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 cluster1 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-refering. 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.
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 \( i(\text{NP}_j, \text{NP}_i) \), 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, \( \text{StrSim}_1 (f_{17}) \) and \( \text{StrSim}_2 (f_{18}) \), to measure the string-matching degree of \( \text{NP}_j \) and \( \text{NP}_i \). Given the following similarity function:

\[
\text{Str}_\text{Similarity}(\text{Str}_1, \text{Str}_2) = 100 \times \frac{|\text{Str}_1 \cap \text{Str}_2|}{\text{Str}_1}
\]

\( \text{Str}_1 \) and \( \text{Str}_2 \) are computed using \( \text{Str}_\text{Similarity}(S_{\text{NP}_j}, S_{\text{NP}_i}) \) and \( \text{Str}_\text{Similarity}(S_{\text{NP}_j}, S_{\text{NP}_i'}) \), respectively. Here \( S_{\text{NP}_i} \) is the token list of \( \text{NP}_i \), 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 \( \text{NP}_j \) in a given text, a positive instance is generated by pairing \( \text{NP}_j \) with its closest antecedent. A set of negative instances is also formed by \( \text{NP}_j \) and each \( \text{NP} \) occurring between \( \text{NP}_j \) and \( \text{NP}_i \).

When the training instances are ready, a classifier is learned by C5.0 algorithm (Quinlan, 1993). During resolution, each encountered noun phrase, \( \text{NP}_j \), is paired in turn with each preceding noun phrase, \( \text{NP}_i \). 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 \( \text{NP}_i \) is coreferential to \( \text{NP}_j \). In our study, two antecedent selection strategies, Most Recent First (MRF) and Best First (BF), are tried to link \( \text{NP}_j \) 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,

\[
\ldots [1 \text{ The hepatitis B virus] encodes] 2 \text{ a transcriptional transactivator protein} \ldots
\]

\[
\ldots \text{We have investigated whether} [3 \text{ this antigen}] \text{ is} [4 \text{ a target structure} \text{ for} [5 \text{ human T-lymphocytes}] \ldots
\]

\[
\ldots \text{We found that} [6 \text{ hepatitis B virus}] \ldots
\]

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, \( \text{NP}_6 \) and its closest antecedent, \( \text{NP}_3 \), will form a positive instance. Nevertheless, such an instance is not informative in that \( \text{NP}_3 \) bears little information related to the entity and thus provides few clues to explain its coreference relationship with \( \text{NP}_6 \).

In fact, this relationship would be clear if [1 The hepatitis B virus], the antecedent of \( \text{NP}_3 \), is taken into consideration. \( \text{NP}_1 \) gives a detailed description of the entity. By comparing the string of \( \text{NP}_6 \) with this description, it is apparent that \( \text{NP}_6 \) belongs to the cluster of \( \text{NP}_1 \), and thus should be coreferential to \( \text{NP}_3 \). 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.
Features describing the relationships between NP\(_j\) and NP\(_i\):

1. **DefNp**\(_j\) 1 if NP\(_j\) is a definite NP; else 0
2. **DemoNP**\(_j\) 1 if NP\(_j\) starts with a demonstrative; else 0
3. **IndefNP**\(_j\) 1 if NP\(_j\) is an indefinite NP; else 0
4. **Pron**\(_j\) 1 if NP\(_j\) is a pronoun; else 0
5. **ProperNP**\(_j\) 1 if NP\(_j\) is a proper NP; else 0
6. **DefNP**\(_i\) 1 if NP\(_i\) is a definite NP; else 0
7. **DemoNP**\(_i\) 1 if NP\(_i\) starts with a demonstrative; else 0
8. **IndefNP**\(_i\) 1 if NP\(_i\) is an indefinite NP; else 0
9. **Pron**\(_i\) 1 if NP\(_i\) is a pronoun; else 0
10. **ProperNP**\(_i\) 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**\(_j\) The string similarity of NP\(_j\) against NP\(_i\)
18. **StrSim**\(_i\) 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_Len** 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:

\[
i\{\text{NP}_j, C_k, \text{NP}_i\}
\]

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_Len**: 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 \(\text{Str\ Similarity}(S_{NP_j}, S_{C_k})\), 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 intu-
ition here is that the NP with the longest string
bears the most description information of the re-
ferred 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 pro-
cess 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 in-
stance 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 Ta-
ble 1 again. For the noun phrase \([6 \text{ hepatitis B virus}]\), the annotated preceding coreferential
clusters are:

- C1: \([5 \text{ human T-lymphocytes}]\)
- C2: \([4 \text{ a target structure}]\)
- C3: \([1 \text{ The hepatitis B virus}, [3 \text{ this antigen}]\)
- C4: \([2 \text{ a transcriptional transactivator protein}]\)

Thus three training instances are generated:

- $i(NP_6, C_1, [5 \text{ human T-lymphocytes}] )$
- $i(NP_6, C_2, [4 \text{ a target structure}] )$
- $i(NP_6, C_3, [3 \text{ this antigen}] )$

Among them, the first two instances are la-
belled 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 doc-
ument, 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 clus-
ter under consideration, we create multiple in-
stances 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 selec-
tion strategies, i.e. MRF and BF, could be ap-
plicated to link $NP_j$ to a proper cluster. For MRF

---

\(^3\)We define the position of a cluster as the position of
the last NP in the cluster.

---

```plaintext
algorithm RESOLVE (a testing document d)
ClusterSet = \{\};
//suppose d has N markable NPs;
for \(j = 1 \text{ to } N\)
foreach cluster in ClusterSet
$$CF_{cluster} = \max_{NP_i \in \text{cluster}} CF_i(NP_j, \text{cluster}, NP_i)$$
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

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).

### 3.4 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 ap-
proach. However, due to the fact that in our
approach a cluster is processed based on its re-
ference 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 con-
sidered as an extension of the baseline approach
by applying additional cluster features as the
properties of NP. 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 capabili-
ties of the original approach could be enhanced.

### 4 Evaluation

#### 4.1 Data collection

Our coreference resolution system is a com-
ponent of our information extraction system
in biomedical domain. For this purpose, we
notated 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
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-
Figure 2: The resulting decision trees for the NP-NP and NP-Cluster based approaches

<table>
<thead>
<tr>
<th>Features</th>
<th>R</th>
<th>P</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{1-21}$</td>
<td>80.3</td>
<td>77.5</td>
<td>78.9</td>
</tr>
<tr>
<td>$f_{1-21}, f_{22}$</td>
<td>84.1</td>
<td>74.4</td>
<td>79.0</td>
</tr>
<tr>
<td>$f_{1-21}, f_{23}$</td>
<td>84.7</td>
<td>78.8</td>
<td>81.6</td>
</tr>
<tr>
<td>$f_{1-21}, f_{24}$</td>
<td>84.3</td>
<td>78.0</td>
<td>81.0</td>
</tr>
<tr>
<td>$f_{1-21}, f_{23}, f_{22}$</td>
<td>84.9</td>
<td>78.9</td>
<td>81.8</td>
</tr>
<tr>
<td>$f_{1-21}, f_{23}, f_{24}, f_{22}$</td>
<td>84.9</td>
<td>78.8</td>
<td>81.7</td>
</tr>
</tbody>
</table>

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

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 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


