Applying Coreference to Improve Name Recognition

Heng JI and Ralph GRISHMAN
Department of Computer Science
New York University
715 Broadway, 7th Floor
New York, NY 10003, U.S.A.
hengji@cs.nyu.edu, grishman@cs.nyu.edu

Abstract

We present a novel method of applying the results of coreference resolution to improve Name Recognition for Chinese. We consider first some methods for gauging the confidence of individual tags assigned by a statistical name tagger. For names with low confidence, we show how these names can be filtered using coreference features to improve accuracy. In addition, we present rules which use coreference information to correct some name tagging errors. Finally, we show how these gains can be magnified by clustering documents and using cross-document coreference in these clusters. These combined methods yield an absolute improvement of about 3.1% in tagger F score.

1 Introduction

The problem of name recognition and classification has been intensively studied since 1995, when it was introduced as part of the MUC-6 Evaluation (Grishman and Sundheim, 1996). A wide variety of machine learning methods have been applied to this problem, including Hidden Markov Models (Bikel et al. 1997), Maximum Entropy methods (Borthwick et al. 1998, Chieu and Ng 2002), Decision Trees (Sekine et al. 1998), Conditional Random Fields (McCallum and Li 2003), Class-based Language Model (Sun et al. 2002), Agent-based Approach (Ye et al. 2002) and Support Vector Machines. However, the performance of even the best of these models[1] has been limited by the amount of labeled training data available to them and the range of features which they employ. In particular, most of these methods classify an instance of a name based on the information about that instance alone, and very local context of that instance – typically, one or two words preceding and following the name. If a name has not been seen before, and appears in a relatively uninformative context, it becomes very hard to classify.

We propose to use more global information to improve the performance of name recognition. Some name taggers have incorporated a name cache or similar mechanism which makes use of names previously recognized in the document. In our approach, we perform coreference analysis and then use detailed evidence from other phrases in the document which are co-referential with this name in order to disambiguate the name. This allows us to perform a richer set of corrections than with a name cache. We then go one step further and process *similar documents* containing instances of the same name, and combine the evidence from these additional instances. At each step we are able to demonstrate a small but consistent improvement in named entity recognition.

The rest of the paper is organized as follows. Section 2 briefly describes the baseline name tagger and coreference resolver used in this paper. Section 3 considers methods for assessing the confidence of name tagging decisions. Section 4 examines the distribution of name errors, as a motivation for using coreference information. Section 5 shows the coreference features we use and how they are incorporated into a statistical name filter. Section 6 describes additional rules using coreference to improve name recognition. Section 7 provides the flow graph of the improved system. Section 8 reports and discusses the experimental results while Section 9 summarizes the conclusions.

2 Baseline Systems

The task we consider in this paper is to identify three classes of names in Chinese text: persons (PER), organizations (ORG), and geo-political entities (GPE). Geo-political entities are locations which have an associated government, such as

[1] The best results reported for Chinese named entity recognition, on the MET-2 test corpus, are 0.92 to 0.95 F-measure for the different name types (Ye et al. 2002).

cities, states, and countries.[2] Name recognition in Chinese poses extra challenges because neither capitalization nor word segmentation clues are explicitly provided, although most of the techniques we describe are more generally applicable.

Our study builds on an extraction system developed for the ACE evaluation, a multi-site evaluation of information extraction organized by the U.S. Government. Following ACE terminology, we will use the term *mention* to refer to a name or noun phrase of one of the types of interest, and the term *entity* for a set of coreferring mentions. We briefly describe in this section the baseline Chinese named entity tagger, as well as the coreference system, used in our experiments.

2.1 Chinese Name Tagger

Our baseline name tagger consists of an HMM tagger augmented with a set of post-processing rules. The HMM tagger generally follows the NYMBLE model (Bikel et al, 1997), but with a larger number of states (12) to handle name prefixes and suffixes, and transliterated foreign names separately. It operates on the output of a word segmenter from Tsinghua University. It uses a trigram model with dynamic backoff. The post-processing rules correct some omissions and systematic errors using name lists (for example, a list of all Chinese last names; lists of organization and location suffixes) and particular contextual patterns (for example, verbs occurring with people's names). They also deal with abbreviations and nested organization names.

2.2 Chinese Coreference Resolver

For this study we have used a rule-based coreference resolver. Table 1 lists the main rules and patterns used. We have extensive rules for name-name coreference, including rules specific to the particular name types. For these experiments, we do not attempt to resolve pronouns, and we only resolve names with nominals when the name and nominal appear in close proximity in a specific structure, as listed in Table 1.

We have used the MUC coreference scoring metric (Vilain et al, 1995) to evaluate this resolver, excluding all pronouns and limiting ourselves to noun phrases of semantic type PER, ORG, and GPE. Using a perfect (hand-generated) set of mentions, we obtain a recall of 82.7% and precision of 95.1%, for an F score of 88.47%. Using the mentions generated by our extraction system, we obtain a recall of 74.3%, a precision of 84.5%, and an F score of 79.07%.[3]

3 Confidence Measures

In order to decide when we need to rely on global (coreference) information for name tagging, we want to have some assessment of the confidence that the name tagger has in individual tagging decisions. In this paper, we use two tools to reach this goal. The first method is to use three manually built proper name lists which include common names of each type (selected from the high frequency names in the user query blog of COMPASS, a Chinese search engine, and name lists provided by Linguistic Data Consortium; the PER list includes 147 names, the GPE list 226 names, and the ORG list 130 names). Names on these lists are accepted without further review.

The second method is to have the HMM tagger compute a probability *margin* for the identification of a particular name as being of a particular type. Scheffer et al. (2001) used a similar method to identify good candidates for tagging in an active learner. During decoding, the HMM tagger seeks the path of maximal probability through the Viterbi lattice. Suppose we wish to evaluate the confidence with which words w_i, …, w_j are identified as a name of type T. We compute

$$\text{Margin}(w_i, \ldots, w_j; T) = \log P_1 - \log P_2$$

Here P_1 is the maximum path probability and P_2 is the maximum probability among all paths for which some word in w_i, …, w_j is assigned a tag other than T.

A large margin indicates greater confidence in the tag assignment. If we exclude names tagged with a margin below a threshold, we can increase the precision of name tagging at some cost in recall. Figure 1 shows the trade-off between margin threshold and name recognition performance. Names with a margin over 3.0 are accepted on this basis.

[2] This class is used in the U.S. Government's ACE evaluations; it excludes locations without governments, such as bodies of water and mountains.

[3] In our scoring, we use the ACE keys and only score mentions which appear in both the key and system response. This therefore includes only mentions identified as being in the ACE semantic categories by both the key and the system response. Thus these scores cannot be directly compared against coreference scores involving all noun phrases.

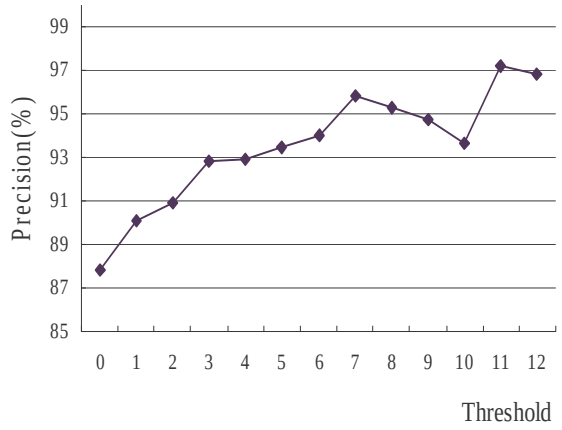

Figure 1: Tradeoff between Margin Threshold and name recognition performance

4 Distribution of Name Errors

We consider now names which did not pass the confidence measure tests: names not on the common name list, which were tagged with a margin below the threshold. We counted the accuracy of these "obscure" names as a function of the number of mentions in an entity; the results are shown in Table 2.

The table shows that the accuracy of name recognition increases as the entity includes more mentions. In other words, if a name has more coref-ed mentions, it is more likely to be correct. This also provides us a linguistic intuition: if people mention an obscure name in a text, they tend to emphasize it later by repeating the same name or describe it with nominal mentions.

The table also indicates that the accuracy of single name entities (singletons) is much lower than the overall accuracy. So, although they constitute only about 10% of all names, increasing their accuracy can significantly improve overall performance. Coreference information can play a great role here. Take the 157 PER singletons as an example; 56% are incorrect names. Among these incorrect names, 73% actually belong to the other two name types. Many of these can be easily fixed by searching for coreference to other mentions without type restriction. Among the correct names, 71% can be confirmed by the presence of a title word or a Chinese last name. From these observations we can conclude that without strong confirmation features, singletons are much less likely to be correct names.

5 Incorporating Coreference Information into Name Recognition

We make use of several features of the coreference relations a name is involved in; the features are listed in Table 3. Using these features, we built an independent classifier to predict if a name identified by the baseline name tagger is correct or not. (Note that this classifier is trained on all name mentions, but during test only 'obscure' names which failed the tests in section 3 are processed by this classifier.) Each name corresponds to a feature vector which consists of the factors described in Table 3. The PER context words are generated from the context patterns described in (Ji and Luo, 2001). We used a Support Vector Machine to implement the classifier, because of its state-of-the-art performance and good generalization ability. We used a polynomial kernel of degree 3.

6 Name Rules based on Coreference

Besides the factors in the above statistical model, additional coreference information can be used to filter and in some cases correct the tagging produced by the HMM. We developed the following rules to correct names generated by the baseline tagger.

6.1 Name Structure Errors

Sometimes the Name tagger outputs names which are too short (incomplete) or too long. We can make use of the relation among mentions in the same entity to fix them. For example, nested ORGs are traditionally difficult to recognize correctly. Errors in ORG names can take the following forms:

(1) *Head Missed*. Examples: "中国艺术（团）/ Chinese Art (Group)", "中国学生（会）/ Chinese Student (Union)", "俄罗斯核动力（所）/ Russian Nuclear Power (Instituition)"

Rule 1: If an ORG name x is coref-ed with other mentions with head y (an ORG suffix), and in the original text x is immediately followed by y, then tag xy instead of x; otherwise discard x.

(2) *Modifier Missed*. Rule 1 can also be used to restore missed modifiers. For example, "（爱丁堡）大学 / (Edinburgh) University"; "（鹏程）有限公司 / (Peng Cheng) Limited Corporation", and some incomplete translated PER names such as "（巴）勒斯坦 / (Pa)lestine".

(3) *Name Too Long*

Rule 2: If a name x has no coref-ed mentions but part of it, x', is identical to a name in another entity y, and y includes at least two mentions; then tag x' instead of x.

.

Rule Type		Rule	Description
Name & Name	All	Ident(i, j)	$Mention_i$ and $Mention_j$ are identical
		Abbrev(i, j)	$Mention_i$ is an abbreviation of $Mention_j$
		Modifier(i, j)	$Mention_j$ = Modifier + "de" + $Mention_i$
		Formal(i, j)	Formal and informal ways of referring to the same entity (Ex. "美国国防部 / American Defense Dept. & 五角大楼/ Pentagon")
	PER	Substring(i, j)	$Mention_i$ is a substring of $Mention_j$
		Title(i, j)	$Mention_j$ = $Mention_i$ + title word; or $Mention_j$ = LastName + title word
	ORG	Head(i, j)	$Mention_i$ and $Mention_j$ have the same head
	GPE	Head(i, j)	$Mention_i$ and $Mention_j$ have the same head
		Capital(i, j)	$Mention_i$: country name; $Mention_j$: name of the capital of this country Applied in restricted context.
		Country(i, j)	$Mention_i$ and $Mention_j$ are different names referring to the same country. (Ex. "中国 / China & 华夏 / Huaxia & 共和国 / Republic")
Name & Nominal	All	RSub(i, j)	$Name_i$ is a right substring of $Nominal_j$
		Apposition(i, j)	$Nominal_j$ is the apposite of $Name_i$
		Modifier2(i, j)	$Nominal_j$ = Determiner/Modifier + $Name_i$/ head
	GPE	Ref(i, j)	$Nominal_j$ = $Name_i$ + GPE Ref Word (examples of GPE Ref Word: "方面 / Side", "政府/Government", "共和国 / Republic", "自治政府/ Municipality")
Nominal& Nominal	All	IdentN(i, j)	$Nominal_i$ and $Nominal_j$ are identical
		Modifier3(i, j)	$Nominal_j$ = Determiner/Modifier + $Nominal_i$

Table1: Main rules used in the Coreference Resolver

Number of mentions per entity / Name Type	1	2	3	4	5	6	7	8	>8
PER	43.94	87.07	91.23	87.95	91.57	91.92	94.74	92.31	97.36
GPE	55.81	88.8	96.07	100	100	100	100	95.83	97.46
ORG	64.71	80.59	89.47	94.29	100	100	--	--	100

Table 2 Accuracy(%) of 'obscure' name recognition

Factor		Description
Coreference Type Weight		Average of weights of coreference relations for which this mention is antecedent: 0.8 for name-name coreference; 0.5 for apposition; 0.3 for other name-nominal coreference
Mention Weight	First Mention	Is first name mention in the entity
	Head	Includes head word of name
	Idiom	Name is part of an idiom
	PER context	For PER Name, has context word in text
	PER title	For PER Name, includes title word
	ORG suffix	For ORG Name, includes suffix word
Entity Weight		Number of mentions in entity / total number of mentions in all entities in document which include a name mention

Table 3 Coreference factors for name recognition

6.2 Name Type Errors

Some names are mistakenly recognized as other name types. For example, the name tagger has difficulty in distinguishing transliterated PER name and transliterated GPE names.

To solve this problem we designed the following rules based on the relation among entities.

Rule 3: If $name_i$ is recognized as type1, the entity it belongs to has only one mention; and $name_j$ is recognized as type2, the entity it belongs to has at least two mentions; and $name_i$ is identical with $name_j$ or $name_i$ is a substring of $name_j$, then correct type1 to type2.

For example, if "克里姆林/ Kremlin" is mistakenly identified as PER, while "克里姆林宫 / Kremlin Palace" is correctly identified as ORG, and in coreference results, "克里姆林 / Kremlin" belongs to a singleton entity, while "克里姆林宫 / Kremlin Palace" has coref-ed mentions, then we correct the type of "克里姆林 / Kremlin" to ORG.

Another common mistake gives rise to the sequence "PER+title+PER", because our name tagger uses the title word as an important context feature for a person name (either preceding or following the title). But this is an impossible structure in Chinese. We can also use coreference information to fix it.

Rule 4: If "PER+title+PER" appears in the name tagger's output, then we discard the PER name with lower coref certainty; and check whether it is coref-ed to other mentions in a GPE entity or ORG entity; if it is, correct the type.

Using this rule we can correctly identify "[斯里兰卡 / Sri Lanka GPE] 总理 / Premier [班达拉耐克 / Bandaranaike PER]", instead of "[斯里兰卡 / Sri Lanka PER] 总理 / Premier [班达拉耐克 / Bandaranaike PER]".

6.3 Name Abbreviation Errors

Name abbreviations are difficult to recognize correctly due to a lack of training data. Usually people adopt a separate list of abbreviations or design separate rules (Sun et al. 2002) to identify them. But many wrong abbreviation names might be produced. We find that coreference information helps to select abbreviations.

Rule 5: If an abbreviation name has no coref-ed mentions and it is not adjacent to another abbreviation (ex. "中/China 美/America"), then we discard it.

7 System Flow

Combining all the methods presented above, the flow of our final system is shown in Figure 2:

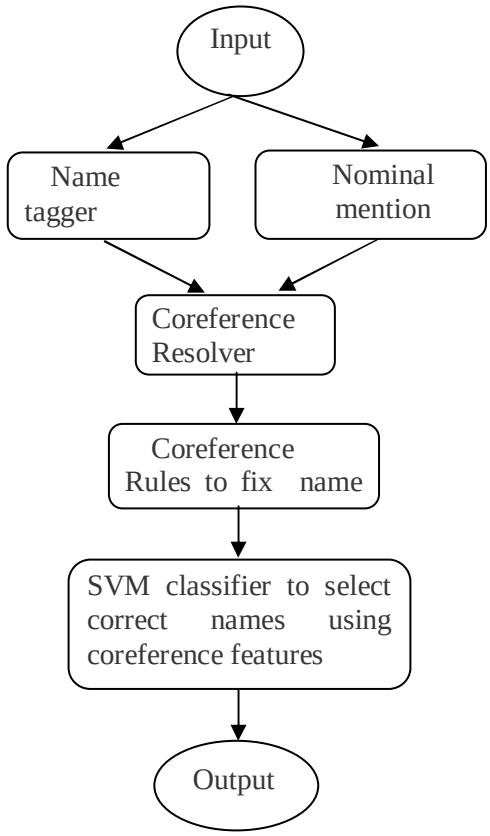

Figure 2 System Flow

8 Experiments

8.1 Training and Test Data

For our experiments, we used the Beijing University Insititute of Computational Linguistics corpus – 2978 documents from *the People's Daily* in 1998, one million words with name tags – and the training corpus for the 2003 ACE evaluation, 223 documents. 153 of our ACE documents were used as our test set.[4] The 153 documents contained 1614 names. Of the system-tagged names, 959 were considered 'obscure': were not on a name list and had a margin below the threshold. These were the names to which the rules and classifier were applied. We ran all the following experiments using the MUC scorer.

[4] The test set was divided into two parts, of 95 documents and 58 documents. We trained two name tagger and classifier models, each time using one part of the test set along with all the other documents, and evaluated on the other part of the test set. The results reported here are the combined results for the entire test set.

8.2 Overall Performance Comparison

Table 4 shows the performance of the baseline system; Table 5 the system with rule-based corrections; and Table 6 the system with both rules and the SVM classifier.

Name	Precision	Recall	F
PER	90.9	88.2	89.5
GPE	82.3	90.8	86.3
ORG	92.1	91.8	91.9
ALL	87.8	90.5	89.1

Table 4 Baseline Name Tagger

Name	Precision	Recall	F
PER	93.3	87.5	90.3
GPE	83.5	90.4	86.8
ORG	90.9	92.1	91.5
ALL	88.5	90.3	89.4

Table 5 Results with Coref Rules Alone

Name	Precision	Recall	F
PER	95.7	84.4	89.7
GPE	88.0	91.7	89.8
ORG	94.5	91.2	92.8
ALL	92.2	89.6	90.9

Table 6 Results for Single Document System

The gains we observed from coreference within single documents suggested that further improvement might be possible by gathering evidence from several related documents.[5] We did this in two stages. First, we clustered the 153 documents in the test set into 38 topical clusters. Most (29) of the clusters had only two documents; the largest had 28 documents. We then applied the same procedures, treating the entire cluster as a single document. This yielded another 1.0% improvement in overall F score (Table 7).

The improvement in F score was consistent for the larger clusters (3 or more documents): the F score improved for 8 of those clusters and remained the same for the 9th. To heighten the multi-document benefit, we took 11 of the small (2 document clusters) and enlarged them by retrieving related documents from `sina.com.cn`. In total, we added 52 texts to these 11 clusters. The net result was a further improvement of 0.3% in F score (Table 8).[6]

Name	Precision	Recall	F
PER	93.3	86.8	90.5
GPE	95.2	90.0	92.5
ORG	92.9	91.7	92.3
ALL	93.8	90.1	91.9

Table 7 Results for Mutiple Document System

Name	Precision	Recall	F
PER	94.7	87.1	90.7
GPE	95.6	89.6	92.5
ORG	95.8	90.3	93.0
ALL	95.4	89.2	92.2

Table 8 Results for Mutiple Document System with additional retrieved texts

8.3 Contribution of Coreference Features

Since feature selection is crucial to SVMs, we did experiments to determine how precision increased as each feature was added. The results are shown in Figure 3. We can see that each feature in the SVM helps to select correct names from the output of the baseline name tagger, although some (like FirstMention) are more crucial than others.

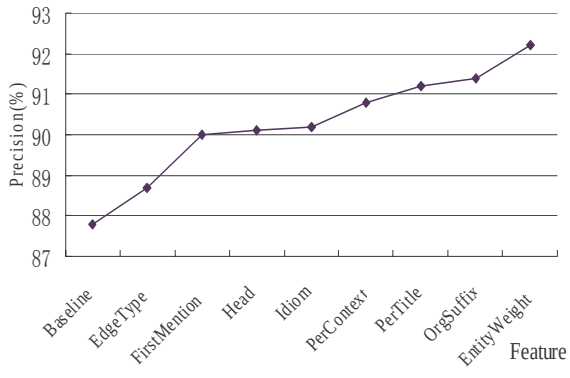

Figure 3 Contributions of features

[5] Borthwick (1999) did use some cross-document information across the entire test corpus, maintaining in effect a name cache for the corpus, in addition to one for the document. No attempt was made to select or cluster documents.

[6] Scores are still computed on the 153 test documents ; the retrieved documents are excluded from the scoring.

8.4 Comparison to Cache Model

Some named entity systems use a name cache, in which tokens or complete names which have been previously assigned a tag are available as features in tagging the remainder of a document. Other systems have made a second tagging pass which uses information on token sequences tagged in the first pass (Borthwick 1999), or have used as features information about features assigned to other instances of the same token (Chieu and Ng 2002). Our system, while more complex, makes use of a richer set of global features, involving the detailed structure of individual mentions, and in particular makes use of both name – name and name – nominal relations.

We have compared the performance of our method (applied to single documents) with a voted cache model, which takes into account the number of times a particular name has been previously assigned each type of tag:

System	Precision	Recall	F
baseline	88.8	90.5	89.1
voted cache	87.6	92.8	90.1
current	92.2	89.6	90.9

Table 9. Comparison with voted cache

Compared to a simple voted cache model, our model provides a greater improvement in name recognition F score; in particular, it can substantially increase the precision of name recognition. The voted cache model can recover some missed names, but at some loss in precision.

9 Conclusions and Future Work

In this paper, we presented a novel idea of applying coreference information to improve name recognition. We used both a statistical filter based on a set of coreference features and rules for correcting specific errors in name recognition. Overall, we obtained an absolute improvement of 3.1% in F score. Put another way, we were able to eliminate about 60% of erroneous name tags with only a small loss in recall.

The methods were tested on a Chinese name tagger, but most of the techniques should be applicable to other languages. More generally, it offers an example of using global and cross-document information to improve local decisions for information extraction. Such methods will be important for breaking the 'performance ceiling' in many areas of information extraction.

In the future, we plan to experiment with improvements in coreference resolution (in particular, adding pronoun resolution) to see if we can obtain further gains in name recognition. We also intend to explore the production of *multiple* tagging hypotheses by our statistical name tagger, with the alternative hypotheses then reranked using global information. This may allow us to replace some of our hand-coded error-correction rules with corpus-trained methods.

10 Acknowledgements

This research was supported by the Defense Advanced Research Projects Agency as part of the Translingual Information Detection, Extraction and Summarization (TIDES) program, under Grant N66001-001-1-8917 from the Space and Naval Warfare Systems Center San Diego, and by the National Science Foundation under Grants IIS-0081962 and 0325657. This paper does not necessarily reflect the position or the policy of the U.S. Government.

References

Daniel M. Bikel, Scott Miller, Richard Schuartz, and Ralph Weischedel. 1999. Nymble: a high-performance Learning Name-finder. *Proc. Fifth Conf. On Applied Natural Language Processing*, Washington, D.C.

Andrew Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition.* Ph.D. Dissertation, Dept. of Computer Science, New York University.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition. *Proc. Sixth Workshop on Very Large Corpora*, Montreal, Canada.

Hai Leong Chieu and Hwee Tou Ng. 2002. Named Entity Recognition: A Maximum Entropy Approach Using Global Information. *Proc.: 17th Int'l Conf. on Computational Linguistics (COLING 2002),* Taipei, Taiwan.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference - 6: A brief history. *Proc. 16th Int'l Conference on Computational Linguistics (COLING 96),* Copenhagen.

Heng Ji, Zhensheng Luo, 2001. A Chinese Name Identifying System Based on Inverse Name Frequency Model and Rules. *Natural Language Processing and Knowledge Engineering (NLPKE) Mini Symposium of 2001 IEEE*

International Conference on Systems, Man, and Cybernetics (SMC2001)

Andrew McCallum and Wei Li. 2003. Early results for Named Entity Recognition With Conditional Random Fields, Feature Induction, and Web-Enhanced Lexicons. *Proc. Seventh Conf. on Computational Natural Language Learning (CONLL-2003)*, Edmonton, Canada.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active Hidden Markov Models for Information Extraction. *Proc. Int'l Symposium on Intelligent Data Analysis (IDA-2001)*.

Satoshi Sekine, Ralph Grishman and Hiroyuki Shinnou. 1998. A Decision Tree Method for Finding and Classifying Names in Japanese Texts. *Proc. Sixth Workshop on Very Large Corpora*; Montreal, Canada.

Jian Sun, Jianfeng Gao, Lei Zhang, Ming Zhou and Changning Huang. 2002. Chinese Named Entity Identification Using Class-based Language Model. *Coling 2002.*

Marc Vilain, John Burger, John Aberdeen, Dennis Connelly, Lynette Hirschman. 1995. A model -- Theoretic Coreference Scoring Scheme. *MUC-6 Proceedings*, Nov. 1995.

Shiren Ye, Tat-Seng Chua, Liu Jimin. 2002. An Agent-based Approach to Chinese Named Entity Recognition. *Coling 2002.*