The Construction of a Chinese Named Entity Tagged Corpus: CNEC1.0

Cheng-Wei Shih, Tzong-Han Tsai, Shih-Hung Wu,
Chiu-Chen Hsieh, and Wen-Lian Hsu

Institute of Information Science, Academia Sinica
{dapi,thtsai,shwu,gladys,hsu}@iis.sinica.edu.tw

Abstract. In order to build an automatic named entity recognition (NER) system for machine learning, a large tagged corpus is necessary. This paper describes the manual construction of a Chinese named entity tagged corpus (CNEC 1.0) that can be used to improve NER performance. In this project, we define five named entity tags: PER (person name), LOC (location name), ORG (organization name), LAO (location as organization), and OAL (organization as location) for named entity categories. In addition, we propose a special tag, DIFF (Difficulty), to annotate ambiguous cases during corpus construction. A, corpus-annotating procedure, a tagging tool, and an original corpus are also introduced. Finally, we demonstrate a part of our manual-tagged corpus.

1 Introduction

Named entity recognition (NER), which includes the identification and classification of certain proper nouns in a text, is an important task in information extraction. It is useful in many natural language processing systems for document indexing and managing data with named entities [Tsai et. al 2004]. Since numerous new proper nouns are generated every day, it is not enough for an IR system to index names from Internet documents or refer to gazettes. Therefore, NER has become an important method for information processing in recent years.

Machine learning (ML) is one of the most popular methods in NER, due to its easy maintenance and portability [Tsai et. al 2004]. Typical machine learning approaches applied in NER include the Hidden Markov Model (HMM) [Bikel et. al 1997], Support Vector Machine (SVM) [Asahara 2003], and Maximum Entropy (ME) [Borthwick 1998]. No matter which approach is used, a tagged named entity corpus with clear annotating criteria is needed in the training phase of building an NER system. However, constructing such a corpus is a labor-intensive task, so few researchers have focused on it. The Automatic Content Extraction program (ACE) executed by the Linguistic Data Consortium (LDC) [Http://wave.ldc.upenn.edu/] annotates seven common entities in English, simplified Chinese, and Arabic. Meanwhile, the IREX (Information Retrieval and Extraction Exercise) [Http://nlp.cs.nyu.edu/irex/] defines 8 kinds of named entities (NE) in Japanese [Sekine and Isahara 2000], and the shared task in CoNLL 2002 and 2003 [Erik 2002] [Erik et. 2002] develops the NER system using four types of NE in English, German, Dutch, and Spanish. However, as none of these methods focus on traditional Chinese, there is an urgent need for a traditional Chinese NE corpus and NE annotating standards to support an automatic Chinese NER system like Mencius [Tsai et. al 2004].

The categories of named entities defined by Message Understanding Conferences (MUC) are the names of persons, organizations, locations, temporal expressions and number expressions [Grishman and Sundheim 1996]. Since temporal and number expressions, such as “the past year” and “40 percent”, are generally used as adjectives to describe other entities, we disregard them and focus on the annotation of person names, organization names, and location names as NEs. We separate organizations and locations into four non-overlapping categories to accommodate common Chinese usage. We also propose a temporary tag, “Difficulty”, to represent named entities that are ambiguous.

The remainder of this paper is organized as follows. Section 2 discusses the main issues of labeling named entities. Section 3 introduces all the NE categories used in CNEC1.0. Section 4 describes our annotation procedure and environment. Finally, in Section 5, we present our conclusion and the direction of future research.

2 Named entities annotation issues

The applications of a corpus determine the kinds of entities to be tagged. In our project, the NER system should support information extraction, question answering, and information retrieval of new documents. The following three issues should be considered before tagging Chinese named entities.

2.1 Proper nouns

We think named entities should be proper nouns. Therefore, each named entity should denote a unique object, so words without uniqueness should not be annotated. For example, in the sentence “他把車停放在停車場/He parked his car in the parking lot.”, we cannot be sure which parking lot he parked in. Therefore, the term “停車場/parking lot” will not be labeled

2.2 Inner feature and Outer feature

Generally, a named entity can be determined in three ways: viewing its literal meaning, checking the context, and semantic understanding as shown in the following example:

“台北市長馬英九/ Mayor of Taipei City, Ma Ying-Jeou” (1)

“台北車站人潮洶湧/Taipei Main Station is crowded.” (2)

“我聽說遠東搬家了/I heard that Yuan-Dong has moved out.” (3)

Obviously, we can easily determine that the term “馬英九/ Ma Ying-jeou” in Sentence (1) is a person name according to the position of the title “台北市長/ Mayor of Taipei City”. “台北車站/Taipei Main Station” in sentence (2) can be identified as a location name because of the term “車站/Station”. However, sometimes we cannot identify or judge a word as a proper noun by its literal meaning. For example, in Sentence (3), we do not know if the term “遠東/ Yuan-Dong” represents a person or a company. These kinds of inaccuracies are due to abbreviations or borrowings. For this reason, two types of feature are used to classify the recognition modes of named entities：the inner feature and the outer feature. In the above examples, sentence (1) is the outer feature type, while sentence (2) can be classified by its inner features. In our work, we do not limit the types of features annotators apply during tagging, but if a named entity cannot be identified by both inner and outer features, as in sentence (3), we ask annotators not to mark it. This eliminates confusing terms and keeps the corpus as clear as possible

2.3 Maximum and minimum semantic unit matching

Named entities are occasionally nested or appear next to one another in a text. In some cases we can combine them to form a larger entity because they may describe the same object. Therefore, determining the boundary of named entities is an important issue. We have found that different named entities have a corresponding annotation policy, which can be classified as maximum and minimum semantic unit matching. Minimum semantic unit matching is recommended for named entities such as person names and location names because these entities singly represent a unique item. For example, the sentence “台北縣板橋市/Ban-Qiao City, Taipei County” is tagged as two place names because “台北縣/Taipei County” and “板橋市/Ban-Qiao City” both denote specific independent entities. On the other hand, named entities such as organizations should apply the maximum unit policy. The term “台北市環保局/Department of Environmental Protection, Taipei City Government” cannot be separated into “台北市/Taipei City” and “環保局/Department of Environmental Protection” for retaining the original meaning.

3 Named entity categories

We propose five target named entity categories for annotating the "unique identifiers" of entities, including organizations, persons and locations, as well as one function tag. These are shown in Table 1 *and explained* in the following sub-sections. For practical purposes, we began our experiment with these NE tags.

Table 1. Tag set of the NE corpus

DIFF	**Difficult problem**
PER	**Person name**
LOC	**Location name**
ORG	**Organization name**
LAO	**Location as Organization**
OAL	**Organization as Location**

3.1 Difficult problem - DIFF

Diff (Difficult problem) is designed to identify problems such as nested, ambiguous, or poorly defined NEs. It addresses controversial items in Chinese named entity identification.

In our opinion, DIFF is essential for identifying ambiguous cases. Named entities that are difficult to classify are isolated from others for data cleansing to ensure that the content of the corpus is clear. DIFF entities will become the future expansion direction of our NER processing domain.

3.2 Person name - PER

Traditionally, the structure of Chinese person names follows the principle that the surname (one or two characters) is placed before the person's chosen names (one or two characters). In our research, the annotation of person names follows this principle. But some entities with "person" meaning as Diff tag such as nicknames, incomplete Chinese person names, foreigners' names and pronouns, are marked as Diff. These exceptions are discussed below.

3.2.1 Nicknames

Nicknames are not only given to people, but are sometimes given to pets or even objects like toys and vehicles. Because of their uniqueness, we mark nicknames as DIFF within a context, as the following example shows.

[小炳<DIFF>] 疼愛的女兒 [央央<DIFF>]

[xiao-bing<DIFF>] loves his daughter [yang-yang<DIFF>]

3.2.2 Incomplete Chinese person names

Following the full name principle, an incomplete Chinese person name indicates that the surname or chosen name may be omitted. For instance, "Shin, Cheng-Wei" is a full person name, but we sometimes only use the chosen name "Cheng-Wei" to address the person. Another example of an incomplete person name is a surname that follows a title or an appellation such as "李先生(Mr. Lee)". Both of these cases are tagged as DIFF.

[陳總統<DIFF>] 上午前往日本北海道遊玩

[President Chen<DIFF>] went to Hokaido for sightseeing this morning.

3.2.3 Foreigners' names

Chinese NER has difficulty dealing with foreigners' names because of the following name constructions: direct translation, Japanese person names, and Korean person names. First, direct translation cannot normally meet the principle of Chinese person names. For example, the Chinese translation of "Mel Gibson" is "梅爾吉勃遜(Mei-er-ji-bo-xun)", which obviously doesn't meet the naming rule. Second, Japanese mostly uses Chinese characters for person names, but, there isn't a surname or composite first name as in Chinese person names. For example, in "日本首相(Japanese prime minister)[小泉純一郎](Junichiro Koizumi)", Koizumi is his surname

and Junichiro is his first name. These names don't match the Chinese person naming principle. Finally, we treat Korean person names as PER because most of them match the Chinese naming rules, e.g. "李英愛 /Lee-Ying-Ai".

As most foreigners' names may cause confusion in Chinese NER, we use DIFF tag as a temporary solution in CNEC 1.0 to solve the problem.

[梅爾吉勃遜<DIFF>] 導演受難記名利雙收

[Mel Gibson<DIFF>] has achieved both fame and wealth by directing the movie "The Passion of the Christ."

3.2.4 Pronouns

Some NER researchers claim their systems can handle pronouns as named entities for person names. However, because of the "uniqueness" of NEs, pronouns are beyond our scope and we do not annotate them as NE tags.

3.3 Location names - LOC

Basically, a location name pinpoints a place's geographical position on an accurate map or in other reference material. Proper names like "Hyde Park", "New York Art Theater" or "Berlin Wall" are suitable for NER, but terms like "a park", "a theater", or "a wall" are not. So the main purpose in tagging location names is to recognize an existent location in geographic.

A location name included in another compound word such as "西班牙海鮮飯"(Spanish seafood rice) is an issue in NE annotation. In Chinese, a term's noun and adjective forms are the same, In this case, "Spanish" and "Spain" are translated as the same Chinese word(西班牙). Therefore, we suggest a syntactic frame: insert "de（的）" between a possible location name and the other words close to (A de(的) B) to solve such cases.

Chinese Word 1:	西班牙海鮮飯
In English:	Spanish seafood rice
[A de B] in Chinese:	[西班牙][的][海鮮飯]
In English:	[Seafood Rice] [of] [Spain]

Maximum and minimum semantic unit selection (section 2.3) is regarded as a Chinese segmentation problem. For example, in the phrase "美國(United States)德州(Texas)奧斯汀(Austin)", there are no spaces between the words in written Chinese. Location follows minimum semantic unit matching to tag the example as [美國(United Sate)<LOC>]、[德州(Texas)<LOC>]、[奧斯汀(Austin)<LOC>]. In addition to the basic tagging rule, several location types have to be labeled

3.3.1 Roads, sections, and addresses

Address' location information should be marked from country name, state, city, road (Boulevard, avenue, street, etc.,) to section. Other information in an address is excluded.

[忠孝東路四段<LOC>] 100號

[Zhong-Xiao East Road, Section 4<LOC>], No. 100

Sometimes a road section can be described in a city-section or area-section. In this case the road name and its description should be tagged separately.

[西濱快速道路<LOC>] [嘉義<LOC>] 段

[Xi-Bin Express Way<LOC>] [Jia-Yi<LOC>] Section

3.3.2 Location abbreviations

Location abbreviations are terms composed of two or more place names in a single entity. The other type of location abbreviation is multi-name expression containing conjoined modifiers. For instance, 中(Taichung)彰(Changhua)投(Nantou)地區(area) is a common way to describe the location of three neighboring cities in Taiwan. We suggest that such cases should be tagged as DIFF.

[桃竹苗<DIFF>] 地區連日豪雨

[Tao-Chu-Miao<DIFF>] area it has been raining torrentially for a couple of days.

Tao-Taoyuan; Chu: Hsihchu; Miao: Miaoli

3.3.3 World place names

World place names can be divided into two sets: locations written in a foreign language and translated location names (written in Chinese). A translated name such as "紐約/New York" can be considered as a location name in CNEC 1.0. But an original name, like Tokyo, should be tagged as DIFF. The following two sentences demonstrate the criteria we set.

[密蘇里州(Missouri State)<LOC>] 的 [聖路易市(St. Louis City)<LOC>]

[Missouri<DIFF>]州的 [St. Louis City<DIFF>].

#[St. Louis City<LOC>] in [Missouri State<LOC>]

3.4 Organization names - ORG

In general, organizations include companies, government bodies, institutes, and other organized groups. We define an organization as having the ability to execute plans and projects. The tagging of ORG has to apply maximum semantic unit matching. A typical case is shown below.

[裕隆汽車三義廠(Yulon Motor Sanyi plant)<ORG>]

The maximum semantic unit is used because this entity cannot be separated into "裕隆汽車(Yulon Motor)" and "三義廠(Sanyi plant)". "Sanyi plant" cannot be tagged as a LOCATION according to the uniqueness characteristic of an NE. Like location names, some ambiguity may occur in the following cases.

3.4.1 Organization abbreviations

Organization abbreviation tagging follows the rules for location abbreviations described in Subsection 3.3.2. For example,

[國親新(Guo-Qin-Xin)<DIFF>] 議員下午到[健保局(BNHI)<ORG>]

Guo-Qin-Xin councillors went to the Bureau of National Health Insurance this afternoon.

Guo: Kuomintang; Qin: People First Party; Xin: New Party

3.4.2 Foreign organization names

As in Section 3.3.3, an original organization name is regarded as DIFF in CNEC 1.0, but a translated name is tagged as a location name.

[惠普<ORG>] 與 [微軟<ORG>] 的共同秘密

The secret of [HP<ORG>] and [Microsoft<ORG>]

3.4.3 Groups, bands, crowds, and teams

Through the tagging process, we have found that most Chinese people have a problem tagging similar concepts like groups, bands, crowds and teams as organizations. Hence, we give a definition of an organization to help annotators determine if a term is an ORG. “An organization has five fundamental parts: a founder, capital, structure (departments, section, class, etc.,), a hierarchy (chief, director, dean) and employees.” According to this definition, we do not mark a term as an organization if it doesn’t have an organized structure.

3.5 LAO and OAL

Location as Organization (LAO) and Organization as Location (OAL) are proposed for semantically meaningful NE in Chinese NEC. In some cases, “location” represents an organization-like role to make decisions or to perform some duties. Compare the following two examples from the China Times news corpus:

1. “[台北市政府<ORG>]同意[總統府<OAL>]前的集會遊行”

[Taipei municipal government<ORG>] agreed to the protest in front of the [presidential plaza <LOC>].

2. [總統府<ORG>]發表一中一台政策

[Presidential Office <LAO>] announced “one China, one Taiwan policy”.

The above examples show that the term “總統府(presidential plaza / presidential office)” has a double meaning in Chinese: location and organization. The first term “總統府” is obviously a place name, but, the second “總統府” is an organization. This use of the same term to indicate a place name and an organization name, we call it “borrowing”, is common in Chinese. Sometimes we cannot sure if a location entity is a real location name or just a borrowing. In order to avoid confusion, we separate the LAO tag and OAL tag in location and organization names.

Differentiating between a borrowing and a true NE depends on an entity’s category., By deciding which category an entity belongs to in common usage, we can tell whether it is a borrowing, or not. For example, a country’s name can refer to its geographical position, but it may also be used as an organization name, as in: “中國昨日警告美國停止對軍售/China warned the United States yesterday to stop selling advanced arms to Taiwan.” Obviously, in this case we can tell the country names are borrowings and should be annotated as LAO as follows:

[中國<LAO>] 昨日警告 [美國<LAO>] 停止對 [台灣<LAO>] 軍售

[China<LAO>] warned the [United States<LAO>] yesterday to stop selling advanced arms to [Taiwan<LAO>]

OAL can be identified in the same way:

這輛巴士有到 [行政院<OAL>]

This bus goes by [the Executive Yuan<OAL>]

4 Manual Annotating Process

The most famous corpus in Chinese NER is MET-2 made by MUC [Chinchor, 1998]. However, it only contains single domain data and is not large enough for building a machine-learning-based NER system [Tsai et. al 2004]. We, therefore, collected over a million sentences without any annotations from the online United Daily News (UDN) and China Times for the period December 2002 to December 2003 as raw data. The sentences extracted from raw data recorded in XML format shown in Figure 1.

```
- <Sentence id="id383442" text="所以成了男女朋友">
    <TagGroup type="NameTagging" />
    <Log type="human-tagging" duration="0" />
  </Sentence>
- <Sentence id="id180560" text="中華文化重視家庭、家族關係">
    <TagGroup type="NameTagging" />
    <Log type="human-tagging" duration="0" />
  </Sentence>
- <Sentence id="id1122061" text="樂透頭獎加碼活動將於今日暫時劃下休止符">
    <TagGroup type="NameTagging" />
    <Log type="human-tagging" duration="0" />
  </Sentence>
- <Sentence id="id103554" text="甚至播放視訊檔案等多媒體用途">
    <TagGroup type="NameTagging" />
    <Log type="human-tagging" duration="0" />
  </Sentence>
```

Fig. 1. Original corpus

We chose high school students as annotators and gave them basic training before they performed the annotations. The training process was:

1. All the students attended courses about the project, including an introduction to named entity recognition, segmentation, and parts-of-speech tagging.
2. Students took a qualifying test to select participants for the tagging task.
3. Participants had to acquaint themselves with the annotating criteria we suggest ed and the operation of the tagging tool program. (Figure 2 shows the interface of the tagging tool.) The participants were then divided into three groups.

Each group was given the same sentence set containing 21,000 randomly extracted sentences; 13,208 from the UDN and 8,892 from the China Times. Table 2 shows the distribution of sentences. Participants were asked to finish the tagging task with the tagging tool program in two weeks. Figure 3 shows a tagged XML file in which the annotations are marked. Tagging results were then collected from each group and checked for consistency.

Fig. 2. Tagging tool

Table 2. Raw sentences count for each domain

大陸新聞	中時電子報	638	文化藝術	聯合新聞網	1362
生活消費	中時電子報	226	本日焦點	聯合新聞網	1019
地方新聞	中時電子報	143	生活消費	聯合新聞網	848
社會新聞	中時電子報	883	地方新聞	聯合新聞網	1662
政治新聞	中時電子報	791	重點新聞	聯合新聞網	2305
重點新聞	中時電子報	2987	旅遊休閒	聯合新聞網	608
旅遊休閒	中時電子報	318	財經產業	聯合新聞網	2303
財經產業	中時電子報	672	意見評論	聯合新聞網	519
意見評論	中時電子報	868	資訊科技	聯合新聞網	598
資訊科技	中時電子報	647	影視娛樂	聯合新聞網	1332
影視娛樂	中時電子報	719	醫療保健	聯合新聞網	652

Fig. 3. Tagged corpus. The marked area shows the annotations the participants made: “pos” means the named entity’s starting location in the sentence and “len” is the length of the NE. The label “name” indicates what kind of NE the word is.

5 Conclusion and Future work

In this paper we describe the construction of a tagged corpus for Chinese NER. We define the criteria of Chinese NE tagging, and design a standard tagging procedure for NE corpus annotation. We also demonstrate an annotator training procedure and the statistics of the corpus. The resulting corpus, CNEC 1.0, can be used to improve the performance of Mencius, our Chinese NER system. We do not use some ambiguous entities, labeled as DIFF, that involve issues such as abbreviations, cross-language loanwords and borrowings for training the NER model. As these entities need to be re-classified, advanced annotation will be executed in the next version of CNEC.

References

[1] D. Bikel, S. Miller, Richard Schwartz and Ralph Weischedel: “Nymble: a High-Performance Learning Name Finder” Proceedings of the Fifth Conference on Applied Natural Language Processing, 1997, pp. 194-201.

[2] M. Asahara and Y. Matsumoto: “Japanese Named Entity Extraction with Redundant Morphological Analysis”, HLT- North American Chapter of the Association for Computational Linguistics, Edmonton, Canada, 2003,
[3] A. Borthwick, J. Sterling, E. Agichtein, R. Grishman, “Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition”, Sixth Workshop on Very Large Corpora, 1998.
[4] S. Sekine, R. Grishman, H. Shinnou “A Decision Tree Method for Finding and Classifying Names in Japanese Texts”, Sixth Workshop on Very Large Corpora, 1998.
[5] F. Erik Sang T.K., "Introduction to the CoNLL-2001 Shared Task: Language-Independent Named Entity Recognization", Proceeding of the 6th Conference on Natural Language Learning 2002 (CoNLL 2002), pp. 155-158
[6] F. Erik Sang T.K., Meulder, F.D., "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition", Proceeding of the 7th Conference on Natural Language Learning 2003 (CoNLL 2003)
[7] R. Grishman, B. Sundheim, “Message Understanding Conference - 6: A Brief History”, Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, June 1996.
[8] N. Chinchor, "MUC-7 Named Entity Task Definition(Version 3.5)" The 7th Message Understanding Conference (MUC), 1998
[9] Tsai, T.H., Wu, S.H., Lee, C.W., Shih, C,W., Hsu, W.L.: “Mencius: A Chinese Named Entity Recognizer Using The Maximum Entropy-Based Hybrid Model” Computational Linguistics and Chinese Language Processing, Vol.9, No.1, 2004, pp.65-82.