An Agent-based Approach to Chinese Named Entity Recognition

Shiren Ye Tat-Seng Chua Liu Jimin
School of Computing, National University of Singapore,
Singapore, 117543
yesr@comp.nus.edu.sg chuats@comp.nus.edu.sg Liujm@comp.nus.edu.sg

Abstract

Chinese NE (Named Entity) recognition is a difficult problem because of the uncertainty in word segmentation and flexibility in language structure. This paper proposes the use of a rationality model in a multi-agent framework to tackle this problem. We employ a greedy strategy and use the NE rationality model to evaluate and detect all possible NEs in the text. We then treat the process of selecting the best possible NEs as a multi-agent negotiation problem. The resulting system is robust and is able to handle different types of NE effectively. Our test on the MET-2 test corpus indicates that our system is able to achieve high F_1 values of above 92% on all NE types.

1. Introduction

Named entity (NE) recognition is a fundamental step to many language processing tasks. It was a basic task of the Message Understanding Conference (MUC) and has been studied intensively. Palma & Day (97) reported that person (PER), location (LOC) and organization (ORG) names are the most difficult sub-tasks as compared to other entities as defined in MUC. This paper thus focuses only on the recognition of PER, LOC and ORG entities.

Recent research on NE recognition has been focused on the machine learning approach, such as the transformation-based learning (Aberdeen 95), hidden Markov model (Bikel et al. 97), decision tree (Sekin et al. 98), collocation statistics (Lin 98), maximum entropy model (Borthwick 99), and EM bootstrapping (Cucerzan & Yarowsky 99). Other than English, several recent works examined the extraction of information from Spanish, Chinese, and Japanese (Isozaki 01). Most approaches for Chinese NE recognition used handcrafted rules, supplemented by word or character frequency statistics. These methods require a lot of resources to model the NEs. Chen et al. (98) used 1-billion person name dictionary and employed mainly internal word statistics with no generalization. Yu et al. (98) employed a common framework to model both the context and information residing within the entities, and performed rule generalization using POS (part-of-speech) and some semantic tags. A similar system is also reported in Luo & Song (01).

Chinese NE recognition is much more difficult than that in English due to two major problems. The first is the word segmentation problem (Sproat et al. 96, Palmer 97). In Chinese, there is no white space to delimit the words, where a word is defined as consisting of one or more characters representing a linguistic token. Word is a vague concept in Chinese, and Palmer (97) showed that even native speakers could only achieve about 75% agreement on “correct” segmentation. As word segmentation is the basic initial step to almost all linguistic analysis tasks, many techniques developed in English NLP cannot be applied to Chinese.

Second, there is no exterior feature (such as the capitalization) to help identify the NEs, which share many common characters with non-NE (or common words). For example, while 中 is normally associated with the country *China*, it could also mean the concepts *in, at* or *hit*; and张 normally refers to the surname *Zhang*, but it also means the concepts *open, sheet* or *spread*. Moreover, proper names in Chinese may contain common words and vice versa.

Because of the above problems, the use of statistical and heuristic rules commonly adopted in most existing systems is inadequate to tackle the Chinese NE recognition problem. In this paper, we consider a new approach of employing a rationality model in a multi-agent framework.

The main ideas of our approach are as follows. First, we use an NE rationality measure to evaluate the probability of a sequence of tokens being a specific NE type, and adopt a greedy approach to detect all possible NEs. Second, we treat the process of selecting the best NEs among a large set of possibilities as a multi-agent negotiation problem. We test our overall approach on the MET-2 test set and the system is able to achieve high F_1 values of over 92% on all NE types. The results are significantly better than most reported systems on MET-2 test set.

The rest of the paper describes the details of our rationality-based and multi-agent negotiation approach to detect and refine NEs.

2. Rationality Model for NE Detection

2.1 Named Entity and Its tokens Feature

For clarity and without lost of generality, we focus our discussion mainly on PER entity. The problems and techniques discussed are applicable to LOC and ORG entities. We consider a simple PER name model comprising the surname followed by the first-name. Given the presence of a surname (as cue-word) in a token sequence, we compute the likelihood of this token playing the role of surname and the next token as the first-name. The pair could be recognized as PER only if both tokens are labeled as positive (or of the right types) as shown in Table 1. If either one of both of the tokens are evaluated negatively, then the pair will not be recognized as PER based on the model defined above.

Sentence	PER?	Label	Remarks

请张飞讲话…	Y	张(+)
		飞(+)
			... invite Zhang Fei to speak ...

一张飞机票…	N	张(-)
		飞(-)
			… a piece of airline ticket …

老张飞上海…	?	张(+)
		飞(-)
			//Illegal PER

…张飞…*	?	张(-)
		飞(+)
			//Illegal PER

* Strictly, *张将军* and Mr. Zhang are not really person names. They are references to person names and should be detected via co-reference.

Table 1: An example of NE and non-NE

Although the example depicted in Table 1 is very simple, the same idea can be extended to the more complex NE Types for ORG and LOCs. The number of tokens in a NE may vary from 2 in PER to about 20 for ORG. One constraint is that the sequencing of tokens and their labels must be consistent with the respective NE type. Also, there are grammatical rules governing the composition of different NE type. For example, LOC may consist of a sequence of LOCs; and ORG may include PER and/or LOC on its left. Thus by considering one pair of tokens at a time, and by extending the token sequence to the adjacent token one at a time, we can draw similar conclusion as that depicted in Table 1 for complex NE types.

2.2 The Rationality Computation

If we know the probability distribution of each type of token in a window, NE recognition is then the procedure of evaluating the rationality or certainty of a sequence of tokens with respect to a NE type. Motivated by the results in Table 1 we view NE recognition as a special coloring problem. Initially, all the tokens in the corpus are considered as a sequence of White balls. Given a chain of tokens appears in a NE window, we want to use the probability distribution of these tokens to re-paint some of the white balls to different colors. A sequence of appropriately colored balls would induce an appropriate NE.

For simplicity, we again focus on PER NE type with 2 tokens. The surname token will be colored red and first-name blue. We assume that the number of PER names in the corpus is N, and the rest of tokens is M. Because there are N surname and N first-name tokens in the corpus, the total number of tokens is M+2N. Hence the marginal probability of PER name is Pr(PER)=N/(2N+M) .

	Red		Blue		White	
	Format	Pr.	Format	Pr.	Format	Pr.
Red	$a_R b_R$	0	$a_R b_B$	1	$a_R b_W$	0
Blue	$a_B b_R$	N/(N+M)	$a_B b_B$	0	$a_B b_W$	M/(N+M)
White	$a_W b_R$	N/(N+M)	$a_W b_B$	0	$a_W b_W$	M/(N+M)

Note: Red – Surname; Blue – First-name; White - Others

Table 2: Possibility combination of neighboring tokens within the corpus for PER

Table 2 shows the possible relationships between the red and blue balls for the PER NE type by considering the grammer that the surname must be followed by a first-name in a formal PER. As we only permit the token pair for PER to be labeled as a red ball followed by a blue ball, the following sequences are not possible under our model: (a) a red (or blue) ball follows by itself; (b) a red ball follows by white ball; and (c) a white ball follows by the blue ball. Thus a_Rb_R (a red follows by a red), a_Rb_W, a_Bb_B, and a_Wb_B are illegal combinations.

Given a pair of tokens *a* and *b* in the corpus, they are labeled as surname $|a_R|$ and $|b_R|$ times, as first-name $|a_B|$ and $|b_B|$ times, and as non-PER $|a_W|$ and $|b_W|$ times respectively. The expected value of a token sequence *ab* representing a PER when *a* is red and *b* is blue is:

$$|a_R b_B| = |a_R| \cdot \frac{|b_B|}{N} = \frac{|a_R| \cdot |b_B|}{N} \qquad (1)$$

The expected value of the cases when the token pair ab is not a PER name is the sum of expected values of four cases: a_Bb_R, a_Bb_W, a_Wb_R, a_Wb_W (see Table 2), which after simplification, is given by:

$$\begin{aligned}|a_{\bar{R}} b_{\bar{B}}| &= |a_B b_R| + |a_B b_W| + |a_W b_R| + |a_W b_W| \\ &= \frac{|a_B| \cdot |b_R|}{N+M} + \frac{|a_B| \cdot |b_W|}{N+M} + \frac{|a_W| \cdot |b_R|}{N+M} + \frac{|a_W| \cdot |b_R|}{N+M} \\ &= \frac{(|a_B| + |a_W|) \cdot (|b_R| + |b_W|)}{N+M} \qquad (2)\end{aligned}$$

The ratio between the cases when ab is a PER versus when ab is not a PER is:

$$\Re^C_{ab} = \frac{|a_R b_B|}{|a_{\bar{R}} b_{\bar{B}}|} = \lambda \cdot \Re^R_a \cdot \Re^B_b \qquad (3)$$

where $\Re^R_a = \frac{|a_R|}{(|a_B| + |a_W|)}; \Re^B_b = \frac{|b_B|}{(|b_R| + |b_W|)}$; and $\lambda = \frac{N+M}{N}$. We call $\Re^c_{ab}$, $\Re^R_a$, *and* $\Re^R_b$ the rationality values of tokens *ab*, *a* and *b* of being a PER, red ball or blue ball respectively.

On the other hand, the probabilities of *a* as a surname (red ball) and *b* as a first-name (blue ball) are:

$$P^R_a = \frac{|a_R|}{|a_R| + (|a_B| + |a_W|)}, P^B_b = \frac{|b_B|}{|b_B| + (|b_R| + |b_W|)}$$

Thus, $\Re^R_a = \frac{P^R_a}{1 - P^R_a}; \Re^B_b = \frac{P^B_b}{1 - P^B_b}$ (4)

The form of Equation (4) is similar to the concept of odds likelihood O(h), first introduced in Duda et al. (79) as a generic term to denote the ratio of the probability and converse probability in the Prospector system, namely:

$$O(h) = \frac{P(h)}{P(\text{-}h)} = \frac{P(h)}{1 - P(h)} \qquad (5)$$

Eq. (5) is used in a modified version of the Bayes theorem to solve the uncertainty reasoning problems. Surprisingly, our approach of rationality $\Re$ for NE with two tokens can be deduced as the product of their odds-likelihood. By linking the concept of odds-likelihood and rationality, we can compute the probability of a sequence of tokens being a specific NE type.

Since the rationality values of tokens could vary from 0 to ∞, it may incur overflow or underflow during the rationality evaluation. This is especially so for unknown tokens where their rationality values will be zero. To resolve this problem, we construct a piecewise function to map the rationality values from the range [0, ∞] to $[\Re_{min}, \Re_{max}]$. Here we set the parameters $\Re_{min}$=0.05 and $\Re_{max}$=50, and ensure that most rationality values will retain their original values after transformation.

2.3 The Context of NEs

In addition to identifying the structural information within the NEs, it is equally important to model the context around the NEs. Context is especially pivotal to language such as the Chinese or Korean where there is no white space and capital characters among the tokens. For PER type, the context tokens are likely to be person titles and action words.

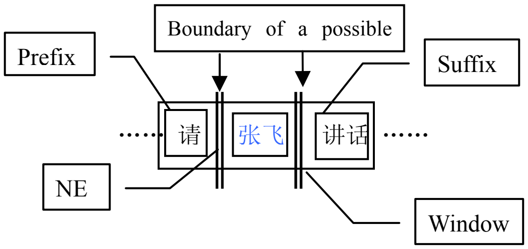


Figure 1: A NE detection window

Thus after we have computed the rationality values of possible NEs, we enlarge the analysis window to cover both the NE candidate and its context. As shown in Figure 1, the window consists of three components: prefix, suffix and the NE candidate. If the NE is at the beginning or end of a paragraph, then the corresponding prefix or suffix is set to void. We can extend the rationality computation for an NE to the context window by incorporating both the prefix and suffix tokens separately.

2.4 The Overall Procedure

The overall procedure for estimating the likelihood of an NE among a sequence of tokens is as follows.

a) Convert prior probability Pr(e) of each token *e* to rationality $\Re$(e). A token *e* may have multiple Pr(e) values, each is dependent on the role token e plays in a possible NE, such as the probability of being a surname, first-name, prefix, suffix, general token or cue-word.

b) At each cue-word position, compute the rationality of a possible NE by considering one pair of tokens at a time, and extending to the next token on the left or right depending on the NE type. The boundaries of PERs are extended forward; while that of ORGs and LOCs are extended backward. Each extension will produce a new NE candidate. The scope of the extension is also determined by the type of NE. The process terminates when the rationality value of the next token falls below a minimum threshold.

c) For all possible NEs, construct the context window and compute its final rationality value within the context window.

The process will result in multiple possible NEs, with most NEs overlapping with one another.

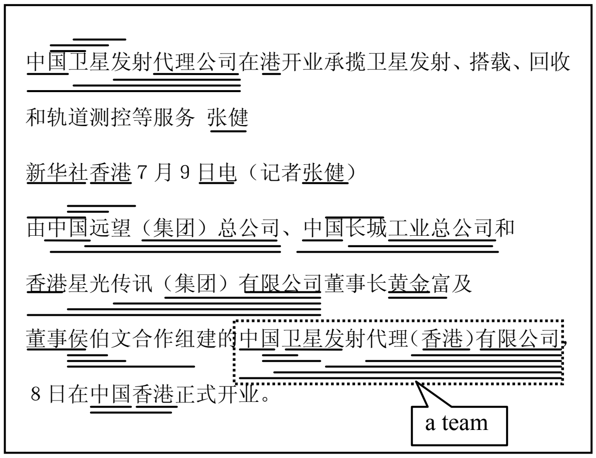
3. Multi-Agent Framework for NE Confirmation

3.1 Relationships between possible NEs

Our greedy approach of identifying all possible NEs using the rationality model results in over segmentation of NEs. Figure 2 shows a list of 80 possible NEs detected from a test article in the MET-2 test corpus. The number of correct NEs in this case is only 13. These possible NEs relate to each other in a complex way. The possible relationships between them are:

a. Overlapping: This is the most common case when the tokens of multiple NEs overlap each other. Examples include "*长城工业总公司*" and "*中国长城工业总公司*". They are both reasonable ORGs if considered separately. However, only one of them can be true.

b. Repetition: Some possible NEs may repeat themselves with same or similar tokens. For example, the NE "*中国卫星发射代理公司*" is similar to "*中国卫星发射代理（香港）有限公司*" in different part of the text. It means that these NEs have same beliefs and could cooperate to enhance each other's belief.



中国卫星发射代理公司在港开业承揽卫星发射、搭载、回收和轨道测控等服务 张健

新华社香港 7 月 9 日电（记者张健）

由中国远望（集团）总公司、中国长城工业总公司和香港星光传讯（集团）有限公司董事长黄金富及董事侯伯文合作组建的中国卫星发射代理(香港)有限公司，8 日在中国香港正式开业。

Figure 2: All possible NEs identified in a test article

c. Unification: When the tokens of two NEs are adjacent to each other in a sentence, they may be unified to become a new NE by combining their tokens. For instance, the NEs "*中国*" and "*香港*" may be combined to form a new NE "*中国香港*". By the way, not all neighboring NEs can be unified because the unification must satisfy the syntactic and semantic specifications of the language. For example, two adjoining PERs cannot be unified, while it is possible for LOCs.

d. Enumerated name list: This is a common language construct to present a list of names. An example of such construct is: "*中国远望(集团)总公司, 中国长城工业总公司,* and *香港星光传讯（集团）有限公司*".

If we knew the relationships between possible NEs, we can use this knowledge to modify the rationality values of possible NEs. The first relationship (overlapping) is of type *competition* while the other three are of type *supporting*. In a *competition* relationship, the rationality values of losing NEs are decremented, whereas in a *supporting* relationship, the rationality of the winning NE can be used to reinforce other NEs.

3.2 Agent-based Reasoning & Negotiation

There is a need to modify the rationality values of possible NEs in order to identify the best possible NEs. One way to achieve this is to employ a decision tree (Sekine 98) to select the best possible candidates. However, it is difficult to use the decision tree to handle multiple relationships between conflicting NEs, and to perform incremental updates of rationality values in situations where the number, distribution and relationships in possible NEs are uncertain. In this work, we adopt a multi-agent approach to refine the rationality of possible NEs and vote the best potential NEs.

Agents are software entities that perform some operations on behalf of their users or another programs with some degree of autonomy, and in so doing, employ some knowledge or representation of the user's goals or desires (Don et al. 96). In our system, we map every possible NE detected to an agent, which acts as the deputy of the NE and depicts all its attributes. Following the approach taken in the DBI system, we use the rationality of the NE as the belief, denoted by Br(A), of agent *A*. Agents are divided into Teams (Decker & Lesser 95) according to their contents and positions in the corpus. The division of agents into teams facilitates the negotiation of agents' beliefs.

The negotiation between agents aims to eliminate underlying conflicts and uncertainty among them. The process of multi-agent negotiation is carried out as follows.

a. We identify agents involved in an unification relationship. These agents will be unified if the constraints of unification are fulfilled. The new agents would inherit the evidences, including the rationality values, of its child agents.

b. We divide the resulting agents into teams. Agents with overlapping tokens will be grouped into same teams, while independent agents will be assigned to different teams.

c. We perform negotiation between agents based on the type of their relationship. For agents that are in *competition* relationship (i.e. those overlapping agents within the same team), we select the agent with the maximal belief (said a_i) as the winner, and decrement the beliefs of the rest of N_t agents in the same team by $\Delta(a_i)$, i.e.

   $$Br(a_j) = Br(a_j) - \Delta(a_i), \text{ for } j=1,.. N_t, \text{ and } j \neq i$$

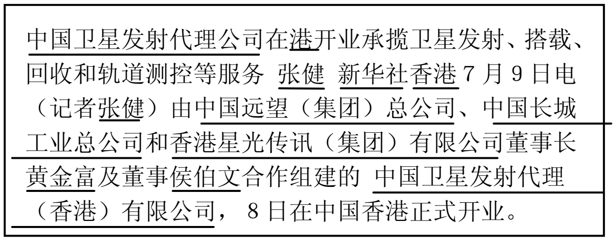
   For agents involved in the *supporting* relations, we again select the agent with the maximal belief (of say a_k) as the winner, but increment the rest of agents in the same set S_k by $\Delta(a_k)$, i.e.

   $$Br(a_j) = Br(a_j) + \Delta(a_k), \text{ for all } j \text{ in } S_k \ \& \ j \neq k$$

d. Repeat step c until the pre-defined rounds of negotiations have been reached.

In order to ensure fairness in the negotiation process, we limit the amount of belief adjustment, $\Delta(a_i)$, during each round of negotiation. If the desired rounds of negotiation is N_R, then the amount of adjustment in each round should be limited to $\Delta(a_i)/N_R$. N_R should be set to allow all agents to have a fair chance to participate in the negotiation process. Here we set N_R to 10.

At the end of negotiation, only agents whose beliefs are greater than the threshold are selected. Figure 3 shows the resulting set of NEs derived from the list given in Figure 2.



中国卫星发射代理公司在港开业承揽卫星发射、搭载、回收和轨道测控等服务 张健 新华社香港７月９日电（记者张健）由中国远望（集团）总公司、中国长城工业总公司和香港星光传讯（集团）有限公司董事长黄金富及董事侯伯文合作组建的 中国卫星发射代理（香港）有限公司，８日在中国香港正式开业。

Fig. 3: NEs after agents-based modification

4. The Overall Process of NE Recognition

Since there is no white space between words in Chinese, the first essential step is to perform preliminary segmentation. Here, we adopt a greedy approach of generating all possible segmentation from the input text by performing the dictionary-based look-up using a common word dictionary. The common word dictionary is generated from the PKU corpus (Yu 99) (see Section 5.1).

Second, we compute the rationality value of each token in the context of being a keyword, general word, or as boundary (prefix or suffix) of a specific NE type.

Third, we identify all possible NE cue-words and use them as seeds of NE candidates. We construct all possible NEs from the cue-word positions through boundary extension and context inclusion.

Forth, we modify the rationality values of all possible NEs using the agent-based negotiation methodology. The conflicts between possible NEs will disappear.

Fifth, we select NEs with the labels of its corresponding seed if their rationality values are above a predefined limit θ. The value θ affects the balance between recall and precision.

5. Experimental Results and Discussions

5.1 The Datasets Used in Our Experiments

We use a number of openly available datasets for our training and testing, including the PKU-corpus (Yu 99), Hownet (Dong & Dong 00), MET2 Chinese resources (Chinchor 02), and two name lists (for foreign and ORG names) collected from the web by using a bootstrapping approach. The PKU is a manually tagged corpus containing one-month of news report from *China's People Daily*. It uses over 30 POS tags including separate tags for surname and first-name. It contains about 37,000 sentences with over 10^6 tokens. From these resources, we generate the following dictionaries and statistics.

a. We use the PKU corpus to build a common word dictionary by removing all words that are tagged as NE. The resulting dictionary contains 37,025 common words.

b. From the PKU corpus, we compute each token's distribution information based on its POS tags, and if it is an NE, its NE type and its role with respect to the NE. Altogether, we obtain the distribution information of about 37,000 different tokens.

c. We maintain a list of LOCs found in the MET-2 test corpus. We do not maintain the PER and ORG lists, because their re-occurrence probabilities are low.

d. We supplement the distribution information derived in step (b) by incorporating tokens obtained from other resources stated above.

The resources we derived are available for down loading at http://www.pris.nus.edu.sg/ie.html

5.2 The Experiment and Results

We test our resulting model on the MET-2 test corpus. Table 3 tabulates the results of our system in terms of recall (Rc), precision (Pr) and F_1 measures. In order to demonstrate the effectiveness of our approach, we perform the tests under 3 different test configurations.

a. We perform the baseline test by simply performing name-dictionary look-up. Notice that we do not use PER dictionary, and hence the performance under PER is left blank (*).

b. We extract all possible NEs by using only the rationality-based approach where the threshold $\Re$ is set to 1.1. If there are conflicts between possible NEs, we simply select the NE with the maximal rationality.

c. We employ the agent-based modification in conjunction with the rationality-based approach to select the best possible NEs.

For comparison purpose, we also list in Table 3 the corresponding results reported in Yu et al. (98) and Chen et al. (98) for the MUC-7 tests.

	Type	N_C	N_P	N_W	N_M	N_S	Rc	Pr	F_1
Base-line test (a)	ORG	79	3	0	295	0	21	98	35.0
	PER	*	*	*	*	*	*	*	*
	LOC	363	84	0	303	26	54	86	66.0
Config (b)	ORG	309	5	28	35	47	83	79	81.0
	PER	154	2	7	11	87	89	62	73.4
	LOC	618	0	29	103	112	82	81	81.7
Config (c)	ORG	356	2	5	14	21	95	93	93.7
	PER	167	1	2	4	9	96	93	94.7
	LOC	703	0	18	29	52	94	91	92.3
Results of Chen et (98)	ORG	393	0	7	77	44	78	83	81.3
	PER	159	0	0	25	56	91	74	81.6
	LOC	583	0	65	102	194	78	69	73.2
Results of Yu et al. (98)	ORG	331	0	14	32	25	88	89	88.5
	PER	160	0	7	7	74	92	66	76.7
	LOC	682	0	1	67	83	91	89	0.0

where $Pr = (N_C + 0.5*N_P)/(N_C + N_W + N_P + N_S)$;
$Rc = (N_C + 0.5*N_P)/(N_C + N_W + N_P + N_M)$;
$F_1 = 2*Pr*Rc/(Pr+Rc)$.
and N_C gives the number of NEs correctly recognized;
N_P denotes the number of NEs partially recognized;
N_W gives the number of NEs incorrectly recognized;
N_M denotes the number of NEs missed; and finally
N_S gives the number of NEs found by the system but not in the tagged list.

Table 3: Results of MET2 under different configurations

Table 3 shows that as we apply the rationality model (Config. b) followed by multi-agent framework (Config. c), the performance of the system improves steadily until it reaches a high performance of over 92% in F_1 value. In fact

Config c results in significant improvements over Conig b in both precision and recall forall NE types. This shows that the agent-based modification could significantly reduce spurious and missing NEs. The performance of our overall system is significantly better than both reported systems as listed in Table 3.

To demonstrate the effectiveness of our approach on general web-based documents, we perform another informal test to recognize NEs on the 100 randomly collected headline news articles from the well-known Chinese web sites (www.sina.com.cn, www.sohu.com, www. zaobao.com, www.Chinese times.com). The topics covered in these articles ranging from politic, economic, society to sports. The informal test shows that our approach could perform well on general web-based articles with F_1 measures of over 90%.

6. Conclusion

Chinese NE recognition is a difficult problem because of the uncertainty in word segmentation. Many existing techniques that require knowledge of word segmentation, and syntactic and semantic tagging of text cannot be applied. In this paper, we propose a new approach of employing a rationality model in a multi-agent framework. We employ a greedy strategy and use the NE rationality measures to detect all possible NEs in the text. We then treat the process of selecting the best possible NEs as the multi-agent negotiation problem. The resulting system is robust and is able to handle different NE models. Our test on the MET-2 test corpus indicates that we could achieve high F_1 values of above 92% on all NE types.

We plan to further test our system on a large-scale test corpus. We will refine our techniques on a wide variety of text corpuses, and apply the bootstrapping technique to tackle the data sparseness problem. Finally, we will extend our research to perform relation and information extraction in multilingual text.

References

Bikel D.M., Schwartz R. & Weischedel R.M. (1999) An Algorithm that Learns What's in a Name. *Machine Learning,* 34(1-3), 211-231

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

Chen H. H., Ding Y. W. Tsai S.C. & Bian, G.W. (1998) Description of the NTU System used for MET-2. In MUC-7 Proc.

Chinchor N. A. (2002), http://www.itl.nist.gov/iaui/894.02/related_projects/muc/.

Cucerzan S. & Yarowsky D. D. (1999) Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In Proc of *1999 Joint SIGDAT Conference on Empirical Methods in NLP & Very Large Corpora*, 90-99.

Decker K., & Lesser V. (1995) Designing a Family of Coordination Algorithm, In Proc Of *1st Int'l Conf. on Multiagent Sys*, 73-80, Menlo Park, CA, AAAI Press.

Don Gilbert, Manny Aparicio, et al (1996) *White paper on intelligent agents* (IBM), http://activist.gpl.ibm.com:81/WhitePaper/ptc2.htm.

Dong Z.D. & Dong Q. (2000) HowNet, available at http://www.keenage.com/zhiwang/e_zhiwang.html.

Duda R., Gaschnig J., & Hart P. (1979) Model design in the prospector consultant system for mineral exploration. In *Expert systems in the micro -electronic age*, Michie D. Ed., Edinburgh Univ. Press, Edinburgh, England.

Isozaki H. (2001) Japanese Named Entity Recognition Based on a Simple Rule Generator and Decision Tree Learning, In ACL'01, 306-313.

Lin D. (1998) Using collocation statistics in information extraction. In MUC-7 Proc.

Luo Z.Y. & Song R. (2001) An Integrated and Fast Approach to Chinese Proper Name Recognition in Chinese Word Segmentation, In Proc. of *Int'l Chinese Computing Conf.*, Singapore 323-328.

Palmer D. D. (1997) A Trainable Rule-Based Algorithm for Word Segmentation, In Proc of *35th of ACL & 8th conf. of EACL*, 321-328.

Sproat R., Shih C., et al (1996) A Stochastic Finite-state Word Segmentation Algorithm for Chinese. *Computational Linguistics*, 22(3), 377-404.

Yu S.H., Bai S.H. & Wu P. (1998) Description of the Kent Ridge Digital Labs System Used For MUC-7, 1998, In MUC-7 Proc.

Yu S.W. (1999) The Specification and Manual of Chinese Word Segmentation and Part of Speech Tagging. http:// www.icl.pku.edu.cn/

Sekine S. (1998) NYU: Description of The Japanese NE System Used for MET-2, in MUC-7 Proc.