A Robust Risk Minimization based Named Entity Recognition System

Tong Zhang
IBM T.J. Watson Research Center
Yorktown Heights
New York, 10598, USA
tzhang@watson.ibm.com

David Johnson
IBM T.J. Watson Research Center
Yorktown Heights
New York, 10598, USA
dejohns@us.ibm.com

Abstract

This paper describes a robust linear classification system for Named Entity Recognition. A similar system has been applied to the CoNLL text chunking shared task with state of the art performance. By using different linguistic features, we can easily adapt this system to other token-based linguistic tagging problems. The main focus of the current paper is to investigate the impact of various local linguistic features for named entity recognition on the CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) shared task data. We show that the system performance can be enhanced significantly with some relative simple token-based features that are available for many languages. Although more sophisticated linguistic features will also be helpful, they provide much less improvement than might be expected.

1 Introduction

An important research area in the field of information extraction is Named Entity Recognition. This topic was a central theme in the message understanding conferences (MUCs). It has become more important nowadays due to the large amount of available electronic text, which makes it necessary to build systems that can automatically process and extract information from text.

In spite of significant work in this area, the problem itself has not been solved. Although some earlier reports suggested accuracy (F1-number) of machine learning based systems to be in the lower 90s with relatively small amount of labeled data (for example, (Bikel et al., 1999; Mikheev et al., 1998; Sundheim, 1995)), these studies were often performed on relatively restricted domains. Our experience indicates that the performance of a statistically based named entity extraction system can vary significantly depending on the underlying domain. There are still open challenges to make the performance of a statistical system consistent across different types of data sources.

In this paper we present a system for named entity recognition based on our earlier work on text chunking (Zhang et al., 2002). One advantage of the proposed system is that it can easily incorporate a large number of linguistic features. This advantage is similar to a number of other approaches, such as the maximum entropy method, which has been widely used to solve NLP problems, see (Borthwick, 1999; Ratnaparkhi, 1999) for example.

The performance of our system can be significantly affected by the choice of available linguistic features. The main focus of this paper is to investigate the impact of some local features. Specifically we show that the system performance can be enhanced significantly with some relatively simple token-based features. More sophisticated linguistic features, although helpful, yield much less improvement in system performance than might be expected.

We believe that this study provides useful insight into the usefulness of various available local linguistic features. Since these simple features are readily available for many languages, it suggests the possibility of setting up a language independent named entity recognition system quickly so that its performance is close to a system that uses much more sophisticated, language dependent features.

2 System description

Following the approach employed in our text chunking system (Zhang et al., 2002), we treat the named entity recognition problem as a sequential token-based tagging problem. We denote by $\{w_i\}$ $(i = 0, 1, \ldots, m)$ the sequence of tokenized text, which is the input to our system. In token-based tagging, the goal is to assign a class-label t_i, taking its value from a predefined set of labels, to every token w_i.

For named entity recognition, and text segmentation in general, the entities (segments) can be encoded as a token-based tagging problem by using various encoding schemes. In this paper, we shall only use the IOB1 encoding scheme which is provided in the CoNLL-2003 shared task.

The goal of our learning system is to predict the class-label value t_i associated with each token w_i. In our system, this is achieved by estimating the conditional probability $P(t_i = c|x_i)$ for every possible class-label value c, where x_i is a feature vector associated with token i. It is essentially a sufficient statistic in our model: we assume that $P(t_i = c|x_i) = P(t_i = c|\{w_i\}, \{t_j\}_{j \le i})$. The feature vector x_i can depend on previously predicted class-labels $\{t_j\}_{j \le i}$, but the dependency is typically assumed to be local. Given such a conditional probability model, in the decoding stage, we estimate the best possible sequence of t_i's using a dynamic programming approach, similar to what is described in (Zhang et al., 2002).

In our system, the conditional probability model has the following parametric form:

$$P(t_i = c|x_i, \{t_{i-\ell}, \dots, t_{i-1}\}) = T(w_c^T x_i + b_c),$$

where $T(y) = \min(1, \max(0, y))$ is the truncation of y into the interval $[0, 1]$. w_c is a linear weight vector and b_c is a constant. Parameters w_c and b_c can be estimated from the training data.

Given training data (x_i, t_i) for $i = 1, \dots, n$. It was shown in (Zhang et al., 2002) that such a model can be estimated by solving the following optimization problem for each c:

$$\inf_{w,b} \frac{1}{n} \sum_{i=1}^{n} f(w_c^T x_i + b_c, y_c^i),$$

where $y_c^i = 1$ when $t_i = c$ and $y_c^i = -1$ otherwise. The function f is defined as:

$$f(p, y) = \begin{cases} -2py & py < -1 \\ \frac{1}{2}(py - 1)^2 & py \in [-1, 1] \\ 0 & py > 1. \end{cases}$$

This risk function is closely related to Huber's loss function in robust estimation. We shall call a classification method that is based on approximately minimizing this risk function *robust risk minimization*. The generalized Winnow method in (Zhang et al., 2002) implements such a method. The numerical algorithm used for experiments in this paper is a variant, and is similar to the one given in (Damerau et al., 2003).

The main purpose of the paper is to investigate the impact of local linguistic features for the Named Entity detection task. The basic linguistic features considered here are all aligned with the tokens. Specifically we will consider features listed in Table 1. These features are represented using a binary encoding scheme where each component of the feature vector x corresponds to an occurrence of a feature that is listed above. We use a window of ± 1 centered at the current token unless indicated otherwise in Table 1.

3 Experimental Results

We study the performance of our system with different feature combinations on the English development set. Our results are presented in Table 2. All of these results are significantly better than the baseline performance of 71.18. We will now discuss the implications of these experimental results.

The small difference between Experiment 1 and Experiment 2 implies that tokens by themselves, whether represented as mixed case text or not, do not significantly affect the system performance.

Experiment 3 shows that even without case information, the performance of a statistical named entity recognition system can be greatly enhanced with token prefix and suffix information. Intuitively, such information allows us to build a character-based token-model which can predict whether an (unseen) English word looks like an entity-type or not. The performance of this experiment is comparable to that of the mixed-case English text plus capitalization feature reported in Experiment 4.

Experiment 4 suggests that capitalization is a very useful feature for mixed case text, and can greatly enhance the perform of a named entity recognition system. With token prefix and suffix information that incorporates a character-based entity model, the system performance is further enhanced, as reported in Experiment 5.

Up to Experiment 5, we have only used very simple token-based linguistic features. Despite their simplicity, these features give very significant performance enhancement. In addition, such features are readily available for many languages, implying that they can be used in a language independent statistical named entity recognition system.

In Experiment 6, we added the provided part-of-speech and chunking information. Clearly they only lead to a relatively small improvement. We believe that most information contained in part-of-speech has already been captured in the capitalization and prefix/suffix features. The chunking information might be more useful, though its value is still quite limited.

By adding the four supplied dictionaries, we observe a small, but statistically significant improvement. The performance is reported in Experiment 7. At this point we have only used information provided by the shared task.

Further performance enhancement can be achieved by using extra information that is not provided in the shared

Feature ID	Feature description
A	Tokens that are turned into all upper-case, in a window of ± 2.
B	Tokens themselves, in a window of ± 2.
C	The previous two predicted tags, and the conjunction of the previous tag and the current token.
D	Initial capitalization of tokens in a window of ± 2.
E	More elaborated word type information: initial capitalization, all capitalization, all digitals, or digitals containing punctuations.
F	Token prefix (length three and four), and token suffix (length from one to four).
G	POS tagged information provided in shared the task.
H	chunking information provided in the shared task: we use a bag-of-word representation of the chunk at the current token.
I	The four dictionaries provided in the shared task: PER, ORG, LOC, and MISC.
J	A number of additional dictionaries from different sources: some trigger words for ORG, PER, LOC; lists of location, person, and organizations.

Table 1: feature definition

Experiment ID	Features used	Precision	Recall	$F_{\beta=1}$
1	A+C	91.94	74.25	82.15
2	B+C	93.70	74.89	83.25
3	A+F	89.96	82.50	86.07
4	B+C+D	88.79	86.01	87.38
5	B+C+D+E+F	90.11	88.67	89.39
6	B+C+D+E+F+G+H	91.00	89.53	90.26
7	B+C+D+E+F+G+H+I	92.14	90.73	91.43
8	B+C+D+E+F+G+H+I+J	92.76	91.42	92.08

Table 2: Performance with different features on the English development set

task. In this study, we will only report performance with additional dictionaries we have gathered from various different sources. With these additional dictionaries, our system achieved a performance of 92, as reported in Experiment 8. Table 4 presents the performance of each entity type separately.

Clearly the construction of extra linguistic features is open ended. It is possible to improve system performance with additional and higher quality dictionaries. Although dictionaries are language dependent, they are often fairly readily available and providing them does not pose a major impediment to customizing a language independent system. However, for more difficult cases, it may be necessary to provide high precision, manually developed rules to capture particular linguistic patterns. Language dependent features of this kind are harder to develop than dictionaries and correspondingly pose a greater obstacle to customizing a language independent system. We have found that such features can appreciably improve the performance of our system, but discussion is beyond the scope of this paper. A related idea is to combine the outputs of different systems. See (Florian et al., 2003) for such a study. Fortunately, as our experiments indicate, special purpose patterns may not be necessary for quite reasonable accuracy.

In Table 4, we report the performance of our system on the German data. We shall note that the performance is significantly lower than the corresponding English performance. Our experience indicates that even for English, the real-world performance of a statistical named entity recognizer can be very low. The performance we reported for the German data is achieved by using the following features: B+C+D+E+F+G+H+I+J (with some small modifications), plus the German word lemma feature provided by the task.

The additional German dictionaries were provided to us by Radu Florian. Without these additional dictionaries (in this case, all information we use is provided by the CoNLL task), the overall performance is listed in Table 3. It is also interesting to note that without any dictionary information, the overall performance drops to an $F_{\beta=1}$ score of 65.5 on the development set, and to 70.2 on the test set. Clearly for this data, dictionary information helps more on the development data than on the test data.

	Precision	Recall	$F_{\beta=1}$
devel. set	82.49	61.22	70.29
test set	81.59	62.73	70.93

Table 3: System performance with only features that are provided by the CoNLL task on German data.

4 Conclusion

In this paper, we presented a general token-based NLP tagging system using a robust risk minimization classification method. The system can take advantage of different kinds of linguistic features.

We have studied the impact of various local linguistic features on the performance of our system. It is interesting to note that most performance improvement can be achieved with some relatively simple token features that are easy to construct. Although more sophisticated linguistic features will also be helpful, they provide much less improvement than might be expected. This observation supports the view that language independent named entity recognition systems can, with relatively small effort, achieve competitive levels of accuracy.

Acknowledgments

The authors would like to thank Radu Florian for preparing the German data and for providing additional German dictionaries that helped to achieve the performance presented in the paper.

English devel.	Precision	Recall	$F_{\beta=1}$
LOC	95.65	94.56	95.10
MISC	90.15	84.38	87.17
ORG	89.03	86.50	87.75
PER	93.76	95.39	94.56
Overall	92.76	91.42	92.08

English test	Precision	Recall	$F_{\beta=1}$
LOC	89.51	89.57	89.54
MISC	77.45	74.36	75.87
ORG	82.57	78.45	80.46
PER	89.67	91.22	90.44
Overall	86.13	84.88	85.50

German devel.	Precision	Recall	$F_{\beta=1}$
LOC	80.69	72.90	76.60
MISC	80.38	42.18	55.32
ORG	83.19	61.40	70.65
PER	86.36	65.95	74.79
Overall	82.98	61.51	70.65

German test	Precision	Recall	$F_{\beta=1}$
LOC	77.71	71.40	74.42
MISC	71.51	39.70	51.06
ORG	79.55	55.37	65.29
PER	91.68	73.81	81.78
Overall	82.00	63.03	71.27

Table 4: System performance with all features on the English and the German data

References

Daniel M. Bikel, Richard L. Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 34(1-3):211–231.

Andrew Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, New York University.

Fred J. Damerau, Tong Zhang, Sholom M. Weiss, and Nitin Indurkhya. 2003. Text categorization for a comprehensive time-dependent benchmark. *Information Processing & Management*, to appear.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings CoNLL-2003*.

A. Mikheev, C. Grover, and M. Moens. 1998. Description of the ltg system used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Adwait Ratnaparkhi. 1999. Learning to parse natural language with maximum entropy models. *Machine Learning*, 34:151–175.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*.

B.M. Sundheim. 1995. Overview of results of the MUC-6 evaluation. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*.

Tong Zhang, Fred Damerau, and David E. Johnson. 2002. Text chunking based on a generalization of Winnow. *Journal of Machine Learning Research*, 2:615–637.