A NP-Cluster Based Approach to Coreference Resolution

Xiaofeng Yang[†‡] **Jian Su**[†] **Guodong Zhou**[†] **Chew Lim Tan**[‡]

[†]Institute for Infocomm Research
21 Heng Mui Keng Terrace,
Singapore, 119613
{xiaofengy,sujian,zhougd}
@i2r.a-star.edu.sg

[‡] Department of Computer Science
National University of Singapore,
Singapore, 117543
{yangxiao,tancl}@comp.nus.edu.sg

Abstract

Traditionally, coreference resolution is done by mining the reference relationships between NP pairs. However, an individual NP usually lacks adequate description information of its referred entity. In this paper, we propose a supervised learning-based approach which does coreference resolution by exploring the relationships between NPs and coreferential clusters. Compared with individual NPs, coreferential clusters could provide richer information of the entities for better rules learning and reference determination. The evaluation done on MEDLINE data set shows that our approach outperforms the baseline NP-NP based approach in both recall and precision.

1 Introduction

Coreference resolution is the process of linking as a cluster[1] multiple expressions which refer to the same entities in a document. In recent years, supervised machine learning approaches have been applied to this problem and achieved considerable success (e.g. Aone and Bennett (1995); McCarthy and Lehnert (1995); Soon et al. (2001); Ng and Cardie (2002b)). The main idea of most supervised learning approaches is to recast this task as a binary classification problem. Specifically, a classifier is learned and then used to determine whether or not two NPs in a document are co-referring. Clusters are formed by linking coreferential NP pairs according to a certain selection strategy. In this way, the identification of coreferential clusters in text is reduced to the identification of coreferential NP pairs.

One problem of such reduction, however, is that the individual NP usually lacks adequate descriptive information of its referred entity. Consequently, it is often difficult to judge whether or not two NPs are talking about the same entity simply from the properties of the pair alone. As an example, consider the pair of a non-pronoun and its pronominal antecedent candidate. The pronoun itself gives few clues for the reference determination. Using such NP pairs would have a negative influence for rules learning and subsequent resolution. So far, several efforts (Harabagiu et al., 2001; Ng and Cardie, 2002b; Ng and Cardie, 2002a) have attempted to address this problem by discarding the "hard" pairs and select only those confident ones from the NP-pair pool. Nevertheless, this eliminating strategy still can not guarantee that the NPs in "confident" pairs bear necessary description information of their referents.

In this paper, we present a supervised learning-based approach to coreference resolution. Rather than attempting to mine the reference relationships between NP pairs, our approach does resolution by determining the links of NPs to the existing coreferential clusters. In our approach, a classifier is trained on the instances formed by a NP and one of its possible antecedent clusters, and then applied during resolution to select the proper cluster for an encountered NP to be linked. As a coreferential cluster offers richer information to describe an entity than a single NP in the cluster, we could expect that such a NP-Cluster framework would enhance the resolution capability of the system. Our experiments were done on the the MEDLINE data set. Compared the baseline approach based on NP-NP framework, our approach yields an improvement of system recall by 4.6%, with still a gain of precision by 1.3%. These results indicate that the NP-Cluster based approach is effective for the coreference resolution task.

The remaining of this paper is organized as follows. Section 2 introduces as the baseline the NP-NP based approach, while Section 3 presents in details our NP-Cluster based approach. Section 4 reports and discusses the experimental results. Section 5 describes related research work.

[1]In this paper the term "cluster" can be interchangeably used as "chain", while the former better emphasizes the equivalence property of coreference relationship.

Finally, conclusion is given in Section 6.

2 Baseline: the NP-NP based approach

2.1 Framework description

We built a baseline coreference resolution system, which adopts the common NP-NP based learning framework as employed in (Soon et al., 2001).

Each instance in this approach takes the form of $\boldsymbol{i}(\mathrm{NP}_j, \mathrm{NP}_i)$, which is associated with a feature vector consisting of 18 features ($f_1 \sim f_{18}$) as described in Table 2. Most of the features come from Soon et al. (2001)'s system. Inspired by the work of (Strube et al., 2002) and (Yang et al., 2004), we use two features, *StrSim1* (f_{17}) and *StrSim2* (f_{18}), to measure the string-matching degree of NP_j and NP_i. Given the following similarity function:

$$Str\_Simlarity(Str_1, Str_2) = 100 \times \frac{|Str_1 \cap Str_2|}{Str_1}$$

StrSim1 and *StrSim2* are computed using Str_Similarity(S_{NP_j}, S_{NP_i}) and Str_Similarity(S_{NP_i}, S_{NP_j}), respectively. Here S_{NP} is the token list of *NP*, which is obtained by applying word stemming, stopword removal and acronym expansion to the original string as described in Yang et al. (2004)'s work.

During training, for each anaphor NP_j in a given text, a positive instance is generated by pairing NP_j with its closest antecedent. A set of negative instances is also formed by NP_j and each NP occurring between NP_j and NP_i.

When the training instances are ready, a classifier is learned by C5.0 algorithm (Quinlan, 1993). During resolution, each encountered noun phrase, NP_j, is paired in turn with each preceding noun phrase, NP_i. For each pair, a testing instance is created as during training, and then presented to the decision tree, which returns a confidence value (CF)[2] indicating the likelihood that NP_i is coreferential to NP_j. In our study, two antecedent selection strategies, Most Recent First (MRF) and Best First (BF), are tried to link NP_j to its a proper antecedent with CF above a threshold (0.5). MRF (Soon et al., 2001) selects the candidate closest to the anaphor, while BF (Aone and Bennett, 1995; Ng and Cardie, 2002b) selects the candidate with the maximal CF.

2.2 Limitation of the approach

Nevertheless, the problem of the NP-NP based approach is that the individual NP usually lacks adequate description information about its referred entity. Consequently, it is often difficult to determine whether or not two NPs refer to the same entity simply from the properties of the pair. See the the text segment in Table 1, for example,

... [$_1$ The hepatitis B virus] encodes [$_2$ a transcriptional transactivator protein] ...
... We have investigated whether [$_3$ this antigen] is [$_4$ a target structure] for [$_5$ human T-lymphocytes] ...
... We found that [$_6$ hepatitis B virus] ...

Table 1: An Example from the data set

In the above text, [$_1$ The hepatitis B virus], [$_3$ this antigen] and [$_6$ hepatitis B virus] are annotated in the same coreferential cluster. According to the above framework, NP_6 and its closest antecedent, NP_3, will form a positive instance. Nevertheless, such an instance is not informative in that NP_3 bears little information related to the entity and thus provides few clues to explain its coreference relationship with NP_6.

In fact, this relationship would be clear if [$_1$ The hepatitis B virus], the antecedent of NP_3, is taken into consideration. NP_1 gives a detailed description of the entity. By comparing the string of NP_6 with this description, it is apparent that NP_6 belongs to the cluster of NP_1, and thus should be coreferential to NP_3. This suggests that we use the coreferential cluster, instead of its single element, to resolve a NP correctly. In our study, we propose an approach which adopts a NP-Cluster based framework to do resolution. The details of the approach are given in the next section.

3 The NP-Cluster based approach

Similar to the baseline approach, our approach also recasts coreference resolution as a binary classification problem. The difference, however, is that our approach aims to learn a classifier which would select the most preferred cluster, instead of the most preferred antecedent, for an encountered NP in text. We will give the framework of the approach, including the instance representation, the training and the resolution procedures, in the following subsections.

[2]The confidence value is obtained by using the smoothed ratio $\frac{p+1}{t+2}$, where p is the number of positive instances and t is the total number of instances contained in the corresponding leaf node.

Features describing the relationships between NP_j and NP_i		
1.	DefNp_1	1 if NP_j is a definite NP; else 0
2.	DemoNP_1	1 if NP_j starts with a demonstrative; else 0
3.	IndefNP_1	1 if NP_j is an indefinite NP; else 0
4.	Pron_1	1 if NP_j is a pronoun; else 0
5.	ProperNP_1	1 if NP_j is a proper NP; else 0
6.	DefNP_2	1 if NP_i is a definite NP; else 0
7.	DemoNP_2	1 if NP_i starts with a demonstrative; else 0
8.	IndefNP_2	1 if NP_i is an indefinite NP; else 0
9.	Pron_2	1 if NP_i is a pronoun; else 0
10.	ProperNP_2	1 if NP_i is a proper NP; else 0
11.	Appositive	1 if NP_i and NP_j are in an appositive structure; else 0
12.	NameAlias	1 if NP_i and NP_j are in an alias of the other; else 0
13.	GenderAgree	1 if NP_i and NP_j agree in gender; else 0
14.	NumAgree	1 if NP_i and NP_j agree in number; else 0
15.	SemanticAgree	1 if NP_i and NP_j agree in semantic class; else 0
16.	HeadStrMatch	1 if NP_i and NP_j contain the same head string; else 0
17.	StrSim_1	The string similarity of NP_j against NP_i
18.	StrSim_2	The string similarity of NP_i against NP_j
Features describing the relationships between NP_j and cluster C_k		
19.	Cluster_NumAgree	1 if C_k and NP_j agree in number; else 0
20.	Cluster_GenAgree	1 if C_k and NP_j agree in gender; else 0
21.	Cluster_SemAgree	1 if C_k and NP_j agree in semantic class; else 0
22.	Cluster_Length	The number of elements contained in C_k
23.	Cluster_StrSim	The string similarity of NP_j against C_k
24.	Cluster_StrLNPSim	The string similarity of NP_j against the longest NP in C_k

Table 2: The features used in our coreference resolution system (Features $f1 \sim 18$ are also used in the baseline system using NP-NP based approach)

3.1 Instance representation

An instance in our approach is composed of three elements like below:

$\boldsymbol{i}\{NP_j, C_k, NP_i\}$

where NP_j, like the definition in the baseline, is the noun phrase under consideration, while C_k is an existing coreferential cluster. Each cluster could be referred by a reference noun phrase NP_i, an element of the cluster. A cluster would probably contain more than one reference NPs and thus may have multiple associated instances.

For a training instance, the label is positive if NP_j is annotated as belonging to C_k, or negative if otherwise.

In our system, each instance is represented as a set of 24 features as shown in Table 2. The features are supposed to capture the properties of NP_j and C_k as well as their relationships. In the table we divide the features into two groups, one describing NP_j and NP_i and the other describing NP_j and C_k. For the former group, we just use the same features set as in the baseline system, while for the latter, we introduce 6 more features:

Cluster_NumAgree, Cluster_GenAgree and Cluster_SemAgree: These three features mark the compatibility of NP_j and C_k in number, gender and semantic agreement, respectively. If NP_j mismatches the agreement with any element in C_k, the corresponding feature is set to 0.

Cluster_Length: The number of NPs in the cluster C_k. This feature reflects the global salience of an entity in the sense that the more frequently an entity is mentioned, the more important it would probably be in text.

Cluster_StrSim: This feature marks the string similarity between NP_j and C_k. Suppose S_{NP_j} is the token set of NP_j, we calculate the feature value using the similarity function Str_Similarity(S_{NP_j}, S_{C_k}), where

$$S_{C_k} = \bigcup_{NP_i \in C_k} S_{NP_i}$$

Cluster_StrLNPSim: It marks the string matching degree of NP_j and the noun phrase in C_k with the most number of tokens. The intuition here is that the NP with the longest string bears the most description information of the refereed entity. The feature is calculated using the similarity function Str_Similarity(S_{NP_j}, S_{NP_k}), where

$NP_k = \arg\max_{NP_i \in C_k} |S_{NP_i}|$

3.2 Training procedure

Given an annotated training document, we process the noun phrases from beginning to end. For each anaphoric noun phrase NP_j, we consider its preceding coreferential clusters from right to left[3]. For each cluster, we create only one instance by taking the last NP in the cluster as the reference NP. The process will not terminate until the cluster to which NP_j belongs is found.

To make it clear, consider the example in Table 1 again. For the noun phrase [$_6$ hepatitis B virus], the annotated preceding coreferential clusters are:

C1:{[$_5$ human T-lymphocytes]}
C2:{[$_4$ a target structure]}
C3:{[$_1$ The hepatitis B virus], [$_3$ this antigen]}
C4:{[$_2$ a transcriptional transactivator protein]}

Thus three training instances are generated:

$\boldsymbol{i}$(NP_6, C1, [$_5$ human T-lymphocytes])
$\boldsymbol{i}$(NP_6, C2, [$_4$ a target structure])
$\boldsymbol{i}$(NP_6, C3, [$_3$ this antigen])

Among them, the first two instances are labelled as negative while the last one is positive.

After the training instances are ready, we use C5.0 learning algorithm to learn a decision tree classifier as in the baseline approach.

3.3 Resolution procedure

The resolution procedure is the counterpart of the training procedure. Given a testing document, for each encountered noun phrase, NP_j, we create a set of instances by pairing NP_j with each cluster found previously. The instances are presented to the learned decision tree to judge the likelihood that NP_j is linked to a cluster. The resolution algorithm is given in Figure 1.

As described in the algorithm, for each cluster under consideration, we create multiple instances by using every NP in the cluster as the reference NP. The confidence value of the cluster is the maximal confidence value of its instances. Similar to the baseline system, two cluster selection strategies, i.e. MRF and BF, could be applied to link NP_j to a proper cluster. For MRF strategy, NP_j is linked to the closest cluster with confidence value above 0.5, while for BF, it is linked to the cluster with the maximal confidence value (above 0.5).

[3]We define the position of a cluster as the position of the last NP in the cluster.

algorithm RESOLVE (a testing document d)
ClusterSet = $\emptyset$;
//suppose d has N markable NPs;
for j = 1 **to** N
  foreach cluster **in** ClusterSet
    $CF_{cluster} = \max_{NP_i \in cluster} CF_{i(NP_j, cluster, NP_i)}$
  select a proper cluster, *BestCluster*, according to a ceterin cluster selection strategy;
  if *BestCluster* != NULL
    $BestCluster = BestCluster \cup \{NP_j\}$;
  else
    //create a new cluster
    $NewCluster = \{ NP_j \}$;
    ClusterSet = ClusterSet $\cup$ {*NewCluster*};

Figure 1: The clusters identification algorithm

3.4 Comparison of NP-NP and NP-Cluster based approaches

As noted above, the idea of the NP-Cluster based approach is different from the NP-NP based approach. However, due to the fact that in our approach a cluster is processed based on its reference NPs, the framework of our approach could be reduced to the NP-NP based framework if the cluster-related features were removed. From this point of view, this approach could be considered as an extension of the baseline approach by applying additional cluster features as the properties of NP_i. These features provide richer description information of the entity, and thus make the coreference relationship between two NPs more apparent. In this way, both rules learning and coreference determination capabilities of the original approach could be enhanced.

4 Evaluation

4.1 Data collection

Our coreference resolution system is a component of our information extraction system in biomedical domain. For this purpose, we annotated 100 MEDLINE[4] abstracts selected from the GENIA[5] data set. The annotation scheme follows that of MUC-6 (1995) and MUC-7 (1998). Each document in the corpus has an average length of 232 words. One characteristic

[4]http://www.medstract.org
[5]http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/index.html

	MRF			BF		
Experiments	R	P	F	R	P	F
Baseline	80.2	77.4	78.8	80.3	77.5	78.9
AllAnte	84.4	70.2	76.6	**85.7**	71.4	77.9
Our Approach	84.4	78.2	81.2	84.9	**78.8**	**81.7**

Table 3: The performance of different coreference resolution systems

of the bio-literature is that pronouns only occupy about 3% among all the NPs. This ratio is quite low compared to that in newswire domain (e.g. above 10% for MUC data set).

A pipeline of NLP components is applied to pre-process an input raw text. Among them, NE recognition, part-of-speech tagging and text chunking adopt the same HMM based engine with error-driven learning capability (Zhou and Su, 2000; Zhou and Su, 2002). The NE recognition component trained on GENIA (Shen et al., 2003) can recognize up to 23 common biomedical entity types with an overall performance of 66.1 F-measure (P=66.5% R=65.7%). In addition, to remove the apparent non-anaphors (e.g., embedded proper nouns) in advance, a heuristic-based anaphoricity identification module is applied, which successfully removes 50.0% non-anaphors with a precision of 83.5% for our data set.

4.2 Experiments and discussions

Our experiments were done on the 100 annotated documents, among them 70 for training and the other 30 for testing. Throughout these experiments, default learning parameters were applied in the C5.0 algorithm. The recall and precision were calculated automatically according to the scoring scheme proposed by Vilain et al. (1995).

In Table 3 we compared the performance of different coreference resolution systems. The first line summarizes the results of the baseline system using traditional NP-NP based approach as described in Section 2. Using BF strategy, *Baseline* obtains 80.3% recall and 77.5% precision. These results are better than the work by Castano et al. (2002) and Yang et al. (2004), which were also tested on the MEDLINE data set and reported a F-measure of about 74 and 69, respectively.

In the experiments, we evaluated another NP-NP based system, *AllAnte*. It adopts a similar learning framework as *Baseline* except that during training it generates the positive instances by paring a NP with all its antecedents instead of only the closest one. The system attempts to use such an instance selection strategy to incorporate the information from coreferential clusters. But the results are nevertheless disappointing: although this strategy boosts the recall by 5.4%, the precision drops considerably by above 6% at the same time. The overall F-measure is even lower than the baseline systems.

The last line of Table 3 demonstrates the results of our NP-Cluster based approach. For BF strategy, the system achieves 84.9% recall and 78.8% precision. As opposed to the baseline system, the recall rises by 4.6% while the precision still gains slightly by 1.3%. Overall, we observe the increase of F-measure by 2.8.

The results in Table 3 also indicate that the BF strategy is superior to the MRF strategy. A similar finding was also reported by Ng and Cardie (2002b) in the MUC data set.

To gain insight into the difference in the performance between our NP-Cluster based system and the NP-NP based system, we compared the decision trees generated in the two systems in Figure 2. In both trees, the string-similarity features occur on the top portion, which supports the arguments by (Strube et al., 2002) and (Yang et al., 2004) that string-matching is a crucial factor for NP coreference resolution. As shown in the figure, the feature *StrSim_1* in left tree is completely replaced by the *Cluster_StrSim* and *Cluster_StrLNPSim* in the right tree, which means that matching the tokens with a cluster is more reliable than with a single NP. Moreover, the cluster length will also be checked when the NP under consideration has low similarity against a cluster. These evidences prove that the information from clusters is quite important for the coreference resolution on the data set.

The decision tree visualizes the importance of the features for a data set. However, the tree is learned from the documents where coreferential clusters are correctly annotated. During resolution, unfortunately, the found clusters are usually not completely correct, and as a result the features important in training data may not be also helpful for testing data. Therefore, in the experiments we were concerned about which fea-

```
HeadMatch = 0:
:...NameAlias = 1: 1 (22/1)
:   NameAlias = 0:
:   :...Appositive = 0: 0 (13095/265)
:       Appositive = 1: 1 (15/4)
HeadMatch = 1:
:...StrSim_1 > 71:
    :...DemoNP_1 = 0: 1 (615/29)
    :   DemoNP_1 = 1:
    :   :...NumAgree = 0: 0 (5)
    :       NumAgree = 1: 1 (26)
    StrSim_1 <= 71:
    :...DemoNP_2 = 1: 1 (12/2)
        DemoNP_2 = 0:
        :...StrSim_2 <= 77: 0 (144/17)
            StrSim_2 > 77:
            :...StrSim_1 <= 33: 0 (42/11)
                StrSim_1 > 33: 1 (38/11)
```

```
HeadMatch = 1:
:...Cluster_StrSim > 66: 1 (663/36)
:   Cluster_StrSim <= 66:
:   :...StrSim_2 <= 85: 0 (140/14)
:       StrSim_2 > 85:
:       :...Cluster_StrLNPSim > 50: 1 (16/1)
:           Cluster_StrLNPSim <= 50:
:           :...Cluster_Length <= 5: 0 (59/17)
:               Cluster_Length > 5: 1 (4)
HeadMatch = 0:
:...NameAlias = 1: 1 (22/1)
    NameAlias = 0:
    :...Appositive = 1: 1 (15/4)
        Appositive = 0:
        :...StrSim_2 <= 54:
            :..
            StrSim_2 > 54:
            :..
```

Figure 2: The resulting decision trees for the NP-NP and NP-Cluster based approaches

Features	R	P	F
$f_{1\sim21}$	80.3	77.5	78.9
$f_{1\sim21}, f_{22}$	84.1	74.4	79.0
$f_{1\sim21}, f_{23}$	84.7	78.8	81.6
$f_{1\sim21}, f_{24}$	84.3	78.0	81.0
$f_{1\sim21}, f_{23}, f_{22}$	84.9	78.6	81.6
$f_{1\sim21}, f_{23}, f_{24}$	**84.9**	**78.9**	**81.8**
$f_{1\sim21}, f_{23}, f_{24}, f_{22}$	**84.9**	78.8	81.7

Table 4: Performance using combined features (f_i refers to the i(th) feature listed in Table 2)

tures really matter for the real coreference resolution. For this purpose, we tested our system using different features and evaluated their performance in Table 4. Here we just considered feature *Cluster_Length* (f_{22}), *Cluster_StrSim* (f_{23}) and *Cluster_StrLNPSim* (f_{24}), as Figure 2 has indicated that among the cluster-related features only these three are possibly effective for resolution. Throughout the experiment, the Best-First strategy was applied.

As illustrated in the table, we could observe that:

1. Without the three features, the system is equivalent to the baseline system in terms of the same recall and precision.
2. *Cluster_StrSim* (f_{23}) is the most effective as it contributes most to the system performance. Simply using this feature boosts the F-measure by 2.7.
3. *Cluster_StrLNPSim* (f_{24}) is also effective by improving the F-measure by 2.1 alone. When combined with f_{23}, it leads to the best F-measure.
4. *Cluster_Length* (f_{22}) only brings 0.1 F-measure improvement. It could barely increase, or even worse, reduces the F-measure when used together with the the other two features.

5 Related work

To our knowledge, our work is the first supervised-learning based attempt to do coreference resolution by exploring the relationship between a NP and coreferential clusters. Cardie and Wagstaff (1999) proposes an unsupervised approach which also incorporates cluster information into consideration. Compared with Cardie and Wagstaff (1999)'s work, our approach has the following advantages:

1. Their approach uses hard constraints to preclude the link of a NP to a cluster mismatching the number, gender or semantic agreements. By contrast, in our approach these agreements together with other features (e.g. cluster-length, string-matching degree,etc) are used as preference factors for cluster selection. Comparatively, the preference based strategy is more reliable for coreference resolution (Mitkov, 1999).
2. Their approach links a NP under consideration to every compatible cluster if their shortest distance is below a threshold, and the link can never be undone. By contrast, our approach would rank all cluster candidates based on their confidence values, and then link the NP to the most preferred one. Comparatively, the algorithm of our approach is less aggressively greedy.

6 Conclusion

In this paper we have proposed a supervised learning-based approach to coreference resolution. Rather than mining the coreferential relationship between NP pairs as in conventional approaches, our approach does resolution by exploring the relationships between a NP and the coreferential clusters. Compared to individual NPs, coreferential clusters provide more information for rules learning and reference determination. In the paper, we first introduced the conventional NP-NP based approach and analyzed its limitation. Then we described in details the framework of our NP-Cluster based approach, including the instance representation, training and resolution procedures. We evaluated our approach in the biomedical domain, and the experimental results showed that our approach outperforms the NP-NP based approach in both recall (4.6%) and precision (1.3%).

While our approach achieves better performance, there is still room for further improvement. For example, the approach just resolves a NP using the cluster information available so far. Nevertheless, the text after the NP would probably give important supplementary information of the clusters. The ignorance of such information may affect the correct resolution of the NP. In the future work, we plan to work out more robust clustering algorithm to link a NP to a globally best cluster.

References

C. Aone and S. W. Bennett. 1995. Evaluating automated and manual acquistion of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the Association for Compuational Linguistics*, pages 122–129.

C. Cardie and K. Wagstaff. 1999. Noun phrase coreference as clustering. In *Proceedings of the Joint Conference on Empirical Methods in NLP and Very Large Corpora.*

J. Castano, J. Zhang, and J. Pustejovsky. 2002. Anaphora resolution in biomedical literature. In *International Symposium on Reference Resolution*, Alicante, Spain.

S. Harabagiu, R. Bunescu, and S. Maiorano. 2001. Text knowledge mining for coreference resolution. In *Proceedings of the 2nd Annual Meeting of the North America Chapter of the Association for Compuational Linguistics*, pages 55–62.

J. McCarthy and Q. Lehnert. 1995. Using decision trees for coreference resolution. In *Proceedings of the 14th International Conference on Artificial Intelligences*, pages 1050–1055.

R. Mitkov. 1999. Anaphora resolution: The state of the art. Technical report, University of Wolverhampton.

MUC-6. 1995. *Proceedings of the Sixth Message Understanding Conference.* Morgan Kaufmann Publishers, San Francisco, CA.

MUC-7. 1998. *Proceedings of the Seventh Message Understanding Conference.* Morgan Kaufmann Publishers, San Francisco, CA.

V. Ng and C. Cardie. 2002a. Combining sample selection and error-driven pruning for machine learning of coreference rules. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pages 55–62, Philadelphia.

V. Ng and C. Cardie. 2002b. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, Philadelphia.

J. R. Quinlan. 1993. *C4.5: Programs for machine learning.* Morgan Kaufmann Publishers, San Francisco, CA.

D. Shen, J. Zhang, G. Zhou, J. Su, and C. Tan. 2003. Effective adaptation of hidden markov model-based named-entity recognizer for biomedical domain. In *Proceedings of ACL03 Workshop on Natural Language Processing in Biomedicine*, Japan.

W. Soon, H. Ng, and D. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

M. Strube, S. Rapp, and C. Muller. 2002. The influence of minimum edit distance on reference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 312–319, Philadelphia.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message understanding Conference (MUC-6)*, pages 45–52, San Francisco, CA. Morgan Kaufmann Publishers.

X. Yang, G. Zhou, J. Su, and C. Tan. 2004. Improving noun phrase coreference resolution by matching strings. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*, Hainan.

G. Zhou and J. Su. 2000. Error-driven HMM-based chunk tagger with context-dependent lexicon. In *Proceedings of the Joint Conference on Empirical Methods on Natural Language Processing and Very Large Corpus*, Hong Kong.

G. Zhou and J. Su. 2002. Named Entity recognition using a HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia.