

# Using a Smoothing Maximum Entropy Model for Chinese Nominal Entity Tagging

**Jinying Chen**

Department of Computer and  
Information Science

University of Pennsylvania  
Philadelphia, PA, 19104

jinying@gradient.ci  
s.upenn.edu

**Nianwen Xue**

Department of Computer and  
Information Science

University of Pennsylvania  
Philadelphia, PA, 19104

xueniwen@linc.cis.up  
enn.edu

**Martha Palmer**

Department of Computer and  
Information Science

University of Pennsylvania  
Philadelphia, PA, 19104

mpalmer@linc.cis.upe  
nn.edu

## Abstract

This paper treats nominal entity tagging as a six-way (five categories plus nonentity) classification problem and applies a smoothing maximum entropy (ME) model with a Gaussian prior to the Chinese nominal entity tagging task. The experimental results show that the model performs consistently better than a ME model using a simple counting cut-off. The results also suggest that simple semantic features extracted from an electronic dictionary improve the model's performance, especially when the training data is insufficient.

## 1 Introduction

Nominal entity tagging refers to the detection and classification of nominal entity mentions in textual data. The task is necessary to support higher-level NLP tasks such as co-reference resolution and relation detection, which is crucial for Information Extraction, Question-Answering and other NLP applications.

This paper treats nominal entity tagging as a classification problem and describes the results of applying a smoothing maximum entropy (ME) model with a Gaussian prior (Chen and Rosenfeld, 1999) to the Chinese nominal entity tagging task.

The experimental results show that the model performs consistently better than a ME model using a simple counting cut-off (Ratnaparkhi, 1998). The results also suggest that simple semantic features extracted from an electronic dictionary improve the model's performance, especially when there is a severe sparse data problem.

In the next two sections, we briefly introduce the task of nominal entity tagging and the smoothing ME model with a Gaussian prior respectively. In Section 4, we describe the features we used in this task. We report our experimental results in Section 5 and conclude our discussion in Section 6.

## 2 Nominal Entity Tagging

It is generally the case that named entity mentions, nominal entity mentions and pronouns that refer to the same entity alternate in discourse. Nominal entity tagging, i.e., detecting and classifying the nominal entity mentions, is among the first steps towards resolving the co-reference among different mentions of the same entity. Therefore, it is a crucial component to any Information Extraction or Question-Answering task that not only extracts entity mentions but also determines which entity mentions are co-referential and the relations between the entities.

Nominal entity tagging is a difficult problem because there is ambiguity involved in both nominal detection and classification. First, the same string can either be a nominal entity<sup>1</sup> or a non-nominal entity depending on the context. For example, “酒店/restaurant” is a nominal entity in “这/this 家/CL 酒店/restaurant”, but it is part of a named entity in “五洲/Wuzhou 大/great 酒店/restaurant”. In the classification task, the same string can belong to different categories based on the context. For example, “中心/center” is a facility in “康复/rehabilitation 中心/center” and a location in “广场/square 中心/center”, depending on the local context of the current word (i.e., its pre-modifier in this case). A harder example involving global contextual information is shown in (1). Here, “港口/port” is a facility in (1a) but a geo-political entity in (1b).

- (1) a. 广州/Guangzhou 的/DE 集装箱/container 港口/port, 去年/last year 前 9 个月/the first nine months 的/DE 吞吐量/throughput 排名/rank 第 5 位/5th

“The throughput of the container port in Guangzhou for the first 9 months of last year 5th.”

- b. 上海/Shanghai 仍然/still 是/is 中国/China 第一大/largest 的/DE 集装箱/container 港口/port.

“Shanghai is still the largest container port in China.”

Clearly, resolving these ambiguities is crucial to successful nominal entity detection and classification. Towards this end, we used a machine-learning approach, namely, a smoothing Maximum Entropy (ME) model with a Gaussian prior to solve this problem. Instead of treating detection and classification as two separate stages, we added a “nonentity” category to the five categories defined by the LDC (Mitchell and Huang, 2003) and recast nominal entity tagging as a six-way classification problem. The five categories are Person (PER), Organization (ORG),

Facility (FAC), Location (LOC) and Geo-political Entity<sup>2</sup> (GPE).

The LDC standard defines nominal entities over words as well as the phrases of which they are heads. The words are called **head words** and the phrases are called the **extents**. For example, in (1b), the GPE mention here includes both the head word “港口/port” and its extent “中国/China 第一大/largest 的/DE 集装箱/container 港口/port”.

In our experiments, we segmented, POS-tagged and parsed the input text with a ME Segmenter (Xue, 2003), a ME POS-tagger (Ratnaparkhi, 1996) and a Chinese generative parser (Bikel, 2002). Since finding the head words is the most important part of the nominal entity tagging task, our focus in this paper is on detecting and classifying nominal words. We first used some heuristics to find the candidates. Briefly speaking, the candidates include all the common nouns as well as the words occurring in a nominal entity list extracted from the training data. We then ran the maximum entropy classifiers to decide which ones are nominal entities and what their categories are.

### 3 A Smoothing Maximum Entropy Model with a Gaussian Prior

Maximum Entropy (ME) modeling is a statistical approach that can be applied to any classification task (Berger et al., 1996). The approach has been used to solve a wide range of NLP tasks, such as prepositional phrase attachment classification (Ratnaparkhi et al. 1994), part-of-speech tagging and parsing (Ratnaparkhi, 1996; Ratnaparkhi, 1998) and word sense disambiguation (Palmer et al., to appear) etc. A ME model combines evidence from different sources (features) without the independence assumption on its features.

In the nominal entity classification task, the ME model produces a probability for each category (PER, ORG, nonentity etc.) of a nominal entity candidate conditioned on the context in which the candidate occurs. The conditional probability is calculated by Equation (1),

<sup>1</sup> In the rest of the paper, we do not explicitly distinguish between *nominal entities* and *nominal entity mentions*. The distinction can be made according to the context.

<sup>2</sup> GPE entities are geographical regions defined by political and/or social groups. A GPE entity subsumes and does not distinguish between a nation, its region, its government, or its people, e.g., 政府/government, 人民/people etc.

$$p(l|c) = \frac{1}{Z(c)} \exp\left(\sum_{j=1}^k \lambda_j f_j(l, c)\right) \quad (1)$$

where  $c$  represents the context containing the nominal candidate and  $l$  is the label for each entity category.  $Z(c)$  is a normalization factor.

$f_j(l, c)$  represents the  $j$ th feature for the candidate and  $k$  is the total number of features used by the model. The features used in our model are all binary-valued feature functions (or indicator functions). A typical feature is shown in equation (2),

$$f_j(GPE, lastChar = \text{国}) = \begin{cases} 1 & \text{iff the candidate is labeled as} \\ & GPE \text{ \& its last character} = \text{国} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Trained with labeled training data, a ME model can be regarded as a maximum likelihood model (Berger et al., 1996). Like other maximum likelihood models, the ME model can suffer from overfitting. Several smoothing methods can be applied for ME models. Empirical evidence has shown that the smoothing ME model with a Gaussian prior performs better than the ME models with other smoothing methods and the basic idea of this approach is to use a zero-mean Gaussian prior with a diagonal covariance matrix to calculate the prior probability of the model (Chen and Rosenfeld, 1999).

In our experiments, we used Mallet, a Machine Learning for Language Toolkit that implements a smoothing ME model with a Gaussian prior<sup>3</sup> (McCallum, 2002).

## 4 Features Used in the Model

The features used in our model include:

### 1. Simple features

- a. the current word  $w_0$ , and its part of speech tag  $p_0$

<sup>3</sup> We choose the default value provided by Mallet, 1, as the value of the variance parameters after examining different values ranging from 0.1 to 10 with step value 0.1 by 10-fold cross-validation on the training data.

- b. the words at positions  $-1$  and  $+1$ ,  $w_{-1}$  and  $w_{+1}$ , relative to  $w_0$  and their part of speech tags  $p_{-1}$  and  $p_{+1}$ .
- c. collocation features  $p_{-1}w_0$ ,  $w_0p_{+1}$ .
- d. the last character  $lc$  of  $w_0$

### 2. Syntactic features

- a. the word  $w'_{-1}$  preceding the minimal NP containing  $w_0$ , and its part of speech tag  $p'_{-1}$ .
- b. the word  $w'_{+1}$  following the minimal NP containing  $w_0$  and its part of speech tag  $p'_{+1}$ .

### 3. Semantic features

- a. the semantic category  $s_0$  of  $w_0$  extracted from the Rocling dictionary, an electronic Chinese dictionary that assigns semantic categories to common nouns, such as *building* for “仓库/warehouse”.
- b. collocation features  $w_{-1}s_0$ ,  $p_{-1}s_0$ ,  $s_0w_{+1}$ ,  $s_0p_{+1}$ .

The first three types of simple features (1a, b, c) are used to capture the local contextual information about the current word. The features in (1d) are used to capture a particular property of many Chinese nouns, i.e., the last characters of many Chinese nouns indicate their semantic categories. For example, “者/suffix” usually indicates the category of person, as in “记者/reporter”, “作者/writer” and “读者/reader” etc.

The simple syntactic features are mainly used to capture the information contained in the pre-modifier, post-modifier of the NP containing the current word, as well as information from the main verb (with this NP as a subject or object) and punctuation information.

The simple semantic features are used to capture the semantic category of the current word. In our model, we use the current word and the surrounding words as lexical features. However, a model using lexical information only generally suffers from the problem of sparse data. In this case, the model needs to back off to more general features. The semantic features extracted from the Rocling dictionary and used here are more general

	The Smoothing ME Model with a Gaussian Prior (%)			The ME Model with Cut-off Threshold (%)		
	Precision	Recall	F	Precision	Recall	F
<b>PER</b>	<b>88.13</b>	<b>72.26</b>	<b>79.41</b>	<b>89.38</b>	<b>60.97</b>	<b>72.49</b>
<b>GPE</b>	<b>80.47</b>	<b>67.95</b>	<b>73.68</b>	<b>74.76</b>	<b>61.54</b>	<b>67.51</b>
<b>ORG</b>	<b>69.83</b>	<b>45.83</b>	<b>55.34</b>	<b>64.98</b>	<b>33.50</b>	<b>44.21</b>
<b>LOC</b>	<b>70.10</b>	<b>46.13</b>	<b>55.64</b>	<b>67.96</b>	<b>22.58</b>	<b>33.90</b>
<b>FAC</b>	<b>66.88</b>	<b>28.31</b>	<b>39.78</b>	<b>68.21</b>	<b>27.25</b>	<b>38.94</b>

Table 1 The performance of the two ME models on the evaluation data set

	All Features (%)			W/o Semantic Features (%)		
	Precision	Recall	F	Precision	Recall	F
<b>PER</b>	<b>88.13</b>	<b>72.26</b>	<b>79.41</b>	<b>89.35</b>	<b>67.95</b>	<b>77.19</b>
<b>GPE</b>	<b>80.47</b>	<b>67.95</b>	<b>73.68</b>	<b>82.14</b>	<b>61.14</b>	<b>70.10</b>
<b>ORG</b>	<b>69.83</b>	<b>45.83</b>	<b>55.34</b>	<b>75.00</b>	<b>40.35</b>	<b>52.47</b>
<b>LOC</b>	<b>70.10</b>	<b>46.13</b>	<b>55.64</b>	<b>70.40</b>	<b>28.39</b>	<b>40.46</b>
<b>FAC</b>	<b>66.88</b>	<b>28.31</b>	<b>39.78</b>	<b>79.10</b>	<b>14.02</b>	<b>23.82</b>

Table 2 The performance of the smoothing ME model with a Gaussian prior with or without semantic features on the evaluation data set

than lexical features but still contain rich information about the word. Intuitively, this information should be very useful for deciding the category of a nominal entity. For example, a common noun representing a person usually has the semantic category *mankind*, and a word with the semantic category *region* is very likely to be a location. The Rocling dictionary contains 4474 entries and 110 semantic categories. It assigns one category to each Chinese noun without considering sense ambiguity. A rough count shows that about 30% of the candidates in the training and test data can be found in this dictionary.

## 5 Experimental Results

In our experiments, we trained and tested the model on the data set provided by LDC for the Automatic Content Extraction (ACE) research program, which contains a 110K-hanzi (Chinese characters) training data set and a 135K-hanzi evaluation data set. We compared our model's performance on the evaluation data set with a simple smoothing ME model that discards features occurring no more than 5 times in the training corpus. The experimental results are shown in

Table 1. The precision is the number of correctly tagged nominal entities divided by the total number of nominal entities predicted by the nominal tagger for each category. The recall is the number of correctly tagged nominal entities divided by the total number of nominal entities in the evaluation data for each category. The scores are calculated based on the official ACE evaluation metric. According to this metric, a predicted nominal entity is correct when the overlap of its head and the head of the gold-standard nominal entity is over 1/3. The evaluation is performed on head words rather than the full extent of the nominal phrase since this is considered to be the most important by this metric. Notice that the scores are calculated only for the five nominal categories. That is, nonentities are not considered.

The results in Table 1 clearly show that the ME model with a Gaussian prior is better than the ME model with a simple counting cut-off, with a 6~7 percentage improvement for PER and GPE and a 10~20 percentage improvement for ORG and LOC.

We also investigated the effect of using simple semantic features discussed in Section 4. The experimental results, given in Table 2, indicate that

	All Features (%)			W/o Semantic Features (%)		
	Precision	Recall	F	Precision	Recall	F
<b>PER</b>	<b>88.83</b>	<b>77.55</b>	<b>82.81</b>	<b>89.82</b>	<b>75.68</b>	<b>82.15</b>
<b>GPE</b>	<b>85.24</b>	<b>73.00</b>	<b>78.65</b>	<b>86.41</b>	<b>72.11</b>	<b>78.62</b>
<b>ORG</b>	<b>68.65</b>	<b>48.46</b>	<b>56.81</b>	<b>68.91</b>	<b>45.94</b>	<b>55.13</b>
<b>LOC</b>	<b>77.54</b>	<b>63.69</b>	<b>69.93</b>	<b>79.65</b>	<b>53.57</b>	<b>64.06</b>
<b>FAC</b>	<b>77.57</b>	<b>36.73</b>	<b>49.85</b>	<b>88.46</b>	<b>30.53</b>	<b>45.39</b>

Table 3 The performance of the ME model with a Gaussian Prior with or without semantic features trained on the enlarged training set

	The Smoothing ME Model with a Gaussian Prior (%)			The ME Model with Cut-off Threshold (%)		
	Precision	Recall	F	Precision	Recall	F
<b>PER</b>	<b>88.83</b>	<b>77.55</b>	<b>82.81</b>	<b>91.31</b>	<b>65.32</b>	<b>76.16</b>
<b>GPE</b>	<b>85.24</b>	<b>73.00</b>	<b>78.65</b>	<b>83.45</b>	<b>65.72</b>	<b>73.53</b>
<b>ORG</b>	<b>68.65</b>	<b>48.46</b>	<b>56.81</b>	<b>66.17</b>	<b>37.25</b>	<b>47.67</b>
<b>LOC</b>	<b>77.54</b>	<b>63.69</b>	<b>69.93</b>	<b>81.13</b>	<b>51.19</b>	<b>62.77</b>
<b>FAC</b>	<b>77.57</b>	<b>36.73</b>	<b>49.85</b>	<b>80.72</b>	<b>29.65</b>	<b>43.37</b>

Table 4 The performance of the two ME models with semantic features trained on the enlarged training set

the semantic features do improve the performance significantly, boosting the performance by 2~3 percent for PER, GPE and ORG and over 15 percent for LOC and FAC.

The relatively low performance of the model on the last three categories (ORG, LOC, FAC) suggests that the training data may be insufficient. We did a rough count of the nominal entities in the training and evaluation data sets and found that the evaluation data set contains more entities than the training set. Furthermore, there are very few location and facility entities (fewer than 250) in the training set. Thus enlarging the training set should improve the model’s performance further, especially for the last two categories.

To verify this, we randomly divided the evaluation data into two sets that have roughly equal number of training files. One set is added to the official training data and the other set is used for testing. Now the training set contains roughly twice as many nominal entities as the test set. We retrained the ME model with a Gaussian prior and the results are shown in Table 3. In this table, we also give the model’s performance with different feature sets (with or without semantic features). As

we can see, the performance improves for all five categories and the improvement for the last two categories (LOC and FAC) are especially significant. Furthermore, the semantic features improve the model’s performance consistently, although the improvement is not as dramatic as in the previous experiments with the smaller training set. This suggests that simple semantic features are especially useful when the training data is insufficient.

Finally, we give the performance of the two ME models trained on the enlarged training set in Table 4. The results show that the smoothing ME model with a Gaussian prior outperforms the simple smoothing ME model significantly, consistent with results in the previous experiments.

An analysis of the tagging output shows that there are two major error types. The first error type comes from the preprocessing. In particular, segmentation errors often prevent the nominal tagger from finding the correct annotation candidates. For example, the nominal entity “发展中/developing 国家/country” was segmented as “发展/develop 中国/China 家/home”. As a result, there is no easy way for the nominal tagger to get

the correct candidate “国家/country”, given that our tagger is word-based and relies on the segmentation output for finding the tagging candidates. The other major error type is due to the inability of the current model to capture global contextual information. For example, the word “港口/port” in the sentence shown in (1b) in section 2, repeated here as (2):

(2) 上海/Shanghai 仍然/still 是/is 中国/China 第一大/largest 的/DE 集装箱/container 港口/port.

“Shanghai is still the largest container port in China.”

One useful piece of information that can help decide that “港口/port” is a GPE entity rather than a facility (as in (1a)) is the fact that “港口/port” refers to the same entity as “上海/Shanghai”, a city in China. The features implemented in this model cannot capture such information.

## 6 Conclusion and Future Work

We have shown that a smoothing ME model with a Gaussian prior outperforms a simple smoothing ME model with a cut-off threshold in the Chinese nominal tagging task. The better performance remains consistent across training sets of different sizes, and the different feature sets (with or without semantic features). We further showed that simple semantic features improve the model’s performance, especially when the training data is insufficient. Given additional data, we expect that further improvement is possible.

For future work, we will explore more complicated features that encode co-reference. In addition, predicate-argument structure information might be also helpful. For example, knowing the semantic role a noun plays with regard to a verb might help determine the nominal category of a noun. We will explore ways to extract such rich linguistic features using automatic methods (Gildea and Jurafsky, 2002; Gildea and Palmer, 2002) and embed them into our model in a robust way in future work.

## References

- Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1): 39-71.
- Daniel M. Bikel. 2002. Design of a Multi-lingual, Parallel-processing Statistical Parsing Engine. In *Proceedings of HLT 2002*, pages 24-27, San Diego, CA, March.
- Stanley. F. Chen and Ronald Rosenfeld. 1999. A Gaussian Prior for Smoothing Maximum Entropy Models. *Technical Report CMU-CS-99-108*, CMU.
- Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics* 28:3, 245-288.
- Daniel Gildea and Martha Palmer, The Necessity of Parsing for Predicate Argument Recognition, In *Proceedings of the 40<sup>th</sup> Meeting of the Association for Computational Linguistics, ACL-02*, Philadelphia, PA, July 7-12, 2002.
- Andrew K. McCallum. 2002. "MALLET: A Machine Learning for Language Toolkit." <http://www.cs.umass.edu/~mccallum/mallet>.
- Alexis Mitchell and Shudong Huang. 2003. *Entity Detection and Tracking - Phase1: EDT and Metonymy Annotation Guidelines*, Version 2.5.
- Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*. To appear.
- Adwait Ratnaparkhi, Jeff Reynar, and Salim Roukos. 1994. A Maximum Entropy Model for Prepositional Phrase Attachment. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 250-255.
- Adwait Ratnaparkhi. 1996. A Maximum Entropy Part-Of-Speech Tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, 1996, University of Pennsylvania.
- Adwait Ratnaparkhi. 1998. Maximum Entropy Models for Natural Language Ambiguity Resolution. *Ph.D. thesis*, University of Pennsylvania.
- Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. *International Journal of Computational Linguistics and Chinese Language Processing*, 8:1:29-48.