Named Entity Recognition in Chinese News Comments on the Web

Xiaojun Wan[a], Liang Zong[b], Xiaojiang Huang[c], Tengfei Ma, Houping Jia,
Yuqian Wu and Jianguo Xiao
Institute of Compute Science and Technology, The MOE Key Laboratory of
Computational Linguistics, Peking University, Beijing 100871, China
{[a]wanxiaojun, [c]huangxiaojiang}@icst.pku.edu.cn
[b]zongliang.cn@gmail.com

Abstract

News comment is a new text genre in the Web 2.0 era. Many people often write comments to express their opinions about recent news events or topics after they read news articles. Because news comments are freely written without checking, they are very different from formal news texts. In particular, named entities in news comments are usually composed of some wrongly written words, informal abbreviations or aliases, which brings great difficulties for machine detection and understanding. This paper addresses the task of named entity recognition in Chinese news comments on the Web. We propose to leverage the entity information in the referred news article to improve named entity recognition in the news comments. Three different schemes are investigated to find useful entities in the news article for new feature generation in the CRFs model. Finally, a dictionary-based correction step is employed to further improve the results. We manually labelled a benchmark dataset with 60 pieces of news and 6000 comments downloaded from a popular Chinese news portal – www.sina.com.cn. The experimental results on the dataset show that our method is effective for this special task.

1 Introduction

Named entity recognition (NER) is one of the fundamental tasks in the field of natural language processing. It has been widely used in the areas of information retrieval, machine translation, question answering, and so on. In most literatures, named entities are defined as entity names (person names, location names and organization names). With the advent of MUC, CONLL, ACE and SIGHAN evaluations, NER has received much attention of the researchers and hence achieved great development.

News comment is a new text genre in the Web 2.0 era, and many people often write and post comments to express their opinions on recent news events or topics after they read news articles. As the roles which people play on the internet have gradually changed from acquirers to suppliers, news comments have become one of the most valuable information resources. For example, on one of the popular Chinese news portals – sina.com.cn, every piece of hot news is associated with over 500 comments. Named entity recognition is the basis of many other news comments understanding and mining applications, including entity relation extraction, opinion holder and target extraction in news comments, and so on.

Because news comments are freely written by different persons with different education backgrounds and writing styles, they are very different from formal news texts. In particular, Chinese news comments have the following properties:

1) The texts in news comments are very informal and noisy, especially for entity names. There are always many noisy pieces of texts in the comments because the comments are written by various users, e.g., wrongly written words, extra spaces, meaningless characters, or informal names, etc. For example, "汇源/Huiyuan" may be wrongly written to the word "汇圆/Huiyuan".

2) News comments are written with various styles. Since the comments are written by different users with different backgrounds, each one has its own writing style. Different users may use different words and phrases to express the same entities. For example, "八一队/Bayi Team" may be written as "81 队".

3) The texts in news comments are usually very short and concise. The average length of each comment is about 20~25 characters in our dataset. Many entity names in news comments are abbreviated in various ways. For example, "信息科学技术学院/School of Electronics En-

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 856–864,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

gineering and Computer Science" may be abbreviated as "信院", "信科" or "信息".

4) Most news comments are relevant to the news topics in the referred news article. There is usually a strong relationship between news comments and the news article. Most news comments are focusing on the entities or events introduced in the news article, or related entities and events. For example, given an news article about "姚明/Yao Ming", some news comments may talk about the players and teams explicitly mentioned in the news article, such as"姚明/Yao Ming" or "休斯敦火箭队/Houston Rockets", and other news comments may talk about related players and teams in NBA, such as "奥尼尔/O'Neal" and "菲尼克斯太阳队/Phoenix Suns".

The first three properties bring great challenges for named entity recognition from Chinese news comments. But the fourth property brings very useful knowledge for the task. In this study, we focus on the task of named entity recognition in Chinese news comments on the web, which has not been investigated yet. Considering the close relationships between named entities in the referred news article and named entities in the news comments, we propose to leverage the entity information in the news article to improve named entity recognition in the news comments. Three schemes are exploited for collecting useful entities from the news article, and then the entity information is incorporated into the CRF-based algorithm for recognizing named entities in the news comments. Finally, an additional correction step is used to further improve the performance.

We manually labeled an evaluation dataset with 60 pieces of news and 6000 news comments from a popular Chinese news portal (www.sina.com.cn). Experimental results on the dataset show that our proposed approach is effective for the task of NER in news comments. The use of focused entities and related entities in the referred news article is very beneficial for the task.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 presents our proposed approach. Section 4 shows the experimental setup and results. Finally, Section 5 summarizes the conclusions.

2 Related Work

Traditional named entity recognition systems use linguistic grammar-based techniques as well as statistical models. The techniques can be categorized into rule-based (Sekine and Nobata, 2004; Chiticariu et al., 2010), machine learning-based (including unsupervised, supervised and semi-supervised methods) (Bikel et al., 1999; Mayfield et al., 2003; McCallum and Li, 2003; Bender and Ney, 2003; Florian et al., 2003; McCallum and Li, 2003; Etzioni et al., 2005; Klementiev and Roth, 2006; Okanohara et al., 2006; Finkel and Manning, 2009; Singh et al., 2010) and hybrid models (Srihari et al., 2001). The most popular statistical models for named entity recognition include Support Vector Machine, Hidden Markov Model, Maximum Entropy Model, Conditional Random Fields, and so on. Background knowledge derived from Wikipedia and WordNet has been used for improving the NER task (Kazama and Torisawa, 2007; Richman and Schone, 2008; Pennacchiotti and Pantel, 2009). Most previous works have investigated the NER task over formal text such as news articles. These kinds of texts are written by professional writers, and thus they are well organized and well-structured, and they have seldom grammatical and spelling errors and noises.

Recently, a few works have investigated the task over informal English texts such as emails and blogs. Huang et al. (2001) address the problem of extracting identity and phone number of the caller from voicemail messages, and they present three typical information extraction methods: hand-crafted rule-based method, maximum entropy models, and probabilistic transducer induction. Jansche and Abney (2002) present a two-phase procedure consisting of a hand-crafted component and a classifier for information extraction from voicemail. Minkov et al. (2005) propose to use email-specific structural features and a recall-enhancing method for improving person name recognition from email. Gruhl et al. (2009) explore the application of restricted relationship graphs and statistical techniques to improve named entity annotation in on-line forum texts discussing popular music.

Generally speaking, the Chinese NER task is harder because Chinese texts have no explicit word segmentation information, and Chinese named entities lacks the capitalization information that plays an important role in signaling named entities, and moreover, the structures of Chinese named entities are more complicated, especially for entity abbreviations. Most Chinese NER systems adopt statistical models or hybrid solutions. Sun et al. (2002) consider the problem of Chinese named entity identification using statistical language model, and they integrate word segmentation and NE identification into a unified framework that consists of several class-based language models. Fang and Sheng (2002) present a hybrid approach, which combines a machine learning method and a rule based method, to improve the Chinese NE system's efficiency. Zhu et al. (2003) adopt the source-channel model framework for the single character named entity recognition. Gao et al. (2005) propose a pragmatic mathematical framework in which segmenting known words and detecting unknown words of different types can be performed simultaneously in a unified way. Wu et al. (2005) present a statistical model with human knowledge which treats NER as a probabilistic tagging problem. Fu and Luke (2005) present a lexicalized HMM-based approach to Chinese NER. Yu et al. (2008) use a Markov Logic Network to combine various types of domain knowledge to correct the output of the Conditional Random Fields model. Zhao and Kit (2008) propose a supervised learning model which combines the unsupervised segmentation results to improve performance. Most of the researches in this field are restricted to formal text corpus, such as newswire articles. To our knowledge, our work is the first attempt to investigate the named entity recognition task over the informal Chinese news comments.

3 Our Proposed Approach

3.1 Overview

As mentioned earlier, named entities in Chinese news comments are more difficult to recognize than those in formal text corpus because of the informal written style of the comments. Given a Chinese news article and its associated comments, our task aims to recognize all named entities (person names, location names and organization names) in the comments.

The basic idea of our proposed approach is to leverage the close relationships between named entities in the news comments and named entities in the news article. We first find a few useful named entities in the news article and then incorporate the entity information into the basic NER tagging algorithm. Finally, we use an additional correction step to further improve the performance. Our approach adopts the CRF-based algorithm for named entity recognition, and we focus on how to find useful entity information from the news article for improving NER in the news comments. Figure 1 shows the framework of our proposed approach. The three key components will be presented in next sections, respectively.

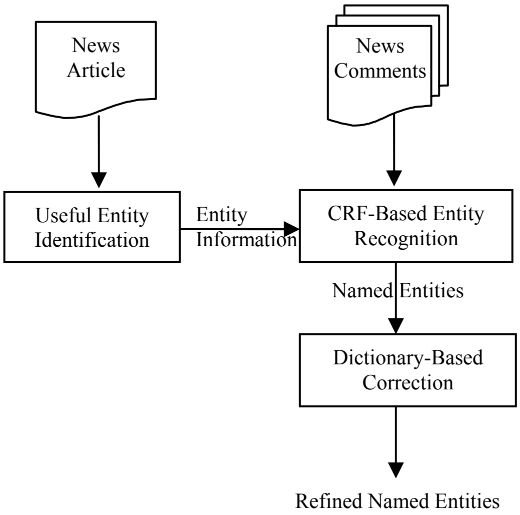
News Article
News Comments
Useful Entity Identification
Entity Information
CRF-Based Entity Recognition
Named Entities
Dictionary-Based Correction
Refined Named Entities

Figure 1: The framework of our proposed approach

3.2 The CRF-Based Entity Recognition Algorithm

The Conditional Random Fields (Lafferty et al., 2001) model is a probabilistic framework to segment and label sequence data, and it has been successfully used in many NLP tasks, including word segmentation, POS tagging, named entity recognition, and so on. The CRFs model defines the conditional distribution $p(Y|X)$ of the labels Y given the observations X with the following formula:

$$P_\lambda(Y \mid X) = \frac{1}{Z(X)} exp(\sum_{c \in C} \sum_{k} \lambda_k f_k(Y_c, X, c))$$

Y is the label sequence; X is the observation sequence; $Z(X)$ is a normalization term; f_k is a feature function; λ_k is the weight of feature function f_k; C is the set of cliques in the undirected graphic model. Given the training data with a set of sentences (characters with their corresponding tags), the parameters of the model are trained by maximizing the conditional log-likelihood. In the testing phase, given a test sentence x, the tagging sequence y is given by $Argmax_y P(y|x)$.

CRFs has been shown to perform well on the task of Chinese named entity recognition (Zhou et al, 2006, Chen et al, 2006, Yu et al, 2008). We treat the Chinese NER task as a character-based sequence labeling problem and use the CRFs model for learning and inference. In this study, we use the linear-chain CRFs model implemented in the CRF++ toolkit[1]. We use four tags (B - beginning of an entity, I - inside of an entity, E – end of an entity, S – a single character entity) for tagging each type of named entity, and thus we have totally 13 tags (including the tag O- outside of any entity). Standard features used in our basic method are listed in Table 1. The segmentation features and POS features are given by our in-house Chinese word segmentation and POS tagging tools. The segmentation features indicate the positions of the Chinese characters in the corresponding Chinese words after word segmentation, and the position of each Chinese character is represented by three types: the first character in a word, the last character in a word, and the middle character in a word. For the POS features, the POS tag of each Chinese character is the same with that of the Chinese word which the character belongs to.

Character features	C_n, $n\in$ [-3,3]
	C_nC_{n+1}, $n\in$ [-3,2]
Segmentation features	S_n, $n\in$ [-2,2]
	S_nS_{n+1}, $n\in$ [-2,1]
POS features	P_n, $n\in$ [-2,2]
	P_nP_{n+1}, $n\in$ [-2,1]

Table 1: Standard features used in the baseline method

Entity features	F_n, $n\in$ [-2,2]
	F_nF_{n+1}, $n\in$ [-2,1]

Table 2: Entity-based features used in our proposed method

The basic method using the above standard features does not work well because the standard features rely only on the news comments, while the named entities in the news comments may appear in different informal or erroneous forms. Since the comments have a very close relationship with the corresponding news article, how to effectively use the named entities' information in the news article is the primary problem in our proposed method. After we extract a few useful entities from the news article, we propose to generate new entity features and add them to the CRFs model.

For each entity type, we collect one prefix list and one suffix list by extracting the prefixes and suffixes from the useful entities. Based on these lists, we assign additional tags to the characters in the news comments to indicate whether the characters are included in a particular list. For instance, the character "刘" \ "Liu" in a person name "刘翔" \ "LiuXiang" is extracted as a prefix of the person name. When a character in a comment is included in a person prefix list, we assign a tag of "Per_Prefix_1", otherwise, we assign a tag of "Per_Prefix_0". The new entity features are listed in Table 2.

The key issue in this study is how to find useful entities from the news article, and we propose three schemes for addressing this problem in next sections.

3.3 Useful Entity Identification

In this section, we propose three schemes for finding useful entities in the news article. The first scheme aims to recognize and use all named entities in the news article. The second scheme aims to extract and use only focused named entities in the news article. The third scheme aims to expand the focused named entities by using web search results.

3.3.1 All NE Recognition

This scheme is the simplest scheme based on the assumption that all the named entities in the news article are useful clues for the NER task in the associated news comments. In this study, we use our in-house Chinese NER tool to tag all named entities in the news article and use all the named entities as useful entities for new feature generation, as mentioned in Section 3.2. Our in-house Chinese NER tool is based on the CRFs model. The F-measure values of the tool over the MSR NER news corpus are 94.14% for person entities, 87.03% for location entities and 84.97% for organization entities.

3.3.2 Focused NE Extraction

This scheme advances the first scheme by finding only a few important named entities in the news article. There are usually many named entities in the news article, and the named entities are unequally important. The associated comments usually focus on discussing a few important entities, which are called focused named entities. The use of all the named entities may introduce some noises, and instead we use only the focused named entities in this scheme.

Focused entities refer to the named entities which are most relevant to the main topic of a news article. Similar to Zhang et al. (2004), we consider the task of focused named entity extraction as a binary classification problem. Given a named entity in the news article, we use a classi-

[1] http://crfpp.sourceforge.net/

fication model to classify it as focused entity or not. For a named entity *A*, the features used in our classification model are listed in Table 3.

Feature Name	Feature Value	Feature Description
Entity Type	Integer (0,1,2)	Entity type of *A*
Entity Frequency	Positive Integer	Occurrence number of *A* in news article
In Title or Not	Boolean (T, F)	Whether *A* appears in the title or not
Entity Document Frequency	Positive Integer	Number of news articles contain A
Entity Distribution	Positive Float	The distribution of *A* in news article

Table 3: Features used for focused NE extraction

In Table 3, the first four features are easy to understand. The entity distribution feature measures how evenly an entity is distributed in a document. The motivation is that if a named entity occurs in many different parts of a document, it is more likely to be an important entity. We use the entropy of the probability distribution to measure it. Considering a document which is divided into *m* sections with equal length, a named entity's probability distribution is represented by $\{p_1, p_2, ..., p_i, \ldots, p_m\}$, where

$$p_i = \frac{occurrencenumber\ inthe\ i\text{-}th\ \mathrm{section}}{total\ occurrence\ nunber\ in\ thedocument}$$

The entity distribution feature value is then computed by $entropy = -\sum_{i=1}^{m} p_i \log p_i$. In our experiments, we simply set $m = 5$.

In the experiments, we used the SVMLight toolkit for classification. We collected 1000 news articles from a popular Chinese news portal - news.sohu.com. Each news article was manually annotated with its focused named entities, and there were totally 1447 focused entities (i.e. 1.4 focused entities per article). We performed 5-fold cross-validation on the dataset and the mean F-measure was 81%.

After we use the classification model to classify all the named entities in the news article into focused entities or not, we do not directly use the classified focused entities, because the number of focused entities is very small. Instead, in order to leverage more useful entities for feature generation, we select the top *K* percent named entities, which are the most confidently classified focused entities in the news article, as useful name entities. *K* is a parameter in our study. We use the output value of the SVM classification model to indicate the classification confidence level.

3.3.3 Related NE Expansion

The above two schemes find useful entities only from the particular news article. However, according to our observation, some named entities in the news comments do not appear in the particular news article at all, but they are closely related to some entities in the news article. This phenomenon is called "topic shifting". For example, when a news article is talking about "中国联通"\"China Unicom", related entities such as "中国移动"\"China Mobile", "中国电信"\"China Telecom", which do not appear in the news article, may be talked about in the associated comments. There related entities are also very useful clues and thus we develop a related NE expansion tool to discover the related entities by using web mining techniques. We use the focused name entities extracted from each news article in Section 3.3.2 as seeds and use our tool to discover related entities for each focused entity. Finally, we use these entities as useful entities for feature generation.

There have been a few researches (Ohshima et al., 2006; Wang and Cohen, 2007, Vyas and Pantel, 2009) related to named entity expansion. One of the most famous online services is *Google Sets*[2]. Motivated by these related researches, our tool consists of the following two key steps for NE expansion of each single focused entity.

1) Given a focused entity *e*, we first submit four queries ["*e* 和"] / ["e and"], ["和 *e*"] / ["and e"], ["*e* 比"] / ["e than"] and ["比 *e*"] / ["than e"] to the Google web search engine and get the top 100 results for each query[3]. Then we split the snippets in the search result into sentences. The character sequences that occur both immediately before the two queries (["和 *e*"] and ["比 *e*"]) and immediately after the two queries (["*e* 和"] and ["*e* 比"]) are extracted as initial candidates, and they are ranked by the geometric mean of the times each one appears immediately before the two queries (["和 *e*"] and ["比 *e*"]) and the times each one appears immediately after the two queries (["*e* 和"] and ["*e* 比"]). The top five candidates with high ranks are selected as the initial expansion results. In this step, we emphasize more on the precision of the candidates by using

[2] http://labs.google.com/sets

[3] The string in [ ] is the complete query string. Note that quotation marks ("") are used in each query string to guarantee that the query characters appear consecutively in the results.

only two strong indicator words “和” / “and” and “比” / “than”.

2) For each candidate *e′* obtained in step 1), we submit a query [“*e*” “*e′*”] to the Google search engine and obtain the top 100 results and the corresponding whole web pages[4]. Similar to Wang and Cohen (2007), in each semi-structured web page, we find the common HTML contexts of the two entities *e* and *e′* as wrappers and use these wrappers to extract more candidates from the page. A graph *G* = <*V*, *E*> is built, where *V* includes all wrappers and candidates and *E* = {(*w*, *c*)| if candidate *c* is extracted by wrapper *w*}. The weight for each edge in *E* was set to 1. A random graph walk with restart (RWR) is then applied to the graph to score the candidates. Finally, the top ranked candidates whose normalized scores are greater than 0.05 are selected as the expansion results. The precision value can reach 75% based on analysis of the expansion results for five focused entities.

The example expansion results by using our tool and *Google Sets* are shown in Table 4 when the seed entity is “休斯顿火箭”\“Houston Rocket”.

Query: 休斯顿火箭(Houston Rocket)	
Google Sets	Our tool
火箭季后赛 (Rockets Playoffs) 火箭 vs (Rockets vs) 火箭迷来踩 (Rockets fans come to see) 火箭的 (Rockets’) 仓储管理 (Storage Management) 迈阿密热火 (Miami Heat) ……	洛杉矶湖人 (LA Lakers) 圣安东尼奥马刺 (San Antonio Spurs) 达拉斯小牛 (Dallas Mavericks) 波士顿凯尔特人 (Boston Celtics) 底特律活塞 (Detroit Pistons) 丹佛掘金 (Denver Nuggets) 芝加哥公牛 (Chicago Bulls) 菲尼克斯太阳 (Phoenix Suns) 奥兰多魔术 (Orlando Magic) 波特兰开拓者 (Portland Trail Blazers) 金州勇士 (Golden State Warriors) 犹他爵士 (Utah Jazz) ……

Table 4: Related NE expansion results

3.4 Dictionary-based Correction

In this section, we present a dictionary-based correction step to address the following two issues:

1) As compared with named entities in news articles, a few named entities in news comments may have different spellings with the same or similar pronunciations, e.g., “谢亚龙” / “Xieyalong” in a news article and “谢鸭龙” / “Xieyalong” in a news comment, “刘翔” / “Liuxiang” in a news article and “刘降” / “Liuxiang” in a news comment.

2) As compared with named entities in news articles, a few named entities in news comments may be replaced with some concise English expressions, e.g., “易趣” / “Yiqu” in a news article and “ebay” in a news comment, “周杰伦” / “Zhoujielun” in a news article and “JAY” in a news comment.

For addressing the first case, we use a Chinese Pinyin dictionary to correct the results. If a Chinese character sequence in a news comment has the same pronunciation as some named entity in the news article, the character sequence is tagged as a named entity with the same type. For example, the character sequence “谢鸭龙” / “Xieyalong” in a news comment has the same pronunciation as the person name “谢亚龙” / “Xieyalong” in the news article, and thus we tag “谢鸭龙” / “Xieyalong” as a person name in our correction step.

For addressing the second case, we use an online English-Chinese bilingual dictionary (http://dict.youdao.com) for correction. The online dictionary can return the translations for most new words, such as “Jay”, “ebay”, and such translations can not be found in traditional dictionaries. For each sequence of continuous non-Chinese characters, we submit the string to the online dictionary, and a list of Chinese translations is returned. We then compare the translations and the entities in the news article, and if a match is found, the correction is performed.

4 Experiment and Analysis

4.1 Experiment Setup

There are no public benchmark datasets for evaluation of named entity recognition in Chinese news comments. Therefore, we manually labeled our dataset for evaluation. We downloaded 60 pieces of news and their associated comments from a popular Chinese news portal – www.sina.com.cn in October, 2008. They belonged to five different domains: politics, economics, sports, entertainment and technology. For each piece of news, we selected the first 100 comments. We then manually annotated the named entities (person name, location name and organization name) in the comments. Two annotators were employed and the conflicting annotations were resolved by discussion. Figure 2 gives

[4] Note that quotation marks (“”) are used for each entity (*e*, *e′*), but not for the whole query string.

two examples in our dataset and Table 5 shows the entity distribution in our dataset. The dataset will be freely downloaded from our website.

①全国绝大多数城市房价暴涨的时候，为什么<ne type="org" ref="中国房地产协会">**中房协**</ne>不站出来建议<ne type="org">**国务院**</ne>调整房地产调控政策？？？
When housing price kept soaring in most cities, why <ne type=”org” ref=”China Real Estate Association”> **zhongfangxie** </ne> didn’t stand out to suggest <netype=”org”> **the State Department** </ne> changing the tactics of adjustment and control of the real estate.
②<ne type="per" ref="周杰伦">**JAY**</ne> 是最棒的 ．永远支持<ne type="per" ref="周杰伦">**杰伦**</ne>!
<ne type="per" ref="zhoujielun"> **JAY** </ne> is the best one, we will stand for <ne type="per" ref="zhoujielun"> **jielun** </ne> forever.

Figure 2: Two examples in our dataset

	Per	Loc	Org	Total
Number	3295	3758	2780	9833

Table 5: Entity distribution in our dataset

In the experiments, we use 5-fold cross validation for evaluation. In each fold, 80% is used for training and the remaining 20% is used for testing. We use the standard F-measure for evaluation.

Finally, the performance values are averaged over the five folds.

4.2 Results and Discussions

Table 6 shows the comparison results for baselines and our proposed methods with different settings. Baseline1 directly uses a public available Chinese NER tool – S-MSRSeg[5] developed by MSR to tag the comments, and it is based on the linear mixture model framework. Baseline2 directly uses our in-house NER tool to tag the comments. Baseline3 uses only the standard features in the CRFs model, which is trained on the comments data via cross-validation, and it does not make use of any entity information in the news article. Different settings are investigated in our proposed method when K is simply set to 50.

We can see that Baseline1 and Baseline2 do not perform well, because the two baselines are developed for NER in formal news texts. Though Baseline3 does not use complex features, it performs much better than the first two baselines, which demonstrates the big difference between news texts and news comments.

The use of the entity information in the referred news article can much improve the performance, especially for person names and organization names. All the three schemes for finding useful entities in the news article are helpful to the task. In particular, the use of focused entities in the news article (Baseline3 + FocusedNE) can much outperform Baseline3 and the method using all named entities (Baseline3 + All NE) for person name recognition and organization name recognition. The results show that not all the named entities in the news article can provide important clues for NER in the news comments, and using all entities’ information may cause some noises. We give an example in the news of “SPORTS_12” in our dataset. “SPORTS_12” talks about Liu Xiang withdrawing from the Olympics. There are totally five different named entities in the news article. Our focused NE extraction model marks “刘翔” / “Liuxiang” and “北京” / “Beijing” as focused NEs when the parameter K = 50. In the associated comments, there are totally 216 NEs. Among them, 112 NEs refer to the two focused NEs, and less than 10 NEs in the comments refer to the other three NEs in the news article. Using these three NEs’ information for NER in comments may cause some noises.

Furthermore, the use of related named entities (Baseline3 + FocusedNE + RelatedNE) can further improve the performance. The F-measure for organization name recognition receives an improvement of 2.3%, while the F-measures for person and location recognition do not change significantly. This is because our related NE expansion tool works very well with an organization name as input. But when the input entity is a person or location name, a few of the expansion results are not named entities, which may introduce many noises to the CRFs model.

Lastly, we can see that the correction step can improve the performance for person name recognition and organization name recognition. Overall, the use of focused entities and related entities as useful entities, together with the correction step, can achieve the best performance in our experiments. The performance for location name recognition cannot be improved very much because the number of location name variants in the news comments is very limited.

[5] The tool can be downloaded from http://research.microsoft.com/en-us/um/people/jfgao/

	Per (%)	Loc (%)	Org (%)
Baseline1 (S-MSRSeg)	60.93	86.13	26.26
Baseline2 (Our In-house NER Tool)	73.61	80.75	37.28
Baseline3 (CRF with Standard Features)	73.73	88.92	59.63
Baseline3 + All NE	74.78	89.40	61.18
Baseline3 + Focused NE	80.53	90.12	65.02
Baseline3 + Focused NE + Related NE	80.80	90.15	67.31
Baseline3 + Focused NE + Correction	82.83	90.30	68.77
Baseline3 + Focused NE + Related NE + Correction	**83.06**	**90.32**	**70.87**

Table 6: Experimental results (F-measure)

In order to better understand the contribution of the focused named entity extraction step, we show the experiment results for the overall method (Baseline3 + FocusedNE + RelatedNE + Correction) with different values for the parameter *K* in Table 7. *K* is varying from 0 to 100 with a step size of 25. *K*=0 means that no name entity information in the news article is used, and *K* = 100 means that all the name entities in the news article are considered. We can see that in our dataset when *K* is set to a number around 50 (between 25 and 75), the overall performance does not change much. Using no entities (*K*=0) and using all entities (*K*=100) will much lower the overall performance, which demonstrates that it is important to leverage appropriate named entities for feature generation in our proposed method.

K	Per (%)	Loc (%)	Org (%)
K = 0	76.71	89.16	65.37
K = 25	81.57	90.21	71.38
K = 50	83.06	90.32	70.87
K = 75	81.16	89.85	69.53
K =100	77.31	89.56	67.59

Table 7: Overall results (F-measure) vs. *K*

5 Conclusion and Future Work

In this paper, we propose to leverage the entity information in the referred news article to improve named entity recognition in the news comments. Three schemes for finding useful entities are presented. Experimental results demonstrate the effectiveness of each component in our proposed method.

In future work, we will explore new features based on the relationships between news article and news comments, and the relationships between news comments. We will also address the co-reference resolution task in news comments.

Acknowledgments

This work was supported by NSFC (60873155), Beijing Nova Program (2008B03) and NCET (NCET-08-0006). We thank the anonymous reviewers for their useful comments.

References

O. Bender, F. J. Och and H. Ney. 2003. Maximum entropy models for named entity recognition. In CoNLL.

D. M. Bikel, R. L. Schwartz, and R. M. Weischedel. 1999. An algorithm that learns what's in a name. In Machine Learning, volume 34, pages 211–231.

A. Chen, F. Peng, R. Shan, and G. Sun. 2006. Chinese named entity recognition with conditional probabilistic models. In 5th SIGHAN Workshop on Chinese Language Processing.

L. Chiticariu, R. Krishnamurthy, Y. Li, F. Reiss and S. Vaithyanathan. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In Proceedings of EMNLP2010.

O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S.Weld, and A. Yates. 2005. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. Artif. Intell., 165 (2005) (1), pp. 91–134.

X. Fang and H. Sheng. 2002. A hybrid approach for Chinese named entity recognition. In Proceedings of Discovery Science'02.

J. R. Finkel and C. D. Manning. 2009. Nested named entity recognition. In EMNLP.

R. Florian, A. Ittycheriah, H. Jing, and T. Zhang. 2003. Named entity recognition through classifier combination. In CoNLL.

G. Fu and K.-K. Luke. 2005. Chinese named entity recognition using lexicalized HMMs. ACM SIGKDD Explorations Newsletter – Natural Language Processing and Text Mining, 7(1).

J. Gao, M. Li, A. Wu, and C.-N. Huang. 2005. Chinese Word segmentation and named entity recognition: a pragmatic approach. Computational Linguistics, 31(4).

D. Gruhl, M. Nagarajan, J. Pieper, C. Robson, and A. Sheth. 2009. Context and domain knowledge enhanced entity spotting in informal text. In ISWC.

J. Huang, G. Zweig, and M. Padmanabhan. 2001. Information extraction from voicemail. In ACL.

M. Jansche and S. Abney. 2002. Information extractionfrom voicemail transcripts. In EMNLP.

J. Kazama and K. Torisawa 2007. Exploiting-Wikipedia as External Knowledge for Named Entity Recognition. Proceedings of EMNLP2007.

A. Klementiev and D. Roth. 2006.Weakly Supervised Named Entity Transliteration and Discovery from Multilingual Comparable Corpora. In Proceeding of ACL2006.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML01.

J. Mayfield, P. McNamee and C. Piatko. Named Entity Recognition using Hundreds of Thousands of Features. In: Proceedings of CoNLL-2003.

A. McCallum and W. Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In CoNLL.

E. Minkov, R. C. Wang, and W. W. Cohen. 2005. Extracting personal names from emails: Applying named entity recognition to informal text. In HLT/EMNLP.

H. Ohshima, S. Oyama, and K. Tanaka. 2006. Searching Coordinate Terms with Their Context from the Web. Proceeding of WISE2006.

D. Okanohara, Y. Miyao, Y. Tsuruoka; J. Tsujii. 2006. Improving the Scalability of Semi-Markov Conditional Random Fields for Named Entity Recognition. In Proceeding of ACL2006.

M. Pennacchiotti and P. Pantel. 2009. Entity extraction via ensemble semantics. In Proceedings of EMNLP2009.

A. E. Richman and P. Schone 2008. Mining Wiki Resources for Multilingual Named Entity Recognition. In Proceeding of ACL2008.

S. Sekine and C. Nobata. 2004. Definition, dictionaries and tagger for extended named entity hierarchy. In Proceedings of Conference on Language Resources and Evaluation.

S. Singh, D. Hillard, and C. Leggeteer. 2010. Minimallysupervised extraction of entities from text advertisements. In NAACL-HLT.

R. Srihari, C. Niu, and W. Li. 2001. A hybrid approachfor named entity and sub-type tagging. In ANLP.

J. Sun, J. Gao, L. Zhang, M, Zhou, and C. Huang. 2002. Chinese named entity identification using class-based language model. In Proceedings of COLING2002.

R. C. Wang, and W. W. Cohen. 2007. Language-Independent Set Expansion of Named Entities using the Web. Proceeding of ICDM2007.

V. Vyas, P. Pantel. 2009. Semi-automatic entity set refinement. Proceedings of HLT-NAACL2009.

Y. Wu, J. Zhao, B. Xu, and H. Yu. 2005. Chinese named entity recognition based on multiple features. In Proceedings of HLT-EMNLP2005.

X. Yu, W. Lam, S.-K. Chan, Y. K. Wu, B. Chen. 2008. Chinese NER Using CRFs and Logic for the Fourth SIGHAN Bakeoff. The Sixth SIGHAN Workshop on Chinese Language Processing.

L. Zhang, Y. Pan, T. Zhang. 2004. Focused named entity recognition using machine learning. In Proceedings of SIGIR2004.

H. Zhao and C. Kit, 2008. Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. The Sixth SIGHAN Workshop on Chinese Language Processing.

J. Zhou, L. He, X. Dai, and J. Chen. 2006. Chinese named entity recognition with a multiphase model. In 5th SIGHAN Workshop on Chinese Language Processing.

X. Zhu, M. Li, J. Gao and C.-N. Huang. 2003. Single character Chinese named entity recognition. In Proceedings of SIGHAN2003.