Probabilistic named entity verification

Yi-Chung Lin and Peng-Hsiang Hung
Advanced Technology Center, Computer and Communications Laboratories,
Industrial Technology Research Institute, Taiwan
{lyc,phhung}@itri.org.tw

Abstract

Named entity (NE) recognition is an important task for many natural language applications, such as Internet search engines, document indexing, information extraction and machine translation. Moreover, in oriental languages (such as Chinese, Japanese and Korean), NE recognition is even more important because it significantly affects the performance of word segmentation, the most fundamental task for understanding the texts in oriental languages. In this paper, a probabilistic verification model is designed for verifying the correctness of a named entity candidate. This model assesses the confidence level of a candidate not only according to the candidate's structure but also according to its context. In our design, the clues for confidence measurement are collected from both positive and negative examples in the training data in a statistical manner. Experimental results show that the proposed method significantly improves the F-measure of Chinese personal name recognition from 86.5% to 94.4%.

Introduction

Named entity (NE) recognition (or proper name recognition) is a task to find the entities of person, location, organization, date, time, percentage and monetary value in text documents. It is an important task for many natural language applications, such as Internet search engines, document indexing, information extraction and machine translation. Moreover, in oriental languages (such as Chinese, Japanese and Korean), NE recognition is even more important because it significantly affects the performance of word segmentation, the most fundamental task for understanding the texts in oriental languages. Therefore, a high-accuracy NE recognition method is highly demanded for most natural language applications in various languages.

There are two major approaches to NE recognition: the handcrafted approach (Grishman, 1995) and the statistical approach (Bikel, 1997; Chen, 1998; Yu, 1998). In the first approach, a system usually relies on a large number of handcrafted rules. This kind of systems can be rapid prototyped but are hard to scale up. In fact, there will be numerous exceptions for most handcrafted rules. It is generally expensive and impossible to code for every exception we can imagine, not to mention those exceptions we are not able to think of. Another serious problem with the handcrafted approach is that systems are hard to be ported across different domains and different languages. Porting a handcrafted system usually means rewriting all its rules. Therefore, the statistical approach is becoming more and more popular because of its cost-effectiveness in scaling up and porting systems.

In general, the statistical approach to NE recognition can be viewed as a two-stage process. First, according to dictionaries and/or pattern matching rules, the input text is tokenized into tokens. Each token may be a word or an NE candidate which can consist of more than one word. Then, the most likely token sequence is selected according to a statistical model, such as Markov model (Bikel, 1997; Yu, 1998) or maximum entropy model (Borthwick, 1999). Although, the statistical NE recognition is much more scalable and portable, its performance is still not satisfactory. The insufficient coverage/precision of pattern matching rules and unknown words are the major sources of errors. Furthermore, the role of the statistical model is to assess the relative possibilities of all possible token sequences and select the most probable

one. The scores obtained from the statistical model can be used for a comparison of competing token sequences, but not for an assessment of the probability that a spotted named entity is correct.

To reduce the recognition errors, we propose a probabilistic verification model to verify the correctness of a named entity. This model assesses the confidence level of a named entity candidate not only according to the candidate's structure but also according to its contexts. In our design, the clues for confidence measurement are collected from both positive and negative examples in the training data. Therefore, the confidence measure has strong discriminant power for judging the correctness of a named entity. In the experiments of Chinese personal name recognition, the proposed verification model increases the F-measure from 86.5% to 94.4%, which corresponds to 58.5% error reduction rate, where "error rate" is defined as "100% – F-measure".

1. Named Entity Verification

As mentioned before, there are several kinds of named entities, including person, location, organization, date, time, percentage and monetary value. In the following description, we use the task of verifying Chinese personal name as an example. However, our proposed method is also applicable on verifying other kinds of named entities in different languages.

Before introducing our approach, we first describe the notations that will be used. In this proposal, a random variable is written with a boldface italic letter. An outcome of a random variable $\boldsymbol{o}$ is denoted as o. If there is no confusion, we usually use $P(o)$ to denote the probability $P(\boldsymbol{o} = o)$. A symbol sequence "$x_1, \cdots, x_n$" is denoted as "x_1^n". Likewise, "$x_{Y,1}^{Y,n}$" denotes the sequence "$x_{Y,1}, \cdots, x_{Y,n}$".

The task of verifying a named entity candidate is viewed as making an acceptance or rejection decision according to the text segment consisting of the candidate and its context. Without loss of generality, a text segment is

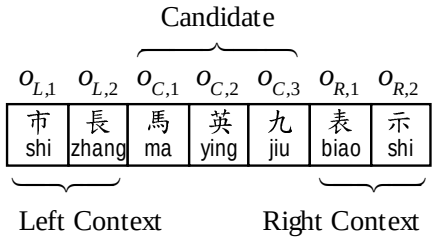

Figure 1: Example of a text segment.

considered as an outcome of the random vector $\boldsymbol{O} = [\boldsymbol{o}_{L,1}^{L,x}, \boldsymbol{o}_{C,1}^{C,y}, \boldsymbol{o}_{R,1}^{R,z}]$. The outcome of each random variable in $\boldsymbol{O}$ is one basic element of text. In Chinese, the basic elements of text are Chinese characters. However, in English, the basic elements are English words. Figure 1 shows an example of a text segment in which the size of the candidate to be verified is 3 (i.e., consists of three Chinese characters) and the sizes of its left context and right context are set to 2 (i.e., two Chinese characters).

Figure 2 depicts the flowchart of our verification approach. First, the candidate in the input text segment is parsed by a predefined grammar. If the candidate is ill-formed (i.e., fail

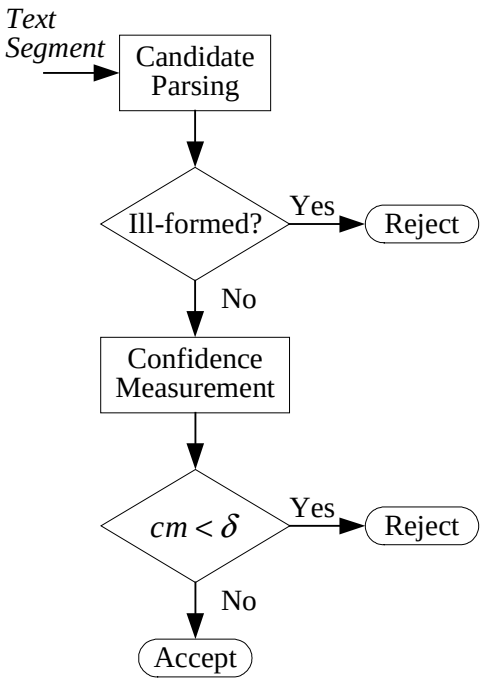

Figure 2: Flowchart of the verification method.

to be parsed), it will be rejected immediately. Otherwise, the text segment is passed to the confidence measurement module to assess the confidence level that the candidate in the text segment is to be a named entity. If the confidence measure is less than a predefined threshold, the candidate will be rejected. Otherwise, it will be accepted.

2. Confidence Measurement

The basic idea of our approach is to formulate the confidence measurement problem as the problem of hypothesis testing. The null hypothesis H_0 in which the candidate is a name is tested against the alternative hypothesis H_1 in which the candidate is not a name. According to Neyman-Pearson Lemma, an optimal hypothesis test involves the evaluation of the following log likelihood ratio:

$$\begin{aligned} &LLR(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z}) \\ &= \log \frac{P(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z} \mid H_0)}{P(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z} \mid H_1)} \\ &= \log P(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z} \mid H_0) \\ &\quad - \log P(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z} \mid H_1) \end{aligned} \tag{1}$$

where $P(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z} \mid H_0)$ is the likelihood of the candidate and its left and right contexts given the hypothesis that the candidate is a name and $P(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z} \mid H_1)$ is the likelihood of the candidate and its left and right contexts given the hypothesis that the candidate is not a name. The hypothesis test is performed by comparing the log likelihood ratio $LLR(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z})$ to a predefined critical threshold δ. If $LLR(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z}) \geq \delta$, the null hypothesis will be accepted. Otherwise, it will be rejected.

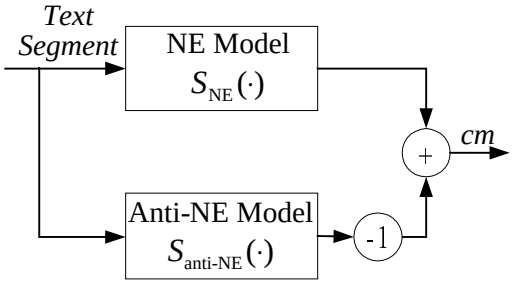

Figure 3: Block diagram of the confidence measurement module.

In our design, as shown in Figure 3, the confidence measurement module consists of two models, named NE model and anti-NE model. The NE model is used to assess the value of $\log P(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z} \mid H_0)$ and the anti-NE model is used to assess the value of $\log P(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z} \mid H_1)$.

2.1. NE Model

The purpose of the NE model is to evaluate the value of $\log P(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z} \mid H_0)$, the log likelihood of the candidate and its left and right contexts given the hypothesis that the candidate is a name. Since it is infeasible to directly estimate the probability $P(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z} \mid H_0)$, it is approximated as follows:

$$\begin{aligned} &P(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z} \mid H_0) \equiv P_0(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z}) \\ &\approx P_0(o_{L,1}^{L,x}) P_0(o_{C,1}^{C,y}) P_0(o_{R,1}^{R,z}) \end{aligned} \tag{2}$$

where the subscript of $P_0(\cdot)$ indicates the probability is evaluated given that the null hypothesis is true. The probability $P_0(o_{L,1}^{L,x})$ is further approximated according to the bigram model as follows:

$$P_0(o_{L,1}^{L,x}) \approx \prod_{i=1}^{x} P_0(o_{L,i} \mid o_{L,i-1}) \tag{3}$$

where $P_0(o_{L,1} \mid o_{L,0}) \equiv P_0(o_{L,1})$. One should notice that we do not assume that the random sequence $\boldsymbol{o}_{L,1}^{L,x}$ is time invariant. For example, the probability $P(\boldsymbol{o}_{L,i} = x \mid \boldsymbol{o}_{L,i-1} = y)$ is not assumed to be equal to $P(\boldsymbol{o}_{L,2} = x \mid \boldsymbol{o}_{L,1} = y)$ for $i \geq 3$. Likewise, the probability $P_0(o_{R,1}^{R,z})$ is also further approximated as follows:

$$P_0(o_{R,1}^{R,z}) \approx \prod_{i=1}^{z} P_0(o_{R,i} \mid o_{R,i-1}) \tag{4}$$

where $P_0(o_{R,1} \mid o_{R,0}) \equiv P_0(o_{R,1})$.

The probability corresponding to the candidate is evaluated by applying the SCFG (Sto-

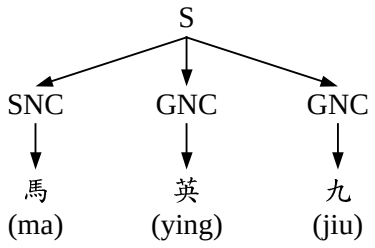

Figure 4: A parse tree of the Chinese personal name "馬英九(ma ying jiu)".

chastic Context-free Grammar) model (Fujisaki, 1989) as follows:

$$P_0(o_{C,1}^{C,y}) = \sum_T P_0(T) \approx \max_T P_0(T) = \max_T \prod_{A\to\alpha\in T} P_0(\alpha \mid A) \quad (5)$$

where T stands for one possible parse tree that derive the candidate, $A \to \alpha$ indicates a rule in the parse tree T, A stands for the left-hand-side symbol of the rule and α stands for the sequence of right-hand-side symbols of the rule. Figure 4 shows an example of a parse tree of the Chinese personal name candidate "馬英九(ma ying jiu)", where "馬(ma)" is the surname and "英九(ying jiu)" is the given name. In this figure, the symbol "S" denotes the start symbol, the symbol "SNG" denotes the nonterminal deriving surname characters and the symbol "GNC" denotes the nonterminal deriving given name characters. As a result, according to equations (2) -(5), the scoring function in the NE model is defined as equation (6) to assess the log likelihood of the text segment "$o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z}$" given the null hypothesis that "$o_{C,1}^{C,y}$" is a name.

$$S_{\text{NE}}(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z}) = \sum_{i=1}^{x} \log P_0(o_{L,i} \mid o_{L,i-1}) + \sum_{i=1}^{z} \log P_0(o_{R,i} \mid o_{R,i-1}) + \max_T \sum_{A\to\alpha\in T} \log P_0(\alpha \mid A) \quad (6)$$

where T is one possible parse tree that derive the candidate "$o_{C,1}^{C,y}$".

2.2. Anti-NE Model

The purpose of the anti-NE model is to evaluate the value of $\log P(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z} \mid H_1)$, the log likelihood of the candidate and its left and right contexts given the hypothesis that the candidate is not a name. Since it is infeasible to directly estimate the probability $P(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z} \mid H_1)$, it is approximated as follows:

$$P(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z} \mid H_1) = P_1(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z}) \approx \prod_{i=1}^{x} P_1(o_{L,i} \mid o_{L,i-1}) \times \prod_{i=1}^{y} P_1(o_{C,i} \mid o_{C,i-1}) \times \prod_{i=1}^{z} P_1(o_{R,i} \mid o_{R,i-1}) \quad (7)$$

where $o_{R,0} \equiv o_{C,y}$, $o_{C,0} \equiv o_{L,x}$, and $P_1(o_{L,1} \mid o_{L,0}) \equiv P_1(o_{L,1})$. Therefore, the following scoring function is used in the anti-NE model to assess the log likelihood of the text segment "$o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z}$" given the alternative hypothesis that "$o_{C,1}^{C,y}$" is not a name.

$$S_{\text{anti-NE}}(o_{L,1}^{L,x}, o_{C,1}^{C,y}, o_{R,1}^{R,z}) = \sum_{i=1}^{x} \log P_1(o_{L,i} \mid o_{L,i-1}) + \sum_{i=1}^{y} \log P_1(o_{C,i} \mid o_{C,i-1}) + \sum_{i=1}^{z} \log P_1(o_{R,i} \mid o_{R,i-1}) \quad (8)$$

3. Experiment Setup

The proposed named entity verification method is used to recognize Chinese personal names from news. In Chinese, most of the personal names consist of three Chinese characters. The first character is a surname. The last two characters are a given name. Therefore, our preliminary experiments are focused on recognizing the personal names of three Chinese characters.

In our experiments, the training corpus consists of about 14,339,000 Chinese characters collected from economy and industry news. This corpus should be annotated to estimate the probabilistic parameters of the scoring functions $S_{\text{NE}}(\cdot)$ and $S_{\text{anti-NE}}(\cdot)$. However, labeling such large amount of data is too costly or prohibited even if it is possible. Therefore, labeling

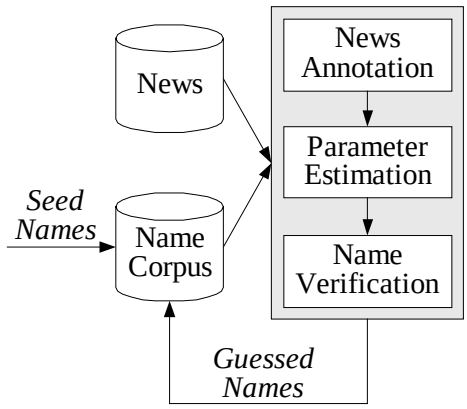

Figure 5: EM-style bootstrapping.

methods that can be bootstrapped from a little seed data or a few seed rules (Collins, 1999; Cucerzan, 1999) are highly demanded to automatically annotate the training data. In the following section, we propose an EM-style bootstrapping procedure (Cucerzan, 1999) for annotating the training data automatically.

3.1. EM-Style Bootstrapping

The Expectation-Maximization (EM) algorithm (Moon, 1996) has been widely used to estimate model parameters from incomplete data in many different applications. In this section, an EM-style bootstrapping procedure is proposed to automatically annotate the named entities in the training corpus. It iteratively uses the proposed verification model to label the training corpus (expectation step), and then uses the labeled training corpus to re-estimate the parameters of the verification model (maximization step). Figure 5 shows the flowchart of the bootstrapping procedure. First, we collect the names of 541 famous people, including government officers and CEOs of big companies. These names are used as seed names of the name corpus. Then, the news is automatically annotated according to the name corpus. The annotated corpus is used to estimate the probabilistic parameters of the scoring functions. Afterward, the proposed verification procedure is used to verify every possible name candidate in the news. The candidates whose confidence measures are larger than a predefined threshold are determined to be names. Currently, if the confidence measures of two overlapped candidates (such as “ma ying jiu” and “ying jiu biao” in Figure 1) pass the threshold, both of them are determined as names. Although this strategy is inadequate, it does not make too much trouble because the chance to get overlapped names is very small in our experiments. Finally, these guessed names are added to the name corpus which will be used to annotate the news in next iteration.

In our case, after four iterations, the size of name corpus is enlarged from 541 to 6,296, as shown in Table 1. The total occurrence frequency of these 6,296 names in the training corpus is 40,345.

3.2. Baseline Model

In the past, many researchers have studied the problem of Chinese personal name recognition. Chang (1994) used the 0-order Markov model to segment a text into words, including Chinese personal names. In his approach, a name probability model is proposed to estimate the probabilities of Chinese personal names. Sproat (1994) proposed to recognize Chinese personal names with the stochastic finite-state word segmentation algorithm. His approach is similar to Chang’s, except that the name probability model is slightly different. In addition to name probability, Chen (1998) also add extra scores to a name candidate according to context clues (such as position, title, speech-act verbs). In the researches mentioned above, the reported F-measure performances on recognizing Chinese personal names are somewhere between 70% and 86%. Since these performances are meas-

Iteration	Nunber of distinct names	Total frequency of names
0	541	18310
1	3389	31157
2	5327	37423
3	6055	39977
4	6296	40345

Table 1: Number of distinct names in the name corpus and total frequency of names in the annotated news during the bootstrapping iteration.

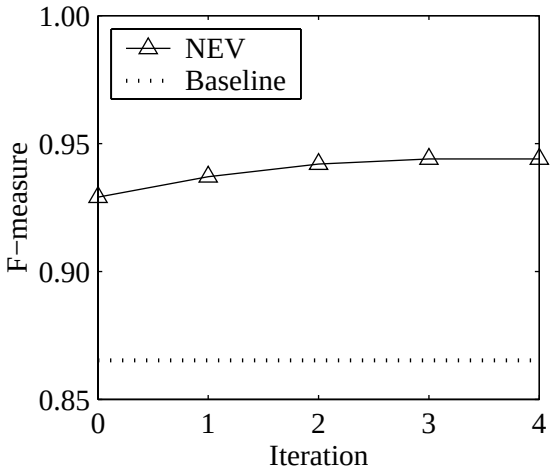

Figure 6: The performances of baseline and name entity verification (NEV).

ured based on different data, higher reported performance does not imply better. In fact, the name probability models used in these researches are very similar. Their performances should be comparable to each other. Therefore, in this paper, Chang's approach, whose reported F-measure is 86%, is chosen as the baseline model.

The baseline model is additionally equipped with a dictionary of 72,333 Chinese words. The prior probabilities of words are estimated from *Academia Sinica Balanced Corpus*, which contains about 2 million Chinese words.

4. Experimental Results and Discussions

Both the baseline model and the proposed name entity verification model (named NEV model) are tested on the same testing corpus. The testing corpus, also collected from economy and industry news, consists of about 737,000 characters. This corpus is annotated manually and contains totally 2,545 Chinese personal names.

The F-measure of the baseline model is 86.5% (as indicated by the dashed line in Figure 6). The precision and recall rates of the baseline model are 79.1% and 95.5% respectively. Although the recall rate of the baseline model is high, the precision rate is pretty low. Over 20% of the name candidates proposed by the baseline model are incorrect.

In our experiments, the sizes of the left- and right-context windows of the NEV model are set to 2. In Figure 6, the solid line with triangle markers depicts the F-measure of the NEV model versus the iteration number of bootstrapping. The F-measure saturates after 3 iterations. After 4 iterations, the F-measure of the NEV model reaches 94.4%. The corresponding precision and recall rates are 96.4% and 92.5% respectively. Compared with the baseline model, the precision rate is greatly improved from 79.1% to 96.4% with a little sacrifice in recall rate. The F-measure is improved from 86.5% to 94.4%, which corresponds to 58.5% error reduction rate, where "error rate" is defined as "100% – F-measure".

Table 2 lists three examples of the misrecognized names made by the baseline model. These examples clearly show that the baseline model tends to incorrectly group consecutive single characters, either from unknown words or single-character words, into names. In the first two examples, the single characters come from the unknown location name "卡羅來納(ga luo lai na; *Carolina*) and the unknown company name "羅技(luo ji; *Logitech*)". The single characters in the last example are single-character words "季(gi; *quarter*)", "全(quan; *all*)" and "美(mei; *USA*)".

Without the inadequate strong tendency of grouping single characters, the NEV model is able to avoid the misrecognition errors made by the baseline model. The NEV model assesses the confidence measure of each name candidate according to the context around the candidate. In Table 2, the name candidates in the shaded boxes are rejected by the NEV model because

以 北 卡 羅 來 納 州 為 例 yi bei ga luo lai na zhou wei li (*take North Carolina State as an example*)
大 廠 羅 技 在 去 年 … da chang luo ji zai qu nian (*In last year, the big company Logitech ...*)
第 一 季 全 美 勞 動 者 … di yi ji quan mei lao dong zhe (*In the first quarter, the workers in all USA ...*)

Table 2: Examples of the incorrect Chinese personal names (in the shaded boxes) produced by the baseline model.

their confidence measures are too low.

To sum up, the experimental results demonstrate that the contextual information, either from positive examples or from negative examples, is very helpful for named entity verification. Besides, the superiority of the NEV model also shows that the proposed probabilistic score functions $S_{\text{NE}}(\cdot)$ and $S_{\text{anti-NE}}(\cdot)$ are effective in providing the scores to produce a reliable confidence measure. Especially, the proposed named entity verification approach does not require any dictionary in advance.

Conclusion

Named entity (NE) recognition is an important task for many natural language applications, such as Internet search engines, document indexing, information extraction and machine translation. Moreover, in oriental languages (such as Chinese, Japanese and Korean), NE recognition is even more important because it significantly affects the performance of word segmentation, the most fundamental task for understanding the texts in oriental languages.

In this paper, a probabilistic verification model is proposed to verify the correctness of a named entity. This model assesses the confidence level of a name candidate not only according to the candidate's structure but also according to its contexts. The clues for confidence measurement are collected from both positive and negative examples in the training data. Therefore, the confidence measure has strong discriminant power for judging the correctness of a named entity. In the experiments of Chinese personal name recognition, the proposed verification model greatly increases the precision rate from 79.1% to 96.4% with a little sacrifice in recall rate. The F-measure is improved from 86.5% to 94.4%, which corresponds to 58.5% error reduction rate, where "error rate" is defined as "100% – F-measure".

Acknowledgements

This paper is a partial result of Project A311XS1211 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, R.O.C. Especially thanks to the CKIP group of Academia Sinica for providing the *Academia Sinica Balanced Corpus*.

References

Bikel, D., Miller S., Schwartz R., and Weischedel R. (1997) *Nymble: A High-performance Learning Name Finder.* In Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 194–201.

Borthwick, A. (1999) *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. Thesis, New York University.

Chang J., Chen S., Ker S., Chen Y. and Liu J. (1994) *A Multiple-Corpus Approach to Recognition of Proper Names in Chinese Texts.* Computer Processing of Chinese and Oriental Languages, Vol. 8, No. 1, pp. 75-85.

Chen, H., Ding Y., Tsai S. and Bian G. (1998) *Description of the NTU System used for MET2*. in Proceedings of the 7th Message Understanding Conference (MUC-7)

Collins, M. and Singer, Y. (1999) *Unsupervised Models for Named Entity Classification*. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 100-110.

Cucerzan S. and Yarowsky D. (1999) *Language independent named entity recognition combining morphological and contextual evidence.* In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 90-99.

Fujisaki, T., Jelinek F., Cocke J., Black E. and Nishino T. (1989), *A Probabilistic Parsing Method for Sentence Disambiguation*. In Proceedings of the International Workshop on Parsing Technologies, pp. 85-94.

Grishman, R. (1995) *The NYU system for MUC-6 or where's the syntax?* In Proceedings of the 6th Message Understanding Conference (MUC-6), pp. 167-175.

Moon, T. K. (1996) *The Expectation-Maximization Algorithm*, IEEE Signal Processing Magazine, November, 1996, pp. 47-60.

Sproat R. and Chang N. (1994) *A Stochastic Finite-State Word-Segmentation Algorithm for Chinese*. In Proceeding of 32nd Annual Meeting of the Association for Computational Linguistics, pp. 66-73.

Yu, S., Bai S. and Wu P. (1998) *Description of the Kent Ridge Digital Labs System Used for MUC-7*. In Proceedings of the 7th Message Understanding Conference (MUC-7)