Improving Pronoun Resolution by Incorporating Coreferential Information of Candidates

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

Coreferential information of a candidate, such as the properties of its antecedents, is important for pronoun resolution because it reflects the salience of the candidate in the local discourse. Such information, however, is usually ignored in previous learning-based systems. In this paper we present a trainable model which incorporates coreferential information of candidates into pronoun resolution. Preliminary experiments show that our model will boost the resolution performance given the right antecedents of the candidates. We further discuss how to apply our model in real resolution where the antecedents of the candidate are found by a separate noun phrase resolution module. The experimental results show that our model still achieves better performance than the baseline.

1 Introduction

In recent years, supervised machine learning approaches have been widely explored in reference resolution and achieved considerable success (Ge et al., 1998; Soon et al., 2001; Ng and Cardie, 2002; Strube and Muller, 2003; Yang et al., 2003). Most learning-based pronoun resolution systems determine the reference relationship between an anaphor and its antecedent candidate only from the properties of the pair. The knowledge about the context of anaphor and antecedent is nevertheless ignored. However, research in centering theory (Sidner, 1981; Grosz et al., 1983; Grosz et al., 1995; Tetreault, 2001) has revealed that the local focusing (or centering) also has a great effect on the processing of pronominal expressions. The choices of the antecedents of pronouns usually depend on the center of attention throughout the local discourse segment (Mitkov, 1999).

To determine the salience of a candidate in the local context, we may need to check the coreferential information of the candidate, such as the existence and properties of its antecedents. In fact, such information has been used for pronoun resolution in many heuristic-based systems. The S-List model (Strube, 1998), for example, assumes that a co-referring candidate is a *hearer-old discourse entity* and is preferred to other *hearer-new* candidates. In the algorithms based on the centering theory (Brennan et al., 1987; Grosz et al., 1995), if a candidate and its antecedent are the *backward-looking centers* of two subsequent utterances respectively, the candidate would be the most preferred since the *CONTINUE* transition is always ranked higher than *SHIFT* or *RETAIN*.

In this paper, we present a supervised learning-based pronoun resolution system which incorporates coreferential information of candidates in a trainable model. For each candidate, we take into consideration the properties of its antecedents in terms of features (henceforth *backward features*), and use the supervised learning method to explore their influences on pronoun resolution. In the study, we start our exploration on the capability of the model by applying it in an ideal environment where the antecedents of the candidates are correctly identified and the backward features are optimally set. The experiments on MUC-6 (1995) and MUC-7 (1998) corpora show that incorporating coreferential information of candidates boosts the system performance significantly. Further, we apply our model in the real resolution where the antecedents of the candidates are provided by separate noun phrase resolution modules. The experimental results show that our model still outperforms the baseline, even with the low recall of the non-pronoun resolution module.

The remaining of this paper is organized as follows. Section 2 discusses the importance of the coreferential information for candidate evaluation. Section 3 introduces the baseline learning framework. Section 4 presents and evaluates the learning model which uses backward fea-

tures to capture coreferential information, while Section 5 proposes how to apply the model in real resolution. Section 6 describes related research work. Finally, conclusion is given in Section 7.

2 The Impact of Coreferential Information on Pronoun Resolution

In pronoun resolution, the center of attention throughout the discourse segment is a very important factor for antecedent selection (Mitkov, 1999). If a candidate is the focus (or center) of the local discourse, it would be selected as the antecedent with a high possibility. See the following example,

<s> $Gitano_1$ has pulled off *a clever illusion*$_2$ with its_3 *advertising*$_4$. <s>
<s> *The campaign*$_5$ gives its_6 clothes a youthful and trendy image to lure consumers into the store. <s>

Table 1: A text segment from MUC-6 data set

In the above text, the pronoun "its_6" has several antecedent candidates, i.e., "$Gitano_1$", "*a clever illusion*$_2$", "its_3", "*its advertising*$_4$" and "*The campaign*$_5$". Without looking back, "*The campaign*$_5$" would be probably selected because of its syntactic role (*Subject*) and its distance to the anaphor. However, given the knowledge that the company *Gitano* is the focus of the local context and "its_3" refers to "$Gitano_1$", it would be clear that the pronoun "its_6" should be resolved to "its_3" and thus "$Gitano_1$", rather than other competitors.

To determine whether a candidate is the "focus" entity, we should check how the status (e.g. grammatical functions) of the entity alternates in the local context. Therefore, it is necessary to track the NPs in the coreferential chain of the candidate. For example, the syntactic roles (i.e., subject) of the antecedents of "its_3" would indicate that "its_3" refers to the most salient entity in the discourse segment.

In our study, we keep the properties of the antecedents as features of the candidates, and use the supervised learning method to explore their influence on pronoun resolution. Actually, to determine the local focus, we only need to check the entities in a short discourse segment. That is, for a candidate, the number of its adjacent antecedents to be checked is limited. Therefore, we could evaluate the salience of a candidate by looking back only its closest antecedent instead of each element in its coreferential chain, with the assumption that the closest antecedent is able to provide sufficient information for the evaluation.

3 The Baseline Learning Framework

Our baseline system adopts the common learning-based framework employed in the system by Soon et al. (2001).

In the learning framework, each training or testing instance takes the form of $\boldsymbol{i}\{ana, candi\}$, where *ana* is the possible anaphor and *candi* is its antecedent candidate[1]. An instance is associated with a feature vector to describe their relationships. As listed in Table 2, we only consider those knowledge-poor and domain-independent features which, although superficial, have been proved efficient for pronoun resolution in many previous systems.

During training, for each anaphor in a given text, a positive instance is created by paring the anaphor and its closest antecedent. Also a set of negative instances is formed by paring the anaphor and each of the intervening candidates. Based on the training instances, a binary classifier is generated using C5.0 learning algorithm (Quinlan, 1993). During resolution, each possible anaphor *ana*, is paired in turn with each preceding antecedent candidate, *candi*, from right to left to form a testing instance. This instance is presented to the classifier, which will then return a positive or negative result indicating whether or not they are co-referent. The process terminates once an instance $\boldsymbol{i}\{ana, candi\}$ is labelled as positive, and *ana* will be resolved to *candi* in that case.

4 The Learning Model Incorporating Coreferential Information

The learning procedure in our model is similar to the above baseline method, except that for each candidate, we take into consideration its closest antecedent, if possible.

4.1 Instance Structure

During both training and testing, we adopt the same instance selection strategy as in the baseline model. The only difference, however, is the structure of the training or testing instances. Specifically, each instance in our model is composed of three elements like below:

[1] In our study candidates are filtered by checking the gender, number and animacy agreements in advance.

Features describing the candidate (*candi*)		
1.	candi_DefNp	1 if *candi* is a definite NP; else 0
2.	candi_DemoNP	1 if *candi* is an indefinite NP; else 0
3.	candi_Pron	1 if *candi* is a pronoun; else 0
4.	candi_ProperNP	1 if *candi* is a proper name; else 0
5.	candi_NE_Type	1 if *candi* is an "organization" named-entity; 2 if "person", 3 if other types, 0 if not a NE
6.	candi_Human	the likelihood (0-100) that *candi* is a human entity (obtained from WordNet)
7.	candi_FirstNPInSent	1 if *candi* is the first NP in the sentence where it occurs
8.	candi_Nearest	1 if *candi* is the candidate nearest to the anaphor; else 0
9.	candi_SubjNP	1 if *candi* is the subject of the sentence it occurs; else 0
Features describing the anaphor (*ana*):		
10.	ana_Reflexive	1 if *ana* is a reflexive pronoun; else 0
11.	ana_Type	1 if *ana* is a third-person pronoun (he, she,...); 2 if a single neuter pronoun (it,...); 3 if a plural neuter pronoun (they,...); 4 if other types
Features describing the relationships between *candi* and *ana*:		
12.	SentDist	Distance between *candi* and *ana* in sentences
13.	ParaDist	Distance between *candi* and *ana* in paragraphs
14.	CollPattern	1 if *candi* has an identical collocation pattern with ana; else 0

Table 2: Feature set for the baseline pronoun resolution system

$\boldsymbol{i}$\{*ana*, *candi*, *ante-of-candi*\}

where *ana* and *candi*, similar to the definition in the baseline model, are the anaphor and one of its candidates, respectively. The new added element in the instance definition, *ante-of-candi*, is the possible closest antecedent of *candi* in its coreferential chain. The *ante-of-candi* is set to NIL in the case when *candi* has no antecedent.

Consider the example in Table 1 again. For the pronoun "it_6", three training instances will be generated, namely, $\boldsymbol{i}\{it_6, The\ compaign_5, \text{NIL}\}$, $\boldsymbol{i}\{it_6, its\ advertising_4, \text{NIL}\}$, and $\boldsymbol{i}\{it_6, it_3, Gitano_1\}$.

4.2 Backward Features

In addition to the features adopted in the baseline system, we introduce a set of backward features to describe the element *ante-of-candi*. The ten features (15-24) are listed in Table 3 with their respective possible values.

Like feature 1-9, features 15-22 describe the lexical, grammatical and semantic properties of *ante-of-candi*. The inclusion of the two features *Apposition* (23) and *candi_NoAntecedent* (24) is inspired by the work of Strube (1998). The feature *Apposition* marks whether or not *candi* and *ante-of-candi* occur in the same appositive structure. The underlying purpose of this feature is to capture the pattern that proper names are accompanied by an appositive. The entity with such a pattern may often be related to the hearers' knowledge and has low preference. The feature *candi_NoAntecedent* marks whether or not a candidate has a valid antecedent in the preceding text. As stipulated in Strube's work, co-referring expressions belong to *hearer-old entities* and therefore have higher preference than other candidates. When the feature is assigned value 1, all the other backward features (15-23) are set to 0.

4.3 Results and Discussions

In our study we used the standard MUC-6 and MUC-7 coreference corpora. In each data set, 30 "dry-run" documents were annotated for training as well as 20-30 documents for testing. The raw documents were preprocessed by a pipeline of automatic NLP components (e.g. NP chunker, part-of-speech tagger, named-entity recognizer) to determine the boundary of the NPs, and to provide necessary information for feature calculation.

In an attempt to investigate the capability of our model, we evaluated the model in an optimal environment where the closest antecedent of each candidate is correctly identified. MUC-6 and MUC-7 can serve this purpose quite well; the annotated coreference information in the data sets enables us to obtain the correct closest

Features describing the antecedent of the candidate (*ante-of-candi*):		
15.	ante-candi_DefNp	1 if *ante-of-candi* is a definite NP; else 0
16.	ante-candi_IndefNp	1 if *ante-of-candi* is an indefinite NP; else 0
17.	ante-candi_Pron	1 if *ante-of-candi* is a pronoun; else 0
18.	ante-candi_Proper	1 if *ante-of-candi* is a proper name; else 0
19.	ante-candi_NE_Type	1 if *ante-of-candi* is an "organization" named-entity; 2 if "person", 3 if other types, 0 if not a NE
20.	ante-candi_Human	the likelihood (0-100) that *ante-of-candi* is a human entity
21.	ante-candi_FirstNPInSent	1 if *ante-of-candi* is the first NP in the sentence where it occurs
22.	ante-candi_SubjNP	1 if *ante-of-candi* is the subject of the sentence where it occurs
Features describing the relationships between the candidate (*candi*) and *ante-of-candi*:		
23.	Apposition	1 if *ante-of-candi* and *candi* are in an appositive structure
Features describing the candidate (*candi*):		
24.	candi_NoAntecedent	1 if *candi* has no antecedent available; else 0

Table 3: Backward features used to capture the coreferential information of a candidate

antecedent for each candidate and accordingly generate the training and testing instances. In the next section we will further discuss how to apply our model into the real resolution.

Table 4 shows the performance of different systems for resolving the pronominal anaphors [2] in MUC-6 and MUC-7. Default learning parameters for C5.0 were used throughout the experiments. In this table we evaluated the performance based on two kinds of measurements:

- "Recall-and-Precision":

  $$\text{Recall} = \frac{\#positive\ instances\ classified\ correctly}{\#positive\ instances}$$

  $$\text{Precision} = \frac{\#positive\ instances\ classified\ correctly}{\#instances\ classified\ as\ positive}$$

  The above metrics evaluate the capability of the learned classifier in identifying positive instances[3]. *F-measure* is the harmonic mean of the two measurements.

- "Success":

  $$\text{Success} = \frac{\#anaphors\ resolved\ correctly}{\#total\ anaphors}$$

  The metric[4] directly reflects the pronoun resolution capability.

The first and second lines of Table 4 compare the performance of the baseline system (*Baseline*) and our system (*Optimal*), where DT_{pron} and $\text{DT}_{pron-opt}$ are the classifiers learned in the two systems, respectively. The results indicate that our system outperforms the baseline system significantly. Compared with *Baseline*, *Optimal* achieves gains in both recall (6.4% for MUC-6 and 4.1% for MUC-7) and precision (1.3% for MUC-6 and 9.0% for MUC-7). For Success, we also observe an apparent improvement by 4.7% (MUC-6) and 3.5% (MUC-7).

```
ante-candi_SubjNP = 1: 1 (49/5)
ante-candi_SubjNP = 0:
:..candi_SubjNP = 1:
   :..SentDist = 2: 0 (3)
   :  SentDist = 0:
   :  :..candi_Human > 0: 1 (39/2)
   :  :  candi_Human <= 0:
   :  :  :..candi_NoAntecedent = 0: 1 (8/3)
   :  :     candi_NoAntecedent = 1: 0 (3)
   :  SentDist = 1:
   :  :..ante-candi_Human <= 50 : 0 (4)
   :     ante-candi_Human > 50 : 1 (10/2)
   :
   candi_SubjNP = 0:
   :..candi_Pron = 1: 1 (32/7)
      candi_Pron = 0:
      :..candi_NoAntecedent = 1:
         :..candi_FirstNPInSent = 1: 1 (6/2)
         :  candi_FirstNPInSent = 0: ...
         candi_NoAntecedent = 0: ...
```

Figure 1: Top portion of the decision tree learned on MUC-6 with the backward features

Figure 1 shows the portion of the pruned decision tree learned for MUC-6 data set. It visualizes the importance of the backward features for the pronoun resolution on the data set. From

[2]The first and second person pronouns are discarded in our study.

[3]The testing instances are collected in the same ways as the training instances.

[4]In the experiments, an anaphor is considered correctly resolved only if the found antecedent is in the same coreferential chain of the anaphor.

Experiments	Testing classifier	Backward feature assigner*	MUC-6				MUC-7			
			R	P	F	S	R	P	F	S
Baseline	DT_{pron}	NIL	77.2	83.4	80.2	70.0	71.9	68.6	70.2	59.0
Optimal	$DT_{pron-opt}$	(Annotated)	**83.6**	84.7	**84.1**	**74.7**	**76.0**	77.6	**76.8**	**62.5**
RealResolve-1	$DT_{pron-opt}$	$DT_{pron-opt}$	75.8	83.8	79.5	73.1	62.3	77.7	69.1	53.8
RealResolve-2	$DT_{pron-opt}$	DT_{pron}	75.8	83.8	79.5	73.1	63.0	**77.9**	69.7	54.9
RealResolve-3	DT'_{pron}	DT_{pron}	79.3	**86.3**	82.7	**74.7**	74.7	67.3	70.8	60.8
RealResolve-4	DT'_{pron}	DT'_{pron}	79.3	**86.3**	82.7	**74.7**	74.7	67.3	70.8	60.8

Table 4: Results of different systems for pronoun resolution on MUC-6 and MUC-7 (*Here we only list backward feature assigner for pronominal candidates. In *RealResolve-1* to *RealResolve-4*, the backward features for non-pronominal candidates are all found by $DT_{non-pron}$.)

the tree we could find that:

1.) Feature *ante-candi_SubjNP* is of the most importance as the root feature of the tree. The decision tree would first examine the syntactic role of a candidate's antecedent, followed by that of the candidate. This nicely proves our assumption that the properties of the antecedents of the candidates provide very important information for the candidate evaluation.

2.) Both features *ante-candi_SubjNP* and *candi_SubjNP* rank top in the decision tree. That is, for the reference determination, the subject roles of the candidate's referent within a discourse segment will be checked in the first place. This finding supports well the suggestion in centering theory that the grammatical relations should be used as the key criteria to rank *forward-looking centers* in the process of focus tracking (Brennan et al., 1987; Grosz et al., 1995).

3.) *candi_Pron* and *candi_NoAntecedent* are to be examined in the cases when the subject-role checking fails, which confirms the hypothesis in the S-List model by Strube (1998) that co-refereing candidates would have higher preference than other candidates in the pronoun resolution.

5 Applying the Model in Real Resolution

In Section 4 we explored the effectiveness of the backward feature for pronoun resolution. In those experiments our model was tested in an ideal environment where the closest antecedent of a candidate can be identified correctly when generating the feature vector. However, during real resolution such coreferential information is not available, and thus a separate module has

algorithm PRON-RESOLVE
input:
$DT_{non-pron}$: classifier for resolving non-pronouns
DT_{pron}: classifier for resolving pronouns
begin:
$M_{1..n}$:= the valid markables in the given document
Ante[1..n] := 0
for i = 1 **to** N
  for j = i - 1 **downto** 0
    if (M_i is a non-pron **and**
      $DT_{non-pron}(i\{M_i, M_j\}) == +$)
      or
      (M_i is a pron **and**
      $DT_{pron}(i\{M_i, M_j, Ante[j]\}) == +$)
    then
      Ante[i] := M_j
      break
return Ante

Figure 2: The pronoun resolution algorithm by incorporating coreferential information of candidates

to be employed to obtain the closest antecedent for a candidate. We describe the algorithm in Figure 2.

The algorithm takes as input two classifiers, one for the non-pronoun resolution and the other for pronoun resolution. Given a testing document, the antecedent of each NP is identified using one of these two classifiers, depending on the type of NP. Although a separate non-pronoun resolution module is required for the pronoun resolution task, this is usually not a big problem as these two modules are often integrated in coreference resolution systems. We just use the results of the one module to improve the performance of the other.

5.1 New Training and Testing Procedures

For a pronominal candidate, its antecedent can be obtained by simply using $DT_{pron-opt}$. For

Training Procedure:
T1. Train a non-pronoun resolution classifier $DT_{non-pron}$ and a pronoun resolution classifier DT_{pron}, using the baseline learning framework (without backward features).
T2. Apply $DT_{non-pron}$ and DT_{pron} to identify the antecedent of each non-pronominal and pronominal markable, respectively, in a given document.
T3. Go through the document again. Generate instances with backward features assigned using the antecedent information obtained in T2.
T4. Train a new pronoun resolution classifier DT'_{pron} on the instances generated in T3.

Testing Procedure:
R1. For each given document, do T2∼T3.
R2. Resolve pronouns by applying DT'_{pron}.

Table 5: New training and testing procedures

a non-pronominal candidate, we built a non-pronoun resolution module to identify its antecedent. The module is a duplicate of the NP coreference resolution system by Soon et al. (2001)[5] , which uses the similar learning framework as described in Section 3. In this way, we could do pronoun resolution just by running PRON-RESOLVE($DT_{non-pron}$, $DT_{pron-opt}$), where $DT_{non-pron}$ is the classifier of the non-pronoun resolution module.

One problem, however, is that $DT_{pron-opt}$ is trained on the instances whose backward features are correctly assigned. During real resolution, the antecedent of a candidate is found by $DT_{non-pron}$ or $DT_{pron-opt}$, and the backward feature values are not always correct. Indeed, for most noun phrase resolution systems, the recall is not very high. The antecedent sometimes can not be found, or is not the closest one in the preceding coreferential chain. Consequently, the classifier trained on the "perfect" feature vectors would probably fail to output anticipated results on the noisy data during real resolution.

Thus we modify the training and testing procedures of the system. For both training and testing instances, we assign the backward feature values based on the results from separate NP resolution modules. The detailed procedures are described in Table 5.

[5]Details of the features can be found in Soon et al. (2001)

algorithm REFINE-CLASSIFIER
begin:
$DT^1_{pron} := DT'_{pron}$
for i = 1 **to** ∞
  Use DT^i_{pron} to update the antecedents of pronominal candidates and the corresponding backward features;
  Train DT^{i+1}_{pron} based on the updated training instances;
  if DT^{i+1}_{pron} is not better than DT^i_{pron} **then**
    break;
return DT^i_{pron}

Figure 3: The classifier refining algorithm

The idea behind our approach is to train and test the pronoun resolution classifier on instances with feature values set in a consistent way. Here the purpose of DT_{pron} and $DT_{non-pron}$ is to provide backward feature values for training and testing instances. From this point of view, the two modules could be thought of as a preprocessing component of our pronoun resolution system.

5.2 Classifier Refining

If the classifier DT'_{pron} outperforms DT_{pron} as expected, we can employ DT'_{pron} in place of DT_{pron} to generate backward features for pronominal candidates, and then train a classifier DT''_{pron} based on the updated training instances. Since DT'_{pron} produces more correct feature values than DT_{pron}, we could expect that DT''_{pron} will not be worse, if not better, than DT'_{pron}. Such a process could be repeated to refine the pronoun resolution classifier. The algorithm is described in Figure 3.

In algorithm REFINE-CLASSIFIER, the iteration terminates when the new trained classifier DT^{i+1}_{pron} provides no further improvement than DT^i_{pron}. In this case, we can replace DT^{i+1}_{pron} by DT^i_{pron} during the i+1(th) testing procedure. That means, by simply running PRON-RESOLVE($DT_{non-pron}$,DT^i_{pron}), we can use for both backward feature computation and instance classification tasks, rather than applying DT_{pron} and DT'_{pron} subsequently.

5.3 Results and Discussions

In the experiments we evaluated the performance of our model in real pronoun resolution.

The performance of our model depends on the performance of the non-pronoun resolution classifier, $DT_{non-pron}$. Hence we first examined the

coreference resolution capability of $DT_{non-pron}$ based on the standard scoring scheme by Vilain et al. (1995). For MUC-6, the module obtains 62.2% recall and 78.8% precision, while for MUC-7, it obtains 50.1% recall and 75.4% precision. The poor recall and comparatively high precision reflect the capability of the state-of-the-art learning-based NP resolution systems.

The third block of Table 4 summarizes the performance of the classifier $DT_{pron-opt}$ in real resolution. In the systems *RealResolve-1* and *RealResolve-2*, the antecedents of pronominal candidates are found by $DT_{pron-opt}$ and DT_{pron} respectively, while in both systems the antecedents of non-pronominal candidates are by $DT_{non-pron}$. As shown in the table, compared with the *Optimal* where the backward features of testing instances are optimally assigned, the recall rates of two systems drop largely by 7.8% for MUC-6 and by about 14% for MUC-7. The scores of recall are even lower than those of *Baseline*. As a result, in comparison with *Optimal*, we see the degrade of the F-measure and the success rate, which confirms our hypothesis that the classifier learned on perfect training instances would probably not perform well on the noisy testing instances.

The system *RealResolve-3* listed in the fifth line of the table uses the classifier trained and tested on instances whose backward features are assigned according to the results from $DT_{non-pron}$ and DT_{pron}. From the table we can find that: (1) Compared with *Baseline*, the system produces gains in recall (2.1% for MUC-6 and 2.8% for MUC-7) with no significant loss in precision. Overall, we observe the increase in F-measure for both data sets. If measured by Success, the improvement is more apparent by 4.7% (MUC-6) and 1.8% (MUC-7). (2) Compared with *RealResolve-1(2)*, the performance decrease of *RealResolve-3* against *Optimal* is not so large. Especially for MUC-6, the system obtains a success rate as high as *Optimal*.

The above results show that our model can be successfully applied in the real pronoun resolution task, even given the low recall of the current non-pronoun resolution module. This should be owed to the fact that for a candidate, its adjacent antecedents, even not the closest one, could give clues to reflect its salience in the local discourse. That is, the model prefers a high precision to a high recall, which copes well with the capability of the existing non-pronoun resolution module.

In our experiments we also tested the classifier refining algorithm described in Figure 3. We found that for both MUC-6 and MUC-7 data set, the algorithm terminated in the second round. The comparison of DT^2_{pron} and DT^1_{pron} (i.e. DT'_{pron}) showed that these two trees were exactly the same. The algorithm converges fast probably because in the data set, most of the antecedent candidates are non-pronouns (89.1% for MUC-6 and 83.7% for MUC-7). Consequently, the ratio of the training instances with backward features changed may be not substantial enough to affect the classifier generation.

Although the algorithm provided no further refinement for DT'_{pron}, we can use DT'_{pron}, as suggested in Section 5.2, to calculate backward features and classify instances by running PRON-RESOLVE($DT_{non-pron}$, DT'_{pron}). The results of such a system, *RealResolve-4*, are listed in the last line of Table 4. For both MUC-6 and MUC-7, *RealResolve-4* obtains exactly the same performance as *RealResolve-3*.

6 Related Work

To our knowledge, our work is the first effort that systematically explores the influence of coreferential information of candidates on pronoun resolution in learning-based ways. Iida et al. (2003) also take into consideration the contextual clues in their coreference resolution system, by using two features to reflect the ranking order of a candidate in Salience Reference List (SRL). However, similar to common centering models, in their system the ranking of entities in SRL is also heuristic-based.

The coreferential chain length of a candidate, or its variants such as occurrence frequency and TFIDF, has been used as a salience factor in some learning-based reference resolution systems (Iida et al., 2003; Mitkov, 1998; Paul et al., 1999; Strube and Muller, 2003). However, for an entity, the coreferential length only reflects its global salience in the whole text(s), instead of the local salience in a discourse segment which is nevertheless more informative for pronoun resolution. Moreover, during resolution, the found coreferential length of an entity is often incomplete, and thus the obtained length value is usually inaccurate for the salience evaluation.

7 Conclusion and Future Work

In this paper we have proposed a model which incorporates coreferential information of candi-

dates to improve pronoun resolution. When evaluating a candidate, the model considers its adjacent antecedent by describing its properties in terms of backward features. We first examined the effectiveness of the model by applying it in an optimal environment where the closest antecedent of a candidate is obtained correctly. The experiments show that it boosts the success rate of the baseline system for both MUC-6 (4.7%) and MUC-7 (3.5%). Then we proposed how to apply our model in the real resolution where the antecedent of a non-pronoun is found by an additional non-pronoun resolution module. Our model can still produce Success improvement (4.7% for MUC-6 and 1.8% for MUC-7) against the baseline system, despite the low recall of the non-pronoun resolution module.

In the current work we restrict our study only to pronoun resolution. In fact, the coreferential information of candidates is expected to be also helpful for non-pronoun resolution. We would like to investigate the influence of the coreferential factors on general NP reference resolution in our future work.

References

S. Brennan, M. Friedman, and C. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Compuational Linguistics*, pages 155–162.

N. Ge, J. Hale, and E. Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the 6th Workshop on Very Large Corpora*.

B. Grosz, A. Joshi, and S. Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual meeting of the Association for Computational Linguistics*, pages 44–50.

B. Grosz, A. Joshi, and S. Weinstein. 1995. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

R. Iida, K. Inui, H. Takamura, and Y. Matsumoto. 2003. Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the 10th Conference of EACL, Workshop "The Computational Treatment of Anaphora"*.

R. Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 17th Int. Conference on Computational Linguistics*, pages 869–875.

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

M. Paul, K. Yamamoto, and E. Sumita. 1999. Corpus-based anaphora resolution towards antecedent preference. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Workshop "Coreference and It's Applications"*, pages 47–52.

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

C. Sidner. 1981. Focusing for interpretation of pronouns. *American Journal of Computational Linguistics*, 7(4):217–231.

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 and C. Muller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 168–175, Japan.

M. Strube. 1998. Never look back: An alternative to centering. In *Proceedings of the 17th Int. Conference on Computational Linguistics and 36th Annual Meeting of ACL*, pages 1251–1257.

J. R. Tetreault. 2001. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520.

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. 2003. Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Japan.