A Machine Learning Approach to Pronoun Resolution in Spoken Dialogue

Michael Strube and **Christoph Müller**
European Media Laboratory GmbH
Villa Bosch
Schloß-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
{michael.strube|christoph.mueller}@eml.villa-bosch.de

Abstract

We apply a decision tree based approach to pronoun resolution in spoken dialogue. Our system deals with pronouns with NP- and non-NP-antecedents. We present a set of features designed for pronoun resolution in spoken dialogue and determine the most promising features. We evaluate the system on twenty Switchboard dialogues and show that it compares well to Byron's (2002) manually tuned system.

1 Introduction

Corpus-based methods and machine learning techniques have been applied to anaphora resolution in written text with considerable success (Soon et al., 2001; Ng & Cardie, 2002, among others). It has been demonstrated that systems based on these approaches achieve a performance that is comparable to hand-crafted systems. Since they can easily be applied to new domains it seems also feasible to port a given corpus-based anaphora resolution system from written text to spoken dialogue. This paper describes the extensions and adaptations needed for applying our anaphora resolution system (Müller et al., 2002; Strube et al., 2002) to pronoun resolution in spoken dialogue.

There are important differences between written text and spoken dialogue which have to be accounted for. The most obvious difference is that in spoken dialogue there is an abundance of (personal and demonstrative) pronouns with non-NP-antecedents or no antecedents at all. Corpus studies have shown that a significant amount of pronouns in spoken dialogue have non-NP-antecedents: Byron & Allen (1998) report that about 50% of the pronouns in the TRAINS93 corpus have non-NP-antecedents. Eckert & Strube (2000) note that only about 45% of the pronouns in a set of Switchboard dialogues have NP-antecedents. The remainder consists of 22% which have non-NP-antecedents and 33% without antecedents. These studies suggest that the performance of a pronoun resolution algorithm can be improved considerably by enabling it to resolve also pronouns with non-NP-antecedents.

Because of the difficulties a pronoun resolution algorithm encounters in spoken dialogue, previous approaches were applied only to tiny domains, they needed deep semantic analysis and discourse processing and relied on hand-crafted knowledge bases. In contrast, we build on our existing anaphora resolution system and incrementally add new features specifically devised for spoken dialogue. That way we are able to determine relatively powerful yet computationally cheap features. To our knowledge the work presented here describes the first implemented system for corpus-based anaphora resolution dealing also with non-NP-antecedents.

2 NP- vs. Non-NP-Antecedents

Spoken dialogue contains more pronouns with non-NP-antecedents than written text does. However, pronouns with NP-antecedents (like 3rd pers. masculine/feminine pronouns, cf. *he* in the example below) still constitute the largest fraction of all coreferential pronouns in the Switchboard corpus.

In spoken dialogue there are considerable numbers of pronouns that pick up different kinds of abstract objects from the previous discourse, e.g. events, states, concepts, propositions or facts (Webber, 1991; Asher, 1993). These anaphors then have VP-antecedents (*"it_j"* in (B6) below) or sentential antecedents (*"$that_k$"* in (B5)).

A1: …[he]$_i$'s nine months old. …

A2: [He]$_i$ likes to dig around a little bit.

A3: [His]$_i$ mother comes in and says, why did you let [him]$_i$ [play in the dirt]$_j$,

A:4 I guess [[he]$_i$'s enjoying himself]$_k$.

B5: [That]$_k$'s right.

B6: [It]$_j$'s healthy, …

A major problem for pronoun resolution in spoken dialogue is the large number of personal and demonstrative pronouns which are either not referential at all (e.g. expletive pronouns) or for which a particular antecedent cannot easily be determined by humans (called *vague* anaphors by Eckert & Strube (2000)).

In the following example, the *"$that_i$"* in utterance (A3) refers back to utterance (A1). As for the first two pronouns in (B4), following Eckert & Strube (2000) and Byron (2002) we assume that referring expressions in disfluencies, abandoned utterances etc. are excluded from the resolution. The third pronoun in (B4) is an expletive. The pronoun in (A5) is different in that it is indeed referential: it refers back to *"$that_i$"* from (A3).

A1: …[There is a lot of theft, a lot of assault dealing with, uh, people trying to get money for drugs.$_i$]

B2: Yeah.

A3: And, uh, I think [that$_i$]'s a national problem, though.

B4: *It, it, it*'s pretty bad here, too.

A5: [It$_i$]'s not unique …

Pronoun resolution in spoken dialogue also has to deal with the whole range of difficulties that come with processing spoken language: disfluencies, hesitations, abandoned utterances, interruptions, backchannels, etc. These phenomena have to be taken into account when formulating constraints on e.g. the search space in which an anaphor looks for its antecedent. E.g., utterance (B2) in the previous example does not contain any referring expressions. So the demonstrative pronoun in (A3) has to have access not only to (B2) but also to (A1).

3 Data

3.1 Corpus

Our work is based on twenty randomly chosen Switchboard dialogues. Taken together, the dialogues contain 30810 tokens (words and punctuation) in 3275 sentences / 1771 turns. The annotation consists of 16601 markables, i.e. sequences of words and attributes associated with them. On the top level, different types of markables are distinguished: *NP*-markables identify referring expressions like noun phrases, pronouns and proper names. Some of the attributes for these markables are derived from the Penn Treebank version of the Switchboard dialogues, e.g. grammatical function, NP form, grammatical case and depth of embedding in the syntactical structure. *VP*-markables are verb phrases, *S*-markables sentences. *Disfluency*-markables are noun phrases or pronouns which occur in unfinished or abandoned utterances. Among other (type-dependent) attributes, markables contain a *member* attribute with the ID of the coreference class they are part of (if any). If an expression is used to refer to an entity that is not referred to by any other expression, it is considered a singleton.

Table 1 gives the distribution of the *npform* attribute for *NP*-markables. The second and third row give the number of non-singletons and singletons respectively that add up to the total number given in the first row.

Table 2 shows the distribution of the *agreement* attribute (i.e. person, gender, and number) for the pronominal expressions in our corpus. The left figure in each cell gives the total number of expressions, the right figure gives the number of non-singletons. Note the relatively high number of singletons among the personal and demonstrative pronouns (223 for *it*, 60 for *they* and 82 for *that*). These pronouns are either expletive or vague, and cause the most trouble for a pronoun resolution algorithm, which will usually attempt to find an antecedent nonetheless. Singleton *they* pronouns, in particular, are typical for spoken language (as opposed to

	defNP	indefNP	NNP	prp	prp$	dtpro
Total	1080	1899	217	1075	70	392
In coreference relation	219	163	94	786	56	309
Singletons	861	1736	123	289	14	83

Table 1: Distribution of *npform* Feature on Markables (w/o 1st and 2nd Persons)

	3m		**3f**		**3n**		**3p**	
prp	67	63	49	47	**541**	**318**	**418**	**358**
prp$	18	15	14	11	**3**	**3**	**35**	**27**
dtpro	0	0	0	0	**380**	**298**	**12**	**11**
Σ	85	78	63	58	**924**	**619**	**465**	**396**

Table 2: Distribution of Agreement Feature on Pronominal Expressions

written text). The same is true for anaphors with non-NP-antecedents. However, while they are far more frequent in spoken language than in written text, they still constitute only a fraction of all coreferential expressions in our corpus. This defines an upper limit for what the resolution of these kinds of anaphors can contribute at all. These facts have to be kept in mind when comparing our results to results of coreference resolution in written text.

3.2 Data Generation

Training and test data instances were generated from our corpus as follows. All markables were sorted in document order, and markables for first and second person pronouns were removed. The resulting list was then processed from top to bottom. If the list contained an *NP*-markable at the current position and if this markable was not an indefinite noun phrase, it was considered a potential anaphor. In that case, pairs of potentially coreferring expressions were generated by combining the potential anaphor with each *compatible*[1] *NP*-markable preceding[2] it in the list. The resulting pairs were labelled *P* if both markables had the same (non-empty) value in their *member* attribute, *N* otherwise. For anaphors with non-NP-antecedents, *additional* training and test data instances had to be generated. This process was triggered by the markable at the current position being *it* or *that*. In that case, a small set of potential non-NP-antecedents was generated by selecting *S*- and *VP*-markables from the last two valid sentences preceding the potential anaphor. The choice of the last *two* sentences was motivated pragmatically by considerations to keep the search space (and the number of instances) small. A sentence was considered valid if it was neither unfinished nor a backchannel utterance (like e.g. *"Uh-huh"*, *"Yeah"*, etc.). From the selected markables, inaccessible non-NP-expressions were automatically removed. We considered an expression inaccessible if it *ended before* the sentence in which it was contained. This was intended to be a rough approximation of the concept of the right frontier (Webber, 1991). The remaining expressions were then combined with the potential anaphor. Finally, the resulting pairs were labelled *P* or *N* and added to the instances generated with *NP*-antecedents.

4 Features

We distinguish two classes of features: NP-level features specify e.g. the grammatical function, NP form, morpho-syntax, grammatical case and the depth of embedding in the syntactical structure. For these features, each instance contains one value for the antecedent and one for the anaphor. Coreference-level features, on the other hand, describe the relation between antecedent and anaphor in terms of e.g. distance (in words, markables and sentences), compatibility in terms of agreement and identity of syntactic function. For these features, each instance contains only one value.

In addition, we introduce a set of features which is partly tailored to the processing of spoken dialogue. The feature ante_exp_type (17) is a rather obvious yet useful feature to distinguish NP- from non-NP-antecedents. The features ana_np_, vp_ and

[1]Markables are considered compatible if they do not mismatch in terms of agreement.

[2]We disregard the phenomenon of *cataphor* here.

	NP-level features	
1.	ante_gram_func	grammatical function of antecedent
2.	ante_npform	form of antecedent
3.	ante_agree	person, gender, number
4.	ante_case	grammatical case of antecedent
5.	ante_s_depth	the level of embedding in a sentence
6.	ana_gram_func	grammatical function of anaphor
7.	ana_npform	form of anaphor
8.	ana_agree	person, gender, number
9.	ana_case	grammatical case of anaphor
10.	ana_s_depth	the level of embedding in a sentence
	Coreference-level features	
11.	agree_comp	compatibility in agreement between anaphor and antecedent
12.	npform_comp	compatibilty in NP form between anaphor and antecedent
13.	wdist	distance between anaphor and antecedent in words
14.	mdist	distance between anaphor and antecedent in markables
15.	sdist	distance between anaphor and antecedent in sentences
16.	syn_par	anaphor and antecedent have the same grammatical function (yes, no)
	Features introduced for spoken dialogue	
17.	ante_exp_type	type of antecedent (NP, S, VP)
18.	ana_np_pref	preference for NP arguments
19.	ana_vp_pref	preference for VP arguments
20.	ana_s_pref	preference for S arguments
21.	mdist_3mf3p	(see text)
22.	mdist_3n	(see text)
23.	ante_tfidf	(see text)
24.	ante_ic	(see text)
25.	wdist_ic	(see text)

Table 3: Our Features

s_pref (18, 19, 20) describe a verb's preference for arguments of a particular type. Inspired by the work of Eckert & Strube (2000) and Byron (2002), these features capture preferences for NP- or non-NP-antecedents by taking a pronoun's predicative context into account. The underlying assumption is that if a verb preceding a personal or demonstrative pronoun preferentially subcategorizes sentences or VPs, then the pronoun will be likely to have a non-NP-antecedent. The features are based on a verb list compiled from 553 Switchboard dialogues.[3] For every verb occurring in the corpus, this list contains up to three entries giving the absolute count of cases where the verb has a direct argument of type *NP*, *VP* or *S*. When the verb list was produced, pronominal arguments were ignored. The features mdist_3mf3p and mdist_3n (21, 22) are refinements of the mdist feature. They measure the distance in markables between antecedent and anaphor, but in doing so they take the agreement value of the anaphor into account. For anaphors with an agreement value of 3mf or 3p, mdist_3mf3p is measured as D = 1 + the number of *NP*-markables between anaphor and potential antecedent. Anaphors with an agreement value of 3n, (i.e. *it* or *that*), on the other hand, potentially have non-NP-antecedents, so mdist_3n is measured as D + the number of anaphorically accessible[4] *S*- and *VP*-markables between anaphor and potential antecedent.

The feature ante_tfifd (23) is supposed to capture the relative importance of an expression for a dialogue. The underlying assumption is that the higher the importance of a non-NP expression, the higher the probability of its being referred back to. For our purposes, we calculated TF for every word by counting its frequency in each of our twenty Switchboard dialogues separately. The calculation of IDF was based on a set of 553 Switchboard dialogues. For every word, we calculated IDF as log(553/N_w), with N_w=number of documents containing the word. For every non-NP-markable, an *average* TF*IDF value was calculated as the TF*IDF sum of all words comprising the markable, divided by the number of

[3]It seemed preferable to compile our own list instead of using existing ones like Briscoe & Carroll (1997).

[4]As mentioned earlier, the definition of accessibility of non-NP-antecedents is inspired by the concept of the right frontier (Webber, 1991).

words in the markable. The feature ante_ic (24) as an alternative to ante_tfidf is based on the same assumptions as the former. The *information content* of a non-NP-markable is calculated as follows, based on a set of 553 Switchboard dialogues: For each word in the markable, the IC value was calculated as the negative log of the total frequency of the word divided by the total number of words in all 553 dialogues. The *average* IC value was then calculated as the IC sum of all words in the markable, divided by the number of words in the markable. Finally, the feature wdist_ic (25) measures the word-based distance between two expressions. It does so in terms of the *sum of the individual words' IC*. The calculation of the IC was done as described for the ante_ic feature.

5 Experiments and Results

5.1 Experimental Setup

All experiments were performed using the decision tree learner *RPART* (Therneau & Atkinson, 1997), which is a CART (Breiman et al., 1984) reimplementation for the S-Plus and R statistical computing environments (we use R, Ihaka & Gentleman (1996), see http://www.r-project.org). We used the standard pruning and control settings for RPART (cp=0.0001, minsplit=20, minbucket=7). All results reported were obtained by performing 20-fold cross-validation.

In the prediction phase, the trained classifier is exposed to unlabeled instances of test data. The classifier's task is to label each instance. When an instance is labeled as coreferring, the IDs of the anaphor and antecedent are kept in a *response list* for the evaluation according to Vilain et al. (1995).

For determining the relevant feature set we followed an iterative procedure similar to the *wrapper* approach for feature selection (Kohavi & John, 1997). We start with a model based on a set of predefined baseline features. Then we train models combining the baseline with all additional features separately. We choose the best performing feature (f-measure according to Vilain et al. (1995)), adding it to the model. We then train classifiers combining the enhanced model with each of the remaining features separately. We again choose the best performing classifier and add the corresponding new feature to the model. This process is repeated as long as significant improvement can be observed.

5.2 Results

In our experiments we split the data in three sets according to the agreement of the anaphor: third person masculine and feminine pronouns (3mf), third person neuter pronouns (3n), and third person plural pronouns (3p). Since only 3n-pronouns have non-NP-antecedents, we were mainly interested in improvements in this data set.

We used the same baseline model for each data set. The baseline model corresponds to a pronoun resolution algorithm commonly applied to written text, i.e., it uses only the features in the first two parts of Table 3. For the baseline model we generated training and test data which included only NP-antecedents.

Then we performed experiments using the features introduced for spoken dialogue. The training and test data for the models using additional features included NP- and non-NP-antecedents. For each data set we followed the iterative procedure outlined in Section 5.1.

In the following tables we present the results of our experiments. The first column gives the number of coreference links correctly found by the classifier, the second column gives the number of all coreference links found. The third column gives the total number of coreference links (1250) in the corpus. During evaluation, the list of all correct links is used as the *key list* against which the *response list* produced by the classifier (cf. above) is compared. The remaining three columns show precision, recall and f-measure, respectively.

Table 4 gives the results for 3mf pronouns. The baseline model performs very well on this data set (the low recall figure is due to the fact that the 3mf data set contains only a small subset of the coreference links expected by the evaluation). The results are comparable to any pronoun resolution algorithm dealing with written text. This shows that our pronoun resolution system could be ported to the spoken dialogue domain without sacrificing performance.

Table 5 shows the results for 3n pronouns. The baseline model does not perform very well. As mentioned above, for evaluating the performance of the

	correct found	total found	total correct	precision	recall	f-measure
baseline, features 1-16	120	150	1250	80.00	9.60	17.14
plus mdist_3mf3p	121	153	1250	79.08	9.68	17.25

Table 4: Results for Third Person Masculine and Feminine Pronouns (3mf)

	correct found	total found	total correct	precision	recall	f-measure
baseline, features 1-16	109	235	1250	46.38	8.72	14.68
plus none	97	232	1250	41.81	7.76	13.09
plus ante_exp_type	137	359	1250	38.16	10.96	17.03
plus wdist_ic	154	389	1250	39.59	12.32	18.79
plus ante_tfidf	158	391	1250	40.41	12.64	19.26

Table 5: Results for Third Person Neuter Pronouns (3n)

baseline model we removed all potential non-NP-antecedents from the data. This corresponds to a naive application of a model developed for written text to spoken dialogue.

First, we applied the same model to the data set containing all kinds of antecedents. The performance drops somewhat as the classifier is exposed to non-NP-antecedents without being able to differentiate between NP- and non-NP-antecedents. By adding the feature ante_exp_type the classifier is enabled to address NP- and non-NP-antecedents differently, which results in a considerable gain in performance. Substituting the wdist feature with the wdist_ic feature also improves the performance considerably. The ante_tfidf feature only contributes marginally to the overall performance. – These results show that it pays off to consider features particularly designed for spoken dialogue.

Table 6 presents the results for 3p pronouns, which do not have non-NP-antecedents. Many of these pronouns do not have an antecedent at all. Others are *vague* in that human annotators felt them to be referential, but could not determine an antecedent. Since we did not address that issue in depth, the classifier tries to find antecedents for these pronouns indiscriminately, which results in rather low precision figures, as compared to e.g. those for 3mf. Only the feature wdist_ic leads to an improvement over the baseline.

Table 7 shows the results for the combined classifiers. The improvement in f-measure is due to the increase in recall while the precision shows only a slight decrease.

Though some of the features of the baseline model (features 1-16) still occur in the decision tree learned, the feature ante_exp_type divides major parts of the tree quite nicely (see Figure 1). Below that node the feature ana_npform is used to distinguish between negative (personal pronouns) and potential positive cases (demonstrative pronouns). This confirms the hypothesis by Eckert & Strube (2000) and Byron (2002) to give high priority to these features. The decision tree fragment in Figure 1 correctly assigns the *P* label to 23-7=16 sentential antecedents.

```
split, n, loss, yval
  * denotes terminal node

...
  anteexptype=s,vp 1110 55 N
    ananpform=prp 747,11 N *
    ananpform=dtpro 363 44 N
      anteexptype=vp 177 3 N *
      anteexptype=s 186 41 N
        udist>=1.5 95 14 N *
        udist<1.5 91 27 N
          wdistic<43.32 33 4 N *
          wdistic>=43.32 58 23 N
            anasdepth>=2.5 23 4 N *
            anasdepth<2.5 35 16 N
              wdistic>=63.62 24 11 N
                wdistic<80.60 12 3 N *
                wdistic>=80.60 12 4 P *
              wdistic<63.62 11 3 P *
```

Figure 1: Decision Tree Fragment

However, the most important problem is the large amount of pronouns without antecedents. The model does find (wrong) antecedents for a lot of pronouns which should not have one. Only a small fraction of these pronouns are true expletives (i.e., they precede a "weather" verb or are in constructions like *"It seems that ... "*. The majority of these cases are referential, but have no antecedent in the data (i.e.,

	correct found	total found	total correct	precision	recall	f-measure
baseline, features 1-16	227	354	1250	64.12	18.16	28.30
plus wdist_ic	230	353	1250	65.16	18.40	28.70

Table 6: Results for Third Person Plural Pronouns (3p)

	correct found	total found	total correct	precision	recall	f-measure
baseline, features 1-16	456	739	1250	61.71	36.48	45.85
combined	509	897	1250	56.74	40.72	47.42

Table 7: Combined Results for All Pronouns

they are *vague* pronouns).

The overall numbers for precision, recall and f-measure are fairly low. One reason is that we did not attempt to resolve anaphoric definite NPs and proper names though these coreference links are contained in the evaluation key list. If we removed them from there, the recall of our experiments would approach the 51% Byron (2002) mentioned for her system using only domain-independent semantic restrictions.

6 Comparison to Related Work

Our approach for determining the feature set for pronoun resolution resembles the so-called *wrapper* approach for feature selection (Kohavi & John, 1997). This is in contrast to the majority of other work on feature selection for anaphora resolution, which was hardly ever done systematically. E.g. Soon et al. (2001) only compared baseline systems consisting of one feature each, only three of which yielded an f-measure greater than zero. Then they combined these features and achieved results which were close to the best overall results they report. While this tells us which features contribute a lot, it does not give any information about potential (positive or negative) influence of the rest. Ng & Cardie (2002) select the set of features by hand, giving a preference to high precision features. They admit that this method is quite subjective.

Corpus-based work about pronoun resolution in spoken dialogue is almost non-existent. However, there are a few papers dealing with neuter pronouns with NP-antecedents. E.g., Dagan & Itai (1991) presented a corpus-based approach to the resolution of the pronoun *it*, but they use a written text corpus and do not mention non-NP-antecedents at all. Paul et al. (1999) presented a corpus-based anaphora resolution algorithm for spoken dialogue. For their experiments, however, they restricted anaphoric relations to those with NP-antecedents.

Byron (2002) presented a symbolic approach which resolves pronouns with NP- and non-NP-antecedents in spoken dialogue in the TRAINS domain. Byron extends a pronoun resolution algorithm (Tetrault, 2001) with *semantic filtering*, thus enabling it to resolve anaphors with non-NP-antecedents as well. Semantic filtering relies on knowledge about semantic restrictions associated with verbs, like semantic compatibility between subject and predicative noun or predicative adjective.

An evaluation on ten TRAINS93 dialogues with 80 3rd person pronouns and 100 demonstrative pronouns shows that semantic filtering and the implementation of different search strategies for personal and demonstrative pronouns yields a success rate of 72%. As Byron admits, the major limitation of her algorithm is its dependence on domain-dependent resources which cover the domain entirely. When evaluating her algorithm with only domain-independent semantics, Byron achieved 51% success rate. What is problematic with her approach is that she assumes the input to her algorithm to be only referential pronouns. This simplifies the task considerably.

7 Conclusions and Future Work

We presented a machine learning approach to pronoun resolution in spoken dialogue. We built upon a system we used for anaphora resolution in written text and extended it with a set of features designed for spoken dialogue. We refined distance features and used metrics from information retrieval for determining non-NP-antecedents. Inspired by the more linguistically oriented work by Eckert & Strube (2000) and Byron (2002) we also evaluated the contribution of features which used the predicative context of the pronoun to be resolved. However, these features did not show up in the final models since they did not lead to an improvement. Instead, rather simple distance metrics were preferred. While we were (almost) satisfied with the performance of these features, the major problem for a spoken dialogue pronoun resolution algorithm is the abundance of pronouns without antecedents. Previous research could avoid dealing with this phenomenon by either applying the algorithm by hand (Eckert & Strube, 2000) or excluding these cases (Byron, 2002) from the evaluation. Because we included these cases in our evaluation we consider our approach at least comparable to Byron's system when she uses only domain-independent semantics. We believe that our system is more robust than hers and that it can more easily be ported to new domains.

Acknowledgements. The work presented here has been partially funded by the German Ministry of Research and Technology as part of the EMBASSI project (01 IL 904 D/2) and by the Klaus Tschira Foundation. We would like to thank Susanne Wilhelm and Lutz Wind for doing the annotations, Kerstin Schürmann, Torben Pastuch and Klaus Rothenhäusler for helping with the data preparation.

References

Asher, Nicholas (1993). *Reference to Abstract Objects in Discourse*. Dordrecht, The Netherlands: Kluwer.

Breiman, Leo, Jerome H. Friedman, Charles J. Stone & R.A. Olshen (1984). *Classification and Regression Trees*. Belmont, Cal.: Wadsworth and Brooks/Cole.

Briscoe, Ted & John Carroll (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing,* Washington, D.C., 31 March – 3 April 1997, pp. 356–363.

Byron, Donna K. (2002). Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* Philadelphia, Penn., 7–12 July 2002, pp. 80–87.

Byron, Donna K. & James F. Allen (1998). Resolving demonstrative pronouns in the TRAINS93 corpus. In *New Approaches to Discourse Anaphora: Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC2)*, pp. 68–81.

Dagan, Ido & Alon Itai (1991). A statistical filter for resolving pronoun references. In Y.A. Feldman & A. Bruckstein (Eds.), *Artificial Intelligence and Computer Vision*, pp. 125–135. Amsterdam: Elsevier.

Eckert, Miriam & Michael Strube (2000). Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Ihaka, Ross & Robert Gentleman (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

Kohavi, Ron & George H. John (1997). Wrappers for feature subset selection. *Artificial Intelligence Journal*, 97(1-2):273–324.

Müller, Christoph, Stefan Rapp & Michael Strube (2002). Applying Co-Training to reference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* Philadelphia, Penn., 7–12 July 2002, pp. 352–359.

Ng, Vincent & Claire Cardie (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* Philadelphia, Penn., 7–12 July 2002, pp. 104–111.

Paul, Michael, Kazuhide Yamamoto & Eiichiro Sumita (1999). Corpus-based anaphora resolution towards antecedent preference. In *Proc. of the 37th ACL, Workshop Coreference and Its Applications,* College Park, Md., 1999, pp. 47–52.

Soon, Wee Meng, Hwee Tou Ng & Daniel Chung Yong Lim (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Strube, Michael, Stefan Rapp & Christoph Müller (2002). The influence of minimum edit distance on reference resolution. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing,* Philadelphia, Pa., 6–7 July 2002, pp. 312–319.

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

Therneau, Terry M. & Elizabeth J. Atkinson (1997). *An introduction to recursive partitioning using the RPART routines*. Technical Report: Mayo Foundation. Distributed with the RPART package.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly & Lynette Hirschman (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pp. 45–52. San Mateo, Cal.: Morgan Kaufmann.

Webber, Bonnie L. (1991). Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.