarrow \{-1,1\})$, so that given two conflicting rules $R^{(i)}$ and $R^{(j)}$, we are able to resolve which rule is better. To learn such a decision function, we need to compare the support and length of both rules by capturing the difference vector D_{ij} containing 6 discrete features:

$$D_{ij}((R^{(i)},R^{(j)}) = \{f_1, f_2, f_3, f_4, f_5, f_6\} \quad (7)$$

where f_1 and f_2 respectively measure the differences between the length of the name part and the whole rule

pattern between $R^{(i)}$ and $R^{(j)}$; 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 name resolutions. At each position, if there are u+v rules $\{R^{(p1)},\ldots R^{(pu)}, R^{(n1)},\ldots R^{(nv)}\}$ that are applicable, in which, u rules $R^{(p1)},\ldots R^{(pu)}$ give the correct names, and v rules $\{R^{(n1)},\ldots R^{(nv)}\}$ give the wrong names. 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 name differences -- $D_{(pi)(nj)}(R^{(pi)},R^{(nj)})$. After we have setup the D_{ij} values for all conflicting cases, we employ the C5.0 algorithm to learn the decision function f.

5. The Overall NE Recognition Process

Given an input document, the overall process of extracting NEs is as follows (see Figure 2).

a. The first essential step is to perform word segmentation, including p-name recognition. As with all other approaches (Yu et al. 98, Chen et al. 98), 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 the simple dictionary-based forward longest matching technique; (iii) identifying possible P-Names; (iv) locating typical S-Names, again, using dictionary look-up. The result is a list of possibly overlapping words and P-Names, and positions of all CWs that indicate the presence of possible names and their types.
b. We next employ the DE (Default exception)-trees to identify possible exceptions on the use of CWs to induce a name. The DE-tree is derived from the training corpus and it enumerates all exceptions found in the usage of a particular CW and its context. For example, while in most cases the pattern "*山" (or * Mountain) induces the name of a mountain, such as "黄山", "昆仑山", there are cases such as "千山万水" (meaning vast territory), "爬山" (meaning climb mountain) etc that the CW "山" does not introduce a mountain name. Thus the DE-tree for "山" would contain all exceptions together with their context. Although it is impossible to enumerate all possible usage of "山" to induce a name, it is possible to find all its exceptions from the training corpus. The details of the generation of DE-tree for each CW can be found in (Liu & Chua 02). As a result of this step, some segmented words many be removed or new word segments are introduced.
c. At each CW or P-Name position, we employ the appropriate name pattern rules (4) to extract the possible names, and refine these names using the context pattern rules (5). The result is a list of possibly overlapping or conflicting names.
d. In case of ambiguities in the list of NEs recognized, we employ the decision tree to find the best possible name in that specific context. The final result is a list of non-overlapping names.
e. When a name occurs for the first time, it is usually expressed in full with the necessary context. However, subsequent occurrences of the same name will have less context and may be abbreviated. To handle such names, we use a separate sub-string matching technique to locate other occurrences of a new name in all likely name positions. In addition, we also employ heuristic rules to detect names that appear in an enumerated list.

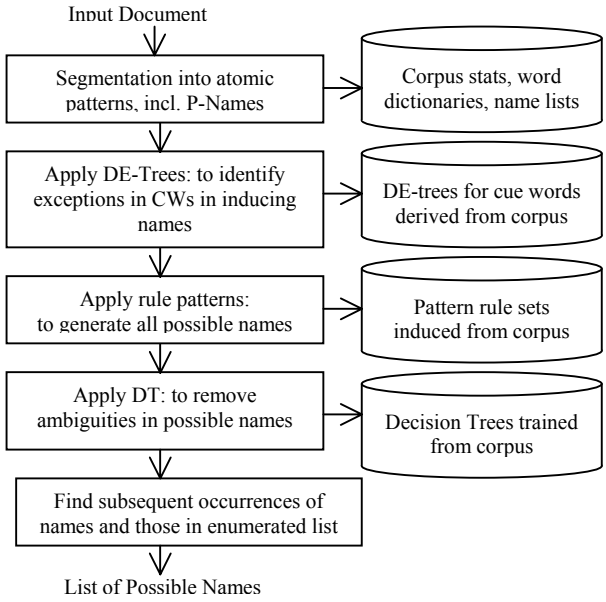

Figure 2: Overall process of extracting NEs

6. Experimental Results and Discussions

6.1 The Datasets used in the Experiments

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 bootstrap approach. The PKU corpus (Yu 1999) contains

one-month of news report from China's People Daily. The corpus is manually segmented into words with appropriate POS tags. It contains 1,121,787 words. We use the PKU corpus and other resources to build the following dictionaries and name lists:

a) We use the PKU corpus to 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 to extract a list of CWs for names. First we select all words tagged as <surname> in a surname list. If a noun occurs before or after a person name, and it has a high occurrence frequency (of over 50), then it is selected as a person context CW. If an organization name or place name composes of more than one word, and the last word is a noun, then the last word is selected as a place or organization suffix. In addition, we supplement this list by using the Chinese name designators derived from the MET2 resource.
c) We obtain a list of Chinese location names from MET2.
d) In order to learn the rules for organization names and the bigram model for P-Name detection, we also collected about 8,000 organization names, and 100,000 P-Names from the web by using a bootstrap approach (Xiao, Liu & Chua 2002).

The resources we derived are available for down loading at http://www.pris.nus.edu.sg/ie.html.

6.2 The Experiment and Results

In our experiment, we use the PKU corpus for training, and the MET2 formal dataset for testing. The results of the experiment are presented in Table 1. For comparison purpose, we also list results reported in Yu et al. (1998) and Chen et al. (1998).

	Type	N_C	N_P	N_W	N_M	N_S	Rc	Pr	F_1
NTU's Results (Chen et al. 98)	Org[n]	293	0	7	77	44	78	85	81.3
	Person	159	0	0	15	56	91	74	81.6
	Place	583	0	65	102	194	78	69	73.2
KDRL's results (Yu et al 98)	Org[n]	331	0	14	32	25	88	89	88.5
	Person	160	0	7	7	74	92	66	76.7
	Place	682	0	1	67	83	91	89	90.0
NUS-PRIS	Org[n]	347	2	14	14	20	92	91	91.5
	Person	171	1	0	2	17	98	91	94.4
	Place	691	0	17	42	61	92	90	91.0

Table 1: The results of MET2 test

The Table tabulates the results in terms of precision (Pr), recall (Rc) and F_1 measures. The definitions of these measures are:

$Pr = (N_C + 0.5*N_P)/(N_C + N_W + N_P + N_M)$
$Rc = (N_C + 0.5*N_P)/(N_C + N_W + N_P + N_S)$
$F_1 = 2*Pr*Rc/(Pr+Rc)$

where N_C, N_P, N_W and N_M respectively give the number of NEs in the test corpus that are correctly recognized, partially recognized, incorrectly recognized, and missed. N_S gives the number of NEs found by the system but not in the tagged list.

The results show that our system significantly outperforms other existing systems in the extraction of all name types, especially for person and organization names. The results demonstrate that our approach in using rule induction learning that is tolerant to word segmentation and name extraction errors is very effective. In particular, our strategy of adopting a greedy approach to locate all likely and overlapping names enables us to find most names together with many false detections. Fortunately, our generalized rule and decision tree approaches are sufficiently robust to remove most false detections. One further point to note is that although we collected large name dictionaries for testing, the base line performance of using these dictionaries directly for name look-up is only less than 30%.

The main sources of errors in our system are:

a) We miss out many Japanese names as our system is not tuned for recognizing Japanese names, which is neither a Chinese name nor P-Name.
b) We miss some person names because of the unexpected format (like the presence of a blank between the surname and the first name) and missing surname in the CW list (like the surname"乂")
c) Some place and organization names are wrongly or inconsistently tagged within and between the training sets.
d) Some common nouns, such as 月球 (moon), 太阳 (sun) and 土星 (Saturn), are tagged as names in the test corpus, they are thus missed out in our results.
e) The Hownet we used is not very complete, and the semantic sets we extracted contain some missing concepts.

Some of these errors are also reported in Chen et al. (1998). We are now in the process of acquiring and testing our system using the 6-month PKU-Corpus. We expect similar results to be achieved on this large-scale corpus.

7. Review of Related Research

Since named entity recognition is almost the first step in Chinese NLP, there have been many researches on this

topic. Most approaches used handcrafted rules, supplemented by word or character frequency statistics. These methods require a lot of resources to model the internal features of each name. Luo & Song (2001) collected large dictionaries for place, person, foreign person, foreign place and organization names. They tested their system on one-month of data from the people's daily and reported high accuracy. However, their resources are not openly available so it is hard to assess the performance of their system. Chen et al. (1998) used 1-billion person name dictionary in their system participated in MET2 test. They used mainly internal word statistics with no generalization. Yu et al. (1998) also collected large person, location and organization name lists. They employed a single framework to model both the context and information residing within the entities, and performed rule generalization using POS and some semantic tags.

In contrast to these systems, we induce rules directly from both the name lists and tagged corpus, and thus require less training resources. We also use lexical chaining to perform rule generalization, instead of POS tags. Finally, we incorporate various approaches to resolve uncertainty and ambiguity in NE recognition.

In many aspects, our system is also similar to those reported in Riloff (1993, 1996), and Soderland et at. (1995) for free text information extraction. However, our system differs from theirs on the main points discussed above.

8. Conclusion

Chinese NE is a difficult problem because of uncertainty in word segmentation and ambiguities in NE location. Many existing techniques that require knowledge on word segmentation, and syntactic or semantic tagging of text cannot be applied. In this paper, we extend the template-based rule induction method popularly used in information and relation extraction tasks. 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 overcoming the problem of data sparseness. Finally, we will extend our work to perform relation and information extraction in Chinese.

Acknowledgments

The authors would like to acknowledge the support of A*STAR, and the Ministry of Education of Singapore for the provision of a research grant RP3989903 under which this research is carried out.

References

Appelt, D.E., et al. 1995. SRI International FASTUS System MUC-6 Test Results and Analysis. *Proc. of the Sixth Message Understanding Conference.* 237–248. Morgan Kaufmann Publishers.

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

Borthwick, A. 1999. A Maximum Entropy Approach to Named Entity Recognition. Ph.D. Thesis, New York Univ.

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 Seventh Message Understanding Conference*

Cucerzan, S. and Yarowsky, D. 1999. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. *Proc. of Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, 90-99

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

Green, S.J. 1999, Lexical Semantics and Automatic Hypertext Construction. *ACM Computing Surveys* 31(4), Dec.

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

Liu, J.-M. and Chua, T.S. 2001. Building Semantic Perceptron Net for Topic Spotting, *Proc. of the Association for Computational Linguistics*, 306-313

Liu, J.-M. and Chua, T.S. 2002. A Hybrid Rule Induction Model for Chinese Name-Entity Recognition. Technical Report, School of Computing, National University of Singapore.

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.

Marsh, E. & Perzanowski, D. 1998. MUC-7 Evaluation of IE Technology: Overview of Results, *Proc. of the*

Seventh Message Understanding Conference, at: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.

Muslea, I. 1999. Extraction Patterns for Information Extraction Tasks: A Survey. *Proc. of AAAI Workshop.*

Palmer, D.D. 1997. A Trainable Rule-based Algorithm for Word Segmentation. *Proc. of the Association for Computational Linguistic, and Eighth Conference of the European Chapter of the Association for Computational Linguistics.*

Riloff, E. 1993. Automatically Constructing a Dictionary for Information Extraction Tasks. *Proc. of AAAI-93*, 811-816

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

Sekine, S., Grishman, R. and Shinnou, H. 1998. A Decision Tree Method for Finding and Classifying Names in Japanese Texts. *Proc. of the 6th Workshop on Very Large Corpora*, Montreal, Canada

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

Sproat, R., Shih, C., Gail, W. and Chang, N. 1996. A Stochastic Finite-state Word Segmentation Algorithm for Chinese. *Computational Linguistics*, 22(3), 377-404

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

Xiao, J., Liu, J.-M. and Chua, T.S. 2002. Extracting Pronunciation-translated Named Entities from Chinese Texts using Bootstrapping Approach. Submitted to *2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'2002).*

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

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

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