Improving Pronoun Resolution Using Statistics-Based Semantic Compatibility Information

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

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

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

Abstract

In this paper we focus on how to improve pronoun resolution using the statistics-based semantic compatibility information. We investigate two unexplored issues that influence the effectiveness of such information: statistics source and learning framework. Specifically, we for the first time propose to utilize the web and the twin-candidate model, in addition to the previous combination of the corpus and the single-candidate model, to compute and apply the semantic information. Our study shows that the semantic compatibility obtained from the web can be effectively incorporated in the twin-candidate learning model and significantly improve the resolution of neutral pronouns.

1 Introduction

Semantic compatibility is an important factor for pronoun resolution. Since pronouns, especially neutral pronouns, carry little semantics of their own, the compatibility between an anaphor and its antecedent candidate is commonly evaluated by examining the relationships between the candidate and the anaphor's context, based on the statistics that the corresponding predicate-argument tuples occur in a particular large corpus. Consider the example given in the work of Dagan and Itai (1990):

(1) *They know full well that companies held tax money aside for collection later on the basis that the government said* ***it***$_1$ *was going to collect* ***it***$_2$.

For anaphor it_1, the candidate *government* should have higher semantic compatibility than *money* because *government_collect* is supposed to occur more frequently than *money_collect* in a large corpus. A similar pattern could also be observed for it_2.

So far, the corpus-based semantic knowledge has been successfully employed in several anaphora resolution systems. Dagan and Itai (1990) proposed a heuristics-based approach to pronoun resolution. It determined the preference of candidates based on predicate-argument frequencies. Recently, Bean and Riloff (2004) presented an unsupervised approach to coreference resolution, which mined the co-referring NP pairs with similar predicate-arguments from a large corpus using a bootstrapping method.

However, the utility of the corpus-based semantics for pronoun resolution is often argued. Kehler et al. (2004), for example, explored the usage of the corpus-based statistics in supervised learning based systems, and found that such information did not produce apparent improvement for the overall pronoun resolution. Indeed, existing learning-based approaches to anaphor resolution have performed reasonably well using limited and shallow knowledge (e.g., Mitkov (1998), Soon et al. (2001), Strube and Muller (2003)). Could the relatively noisy semantic knowledge give us further system improvement?

In this paper we focus on improving pronominal anaphora resolution using automatically computed semantic compatibility information. We propose to enhance the utility of the statistics-based knowledge from two aspects:

Statistics source. Corpus-based knowledge usually suffers from data sparseness problem. That is, many predicate-argument tuples would be unseen even in a large corpus. A possible solution is the

Proceedings of the 43rd Annual Meeting of the ACL, pages 165–172,
Ann Arbor, June 2005.

web. It is believed that the size of the web is thousands of times larger than normal large corpora, and the counts obtained from the web are highly correlated with the counts from large balanced corpora for predicate-argument bi-grams (Keller and Lapata, 2003). So far the web has been utilized in nominal anaphora resolution (Modjeska et al., 2003; Poesio et al., 2004) to determine the semantic relation between an anaphor and candidate pair. However, to our knowledge, using the web to help pronoun resolution still remains unexplored.

Learning framework. Commonly, the predicate-argument statistics is incorporated into anaphora resolution systems as a feature. What kind of learning framework is suitable for this feature? Previous approaches to anaphora resolution adopt the single-candidate model, in which the resolution is done on an anaphor and one candidate at a time (Soon et al., 2001; Ng and Cardie, 2002). However, as the purpose of the predicate-argument statistics is to evaluate the preference of the candidates in semantics, it is possible that the statistics-based semantic feature could be more effectively applied in the twin-candidate (Yang et al., 2003) that focusses on the preference relationships among candidates.

In our work we explore the acquisition of the semantic compatibility information from the corpus and the web, and the incorporation of such semantic information in the single-candidate model and the twin-candidate model. We systematically evaluate the combinations of different statistics sources and learning frameworks in terms of their effectiveness in helping the resolution. Results on the MUC data set show that for neutral pronoun resolution in which an anaphor has no specific semantic category, the web-based semantic information would be the most effective when applied in the twin-candidate model: Not only could such a system significantly improve the baseline without the semantic feature, it also outperforms the system with the combination of the corpus and the single-candidate model (by 11.5% success).

The rest of this paper is organized as follows. Section 2 describes the acquisition of the semantic compatibility information from the corpus and the web. Section 3 discusses the application of the statistics in the single-candidate and twin-candidate learning models. Section 4 gives the experimental results, and finally, Section 5 gives the conclusion.

2 Computing the Statistics-based Semantic Compatibility

In this section, we introduce in detail how to compute the semantic compatibility, using the predicate-argument statistics obtained from the corpus or the web.

2.1 Corpus-Based Semantic Compatibility

Three relationships, possessive-noun, subject-verb and verb-object, are considered in our work. Before resolution a large corpus is prepared. Documents in the corpus are processed by a shallow parser that could generate predicate-argument tuples of the above three relationships[1].

To reduce data sparseness, the following steps are applied in each resulting tuple, automatically:

- Only the nominal or verbal heads are retained.
- Each Named-Entity (NE) is replaced by a common noun which corresponds to the semantic category of the NE (e.g. "IBM" → "company") [2].
- All words are changed to their base morphologic forms (e.g. "companies → company").

During resolution, for an encountered anaphor, each of its antecedent candidates is substituted with the anaphor . According to the role and type of the anaphor in its context, a predicate-argument tuple is extracted and the above three steps for data-sparse reduction are applied. Consider the sentence (1), for example. The anaphors "it_1" and "it_2" indicate a subject_verb and verb_object relationship, respectively. Thus, the predicate-argument tuples for the two candidates "*government*" and "*money*" would be *(collect (subject government))* and *(collect (subject money))* for "it_1", and *(collect (object government))* and *(collect (object money))* for "it_2".

Each extracted tuple is searched in the prepared tuples set of the corpus, and the times the tuple occurs are calculated. For each candidate, its semantic

[1]The possessive-noun relationship involves the forms like "NP_2 *of* NP_1" and "NP_1*'s* NP_2".

[2]In our study, the semantic category of a NE is identified automatically by the pre-processing NE recognition component.

compatibility with the anaphor could be represented simply in terms of *frequency*

$$StatSem(candi, ana) = count(candi, ana) \quad (1)$$

where $count(candi, ana)$ is the count of the tuple formed by *candi* and *ana*, or alternatively, in terms of *conditional probability* ($P(candi, ana|candi)$), where the count of the tuple is divided by the count of the single candidate in the corpus. That is

$$StatSem(candi, ana) = \frac{count(candi, ana)}{count(candi)} \quad (2)$$

In this way, the statistics would not bias candidates that occur frequently in isolation.

2.2 Web-Based Semantic Compatibility

Unlike documents in normal corpora, web pages could not be preprocessed to generate the predicate-argument reserve. Instead, the predicate-argument statistics has to be obtained via a web search engine like Google and Altavista. For the three types of predicate-argument relationships, queries are constructed in the forms of "NP$_{candi}$ VP" (for subject-verb), "VP NP$_{candi}$" (for verb-object), and "NP$_{candi}$'s NP" or "NP of NP$_{candi}$" (for possessive-noun). Consider the following sentence:

(2) *Several experts suggested that IBM's accounting grew much more liberal since the mid 1980s as **its** business turned sour.*

For the pronoun "*its*" and the candidate "*IBM*", the two generated queries are "*business of IBM*" and "*IBM's business*".

To reduce data sparseness, in an initial query only the nominal or verbal heads are retained. Also, each NE is replaced by the corresponding common noun. (e.g, "*IBM's business*" → "*company's business*" and "*business of IBM*" → "*business of company*").

A set of inflected queries is generated by expanding a term into all its possible morphological forms. For example, in Sentence (1), "*collect money*" becomes "*collected|collecting|... money*", and in (2) "*business of company*" becomes "*business of company|companies*"). Besides, determiners are inserted for every noun. If the noun is the candidate under consideration, only the definite article *the* is inserted. For other nouns, instead, *a/an, the* and the empty determiners (for bare plurals) would be added (e.g., "*the|a business of the company|companies*").

Queries are submitted to a particular web search engine (Google in our study). All queries are performed as exact matching. Similar to the corpus-based statistics, the compatibility for each candidate and anaphor pair could be represented using either *frequency* (Eq. 1) or *probability* (Eq. 2) metric. In such a situation, $count(candi, ana)$ is the hit number of the inflected queries returned by the search engine, while $count(candi)$ is the hit number of the query formed with only the head of the candidate (i.e.,"*the* + *candi*").

3 Applying the Semantic Compatibility

In this section, we discuss how to incorporate the statistics-based semantic compatibility for pronoun resolution, in a machine learning framework.

3.1 The Single-Candidate Model

One way to utilize the semantic compatibility is to take it as a feature under the single-candidate learning model as employed by Ng and Cardie (2002).

In such a learning model, each training or testing instance takes the form of ***i***$\{C, ana\}$, where *ana* is the possible anaphor and *C* is its antecedent candidate. An instance is associated with a feature vector to describe their relationships.

During training, for each anaphor in a given text, a positive instance is created by pairing the anaphor and its closest antecedent. Also a set of negative instances is formed by pairing the anaphor and each of the intervening candidates. Based on the training instances, a binary classifier is generated using a certain learning algorithm, like C5 (Quinlan, 1993) in our work.

During resolution, given a new anaphor, a test instance is created for each candidate. This instance is presented to the classifier, which then returns a positive or negative result with a confidence value indicating the likelihood that they are co-referent. The candidate with the highest confidence value would be selected as the antecedent.

3.2 Features

In our study we only consider those domain-independent features that could be obtained with low

Feature	Description
DefNp	1 if the candidate is a definite NP; else 0
Pron	1 if the candidate is a pronoun; else 0
NE	1 if the candidate is a named entity; else 0
SameSent	1 if the candidate and the anaphor is in the same sentence; else 0
NearestNP	1 if the candidate is nearest to the anaphor; else 0
ParalStuct	1 if the candidate has an parallel structure with ana; else 0
FirstNP	1 if the candidate is the first NP in a sentence; else 0
Reflexive	1 if the anaphor is a reflexive pronoun; else 0
Type	Type of the anaphor (0: Single neuter pronoun; 1: Plural neuter pronoun; 2: Male personal pronoun; 3: Female personal pronoun)
StatSem*	the statistics-base semantic compatibility of the candidate
SemMag**	the semantic compatibility difference between two competing candidates

Table 1: Feature set for our pronoun resolution system(*ed feature is only for the single-candidate model while **ed feature is only for the twin-candidate mode)

computational cost but with high reliability. Table 1 summarizes the features with their respective possible values. The first three features represent the lexical properties of a candidate. The POS properties could indicate whether a candidate refers to a hearer-old entity that would have a higher preference to be selected as the antecedent (Strube, 1998). *SameSent* and *NearestNP* mark the distance relationships between an anaphor and the candidate, which would significantly affect the candidate selection (Hobbs, 1978). *FirstNP* aims to capture the salience of the candidate in the local discourse segment. *ParalStuct* marks whether a candidate and an anaphor have similar surrounding words, which is also a salience factor for the candidate evaluation (Mitkov, 1998).

Feature *StatSem* records the statistics-based semantic compatibility computed, from the corpus or the web, by either *frequency* or *probability* metric, as described in the previous section. If a candidate is a pronoun, this feature value would be set to that of its closest nominal antecedent.

As described, the semantic compatibility of a candidate is computed under the context of the current anaphor. Consider two occurrences of anaphors "$\ldots it_1$ *collected* $\ldots$" and "$\ldots it_2$ *said* $\ldots$". As "NP *collected*" should occur less frequently than "NP *said*", the candidates of it_1 would generally have predicate-argument statistics lower than those of it_2. That is, a positive instance for it_1 might bear a lower semantic feature value than a negative instance for it_2. The consequence is that the learning algorithm would think such a feature is not that "indicative" and reduce its salience in the resulting classifier.

One way to tackle this problem is to normalize the feature by the frequencies of the anaphor's context, e.g., "*count(collected)*" and "*count(said)*". This, however, would require extra calculation. In fact, as candidates of a specific anaphor share the same anaphor context, we can just normalize the semantic feature of a candidate by that of its competitor:

$$StatSem_N(C, ana) = \frac{StatSem(C, ana)}{\max\limits_{c_i \in candi\_set(ana)} StatSem(c_i, ana)}$$

The value ($0 \sim 1$) represents the rank of the semantic compatibility of the candidate *C* among $candi\_set(ana)$, the current candidates of ana.

3.3 The Twin-Candidate Model

Yang et al. (2003) proposed an alternative twin-candidate model for anaphora resolution task. The strength of such a model is that unlike the single-candidate model, it could capture the preference relationships between competing candidates. In the model, candidates for an anaphor are paired and features from two competing candidates are put together for consideration. This property could nicely deal with the above mentioned training problem of different anaphor contexts, because the semantic feature would be considered under the current candidate set only. In fact, as semantic compatibility is

a preference-based factor for anaphor resolution, it would be incorporated in the twin-candidate model more naturally.

In the twin-candidate model, an instance takes a form like ***i***$\{C_1, C_2, ana\}$, where C_1 and C_2 are two candidates. We stipulate that C_2 should be closer to *ana* than C_1 in distance. The instance is labelled as "10" if C_1 the antecedent, or "01" if C_2 is.

During training, for each anaphor, we find its closest antecedent, C_{ante}. A set of "10" instances, ***i***$\{C_{ante}, C, ana\}$, is generated by pairing C_{ante} and each of the interning candidates *C*. Also a set of "01" instances, ***i***$\{C, C_{ante}, ana\}$, is created by pairing C_{ante} with each candidate before C_{ante} until another antecedent, if any, is reached.

The resulting pairwise classifier would return "10" or "01" indicating which candidate is preferred to the other. During resolution, candidates are paired one by one. The score of a candidate is the total number of the competitors that the candidate wins over. The candidate with the highest score would be selected as the antecedent.

Features The features for the twin-candidate model are similar to those for the single-candidate model except that a duplicate set of features has to be prepared for the additional candidate. Besides, a new feature, *SemMag*, is used in place of *StatSem* to represent the difference magnitude between the semantic compatibility of two candidates. Let $mag = StatSem(C_1, ana)/StatSem(C_2, ana)$, feature *SemMag* is defined as follows,

$$SemMag(C_1, C_2, ana) = \begin{cases} mag - 1 & : mag >= 1 \\ 1 - mag^{-1} & : mag < 1 \end{cases}$$

The positive or negative value marks the times that the statistics of C_1 is larger or smaller than C_2.

4 Evaluation and Discussion

4.1 Experiment Setup

In our study we were only concerned about the third-person pronoun resolution. With an attempt to examine the effectiveness of the semantic feature on different types of pronouns, the whole resolution was divided into neutral pronoun (*it* & *they*) resolution and personal pronoun (*he* & *she*) resolution.

The experiments were done on the newswire domain, using MUC corpus (Wall Street Journal articles). The training was done on 150 documents from MUC-6 coreference data set, while the testing was on the 50 formal-test documents of MUC-6 (30) and MUC-7 (20). Throughout the experiments, default learning parameters were applied to the C5 algorithm. The performance was evaluated based on *success*, the ratio of the number of correctly resolved anaphors over the total number of anaphors.

An input raw text was preprocessed automatically by a pipeline of NLP components. The noun phrase identification and the predicate-argument extraction were done based on the results of a chunk tagger, which was trained for the shared task of CoNLL-2000 and achieved 92% accuracy (Zhou et al., 2000). The recognition of NEs as well as their semantic categories was done by a HMM based NER, which was trained for the MUC NE task and obtained high F-scores of 96.9% (MUC-6) and 94.3% (MUC-7) (Zhou and Su, 2002).

For each anaphor, the markables occurring within the current and previous two sentences were taken as the initial candidates. Those with mismatched number and gender agreements were filtered from the candidate set. Also, pronouns or NEs that disagreed in person with the anaphor were removed in advance. For the training set, there are totally 645 neutral pronouns and 385 personal pronouns with non-empty candidate set, while for the testing set, the number is 245 and 197.

4.2 The Corpus and the Web

The corpus for the predicate-argument statistics computation was from the TIPSTER's Text Research Collection (v1994). Consisting of 173,252 Wall Street Journal articles from the year 1988 to 1992, the data set contained about 76 million words. The documents were preprocessed using the same POS tagging and NE-recognition components as in the pronoun resolution task. Cass (Abney, 1996), a robust chunker parser was then applied to generate the shallow parse trees, which resulted in 353,085 possessive-noun tuples, 759,997 verb-object tuples and 1,090,121 subject-verb tuples.

We examined the capacity of the web and the corpus in terms of zero-count ratio and count number. On average, among the predicate-argument tuples that have non-zero corpus-counts, above 93% have also non-zero web-counts. But the ratio is only around 40% contrariwise. And for the predicate-

Learning Model	System	Neutral Pron		Personal Pron		Overall	
		Corpus	Web	Corpus	Web	Corpus	Web
Single-Candidate	baseline	65.7		86.8		75.1	
	+frequency	67.3	69.9	86.8	86.8	76.0	76.9
	+normalized frequency	66.9	67.8	86.8	86.8	75.8	76.2
	+probability	65.7	65.7	86.8	86.8	75.1	75.1
	+normalized probability	67.7	70.6	86.8	86.8	76.2	77.8
Twin-Candidate	baseline	73.9		91.9		81.9	
	+frequency	76.7	**79.2**	91.4	91.9	83.3	**84.8**
	+probability	75.9	78.0	91.4	**92.4**	82.8	84.4

Table 2: The performance of different resolution systems

Relationship	N-Pron	P-Pron
Possessive-Noun	0.508	0.517
Verb-Object	0.503	0.526
Subject-Verb	0.619	0.676

Table 3: Correlation between web and corpus counts on the seen predicate-argument tuples

argument tuples that could be seen in both data sources, the count from the web is above 2000 times larger than that from the corpus.

Although much less sparse, the web counts are significantly noisier than the corpus count since no tagging, chunking and parsing could be carried out on the web pages. However, previous study (Keller and Lapata, 2003) reveals that the large amount of data available for the web counts could outweigh the noisy problems. In our study we also carried out a correlation analysis[3] to examine whether the counts from the web and the corpus are linearly related, on the predicate-argument tuples that can be seen in both data sources. From the results listed in Table 3, we observe moderately high correlation, with coefficients ranging from 0.5 to 0.7 around, between the counts from the web and the corpus, for both neutral pronoun (N-Pron) and personal pronoun (P-Pron) resolution tasks.

4.3 System Evaluation

Table 2 summarizes the performance of the systems with different combinations of statistics sources and learning frameworks. The systems without the semantic feature were used as the baseline. Under the single-candidate (SC) model, the baseline system obtains a success of 65.7% and 86.8% for neutral pronoun and personal pronoun resolution, respectively. By contrast, the twin-candidate (TC) model achieves a significantly ($p \leq 0.05$, by two-tailed t-test) higher success of 73.9% and 91.9%, respectively. Overall, for the whole pronoun resolution, the baseline system under the TC model yields a success 81.9%, 6.8% higher than SC does[4]. The performance is comparable to most state-of-the-art pronoun resolution systems on the same data set.

Web-based feature vs. Corpus-based feature The third column of the table lists the results using the web-based compatibility feature for neutral pronouns. Under both SC and TC models, incorporation of the web-based feature significantly boosts the performance of the baseline: For the best system in the SC model and the TC model, the success rate is improved significantly by around 4.9% and 5.3%, respectively. A similar pattern of improvement could be seen for the corpus-based semantic feature. However, the increase is not as large as using the web-based feature: Under the two learning models, the success rate of the best system with the corpus-based feature rises by up to 2.0% and 2.8% respectively, about 2.9% and 2.5% less than that of the counterpart systems with the web-based feature. The larger size and the better counts of the web against the corpus, as reported in Section 4.2,

[3] All the counts were log-transformed and the correlation coefficients were evaluated based on Pearsons' r.

[4] The improvement against SC is higher than that reported in (Yang et al., 2003). It should be because we now used 150 training documents rather than 30 ones as in the previous work. The TC model would benefit from larger training data set as it uses more features (more than double) than SC.

should contribute to the better performance.

Single-candidate model vs. Twin-Candidate model The difference between the SC and the TC model is obvious from the table. For the N-Pron and P-Pron resolution, the systems under TC could outperform the counterpart systems under SC by above 5% and 8% success, respectively. In addition, the utility of the statistics-based semantic feature is more salient under TC than under SC for N-Pron resolution: the best gains using the corpus-based and the web-based semantic features under TC are 2.9% and 5.3% respectively, higher than those under the SC model using either un-normalized semantic features (1.6% and 3.3%), or normalized semantic features (2.0% and 4.9%). Although under SC, the normalized semantic feature could result in a gain close to under TC, its utility is not stable: with metric *frequency*, using the normalized feature performs even worse than using the un-normalized one. These results not only affirm the claim by Yang et al. (2003) that the TC model is superior to the SC model for pronoun resolution, but also indicate that TC is more reliable than SC in applying the statistics-based semantic feature, for N-Pron resolution.

Web+TC vs. Other combinations The above analysis has exhibited the superiority of the web over the corpus, and the TC model over the SC model. The experimental results also reveal that using the the web-based semantic feature together with the TC model is able to further boost the resolution performance for neutral pronouns. The system with such a Web+TC combination could achieve a high success of 79.2%, defeating all the other possible combinations. Especially, it considerably outperforms (up to 11.5% success) the system with the Corpus+SC combination, which is commonly adopted in previous work (e.g., Kehler et al. (2004)).

Personal pronoun resolution vs. Neutral pronoun resolution Interestingly, the statistics-based semantic feature has no effect on the resolution of personal pronouns, as shown in the table 2. We found in the learned decision trees such a feature did not occur (SC) or only occurred in bottom nodes (TC). This should be because personal pronouns have strong restriction on the semantic category (i.e., *human*) of the candidates. A non-human candidate, even with a high predicate-argument statistics, could not be used as the antecedent (e.g. *company_said* in the sentence "*... the company ... he said ...*"). In fact, our analysis of the current data set reveals that most P-Prons refer back to a P-Pron or NE candidate whose semantic category (*human*) has been determined. That is, simply using features *NE* and *Pron* is sufficient to guarantee a high success, and thus the relatively weak semantic feature would not be taken in the learned decision tree for resolution.

Feature Group	Isolated	Combined
SemMag (Web-based)	61.2	61.2
Type+Reflexive	53.1	61.2
ParaStruct	53.1	61.2
Pron+DefNP+InDefNP+NE	57.1	67.8
NearestNP+SameSent	53.1	70.2
FirstNP	**65.3**	**79.2**

Table 4: Results of different feature groups under the TC model for N-pron resolution

```
SameSent_1 = 0:
:..SemMag > 0:
:  :..Pron_2 = 0: 10 (200/23)
:  :  Pron_2 = 1: ...
:  SemMag <= 0:
:  :..Pron_2 = 1: 01 (75/1)
:     Pron_2 = 0:
:     :..SemMag <= -28: 01 (110/19)
:        SemMag > -28: ...
SameSent_1 = 1:
:..SameSent_2 = 0: 01 (1655/49)
   SameSent_2 = 1:
   :..FirstNP_2 = 1: 01 (104/1)
      FirstNP_2 = 0:
      :..ParaStruct_2 = 1: 01 (3)
         ParaStruct_2 = 0:
         :..SemMag <= -151: 01 (27/2)
            SemMag > -151:...
```

Figure 1: Top portion of the decision tree learned under TC model for N-pron resolution (features ended with "_1" are for the first candidate C_1 and those with "_2" are for C_2.)

4.4 Feature Analysis

In our experiment we were also concerned about the importance of the web-based compatibility feature (using *frequency* metric) among the feature set. For this purpose, we divided the features into groups, and then trained and tested on one group at a time. Table 4 lists the feature groups and their respective results for N-Pron resolution under the TC model.

The second column is for the systems with only the current feature group, while the third column is with the features combined with the existing feature set. We see that used in isolation, the semantic compatibility feature is able to achieve a success up to 61% around, just 4% lower than the best indicative feature *FirstNP*. In combination with other features, the performance could be improved by as large as 18% as opposed to being used alone.

Figure 1 shows the top portion of the pruned decision tree for N-Pron resolution under the TC model. We could find that: (i) When comparing two candidates which occur in the same sentence as the anaphor, the web-based semantic feature would be examined in the first place, followed by the lexical property of the candidates. (ii) When two non-pronominal candidates are both in previous sentences before the anaphor, the web-based semantic feature is still required to be examined after *FirstNP* and *ParaStruct*. The decision tree further indicates that the web-based feature plays an important role in N-Pron resolution.

5 Conclusion

Our research focussed on improving pronoun resolution using the statistics-based semantic compatibility information. We explored two issues that affect the utility of the semantic information: statistics source and learning framework. Specifically, we proposed to utilize the web and the twin-candidate model, in addition to the common combination of the corpus and single-candidate model, to compute and apply the semantic information.

Our experiments systematically evaluated different combinations of statistics sources and learning models. The results on the newswire domain showed that the web-based semantic compatibility could be the most effectively incorporated in the twin-candidate model for the neutral pronoun resolution. While the utility is not obvious for personal pronoun resolution, we can still see the improvement on the overall performance. We believe that the semantic information under such a configuration would be even more effective on technical domains where neutral pronouns take the majority in the pronominal anaphors. Our future work would have a deep exploration on such domains.

References

S. Abney. 1996. Partial parsing via finite-state cascades. In *Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information*, pages 8–15.

D. Bean and E. Riloff. 2004. Unsupervised learning of contextual role knowledge for coreference resolution. In *Proceedings of 2004 North American chapter of the Association for Computational Linguistics annual meeting*.

I. Dagan and A. Itai. 1990. Automatic processing of large corpora for the resolution of anahora references. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 330–332.

J. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:339–352.

A. Kehler, D. Appelt, L. Taylor, and A. Simma. 2004. The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of 2004 North American chapter of the Association for Computational Linguistics annual meeting*.

F. Keller and M. Lapata. 2003. Using the web to obtain freqencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

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

N. Modjeska, K. Markert, and M. Nissim. 2003. Using the web in machine learning for other-anaphora resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 176–183.

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. Poesio, R. Mehta, A. Maroudas, and J. Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of 42th Annual Meeting of the Association for Computational Linguistics*.

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

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.

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.

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.

G. Zhou, J. Su, and T. Tey. 2000. Hybrid text chunking. In *Proceedings of the 4th Conference on Computational Natural Language Learning*, pages 163–166, Lisbon, Portugal.