Automatic Detection of Nonreferential *It* in Spoken Multi-Party Dialog

Christoph Müller
EML Research gGmbH
Villa Bosch
Schloß-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
christoph.mueller@eml-research.de

Abstract

We present an implemented machine learning system for the automatic detection of nonreferential *it* in spoken dialog. The system builds on shallow features extracted from dialog transcripts. Our experiments indicate a level of performance that makes the system usable as a preprocessing filter for a coreference resolution system. We also report results of an annotation study dealing with the classification of *it* by naive subjects.

1 Introduction

This paper describes an implemented system for the detection of nonreferential *it* in spoken multi-party dialog. The system has been developed on the basis of meeting transcriptions from the ICSI Meeting Corpus (Janin et al., 2003), and it is intended as a preprocessing component for a coreference resolution system in the DIANA-Summ dialog summarization project. Consider the following utterance:

MN059: Yeah. Yeah. Yeah. I'm sure I could learn a lot about um, yeah, just how to - how to come up with these structures, cuz **it**'s - **it**'s very easy to whip up something quickly, but **it** maybe then makes sense to - to me, but not to anybody else, and - and if we want to share and integrate things, they must - well, they must be well designed really. (Bed017)

In this example, only one of the three instances of *it* is a referential pronoun: The first *it* appears in the *reparandum* part of a speech repair (Heeman & Allen, 1999). It is replaced by a subsequent *alteration* and is thus not part of the final utterance. The second *it* is the subject of an extraposition construction and serves as the placeholder for the postposed infinitive phrase *to whip up something quickly*. Only the third *it* is a referential pronoun which anaphorically refers to *something*.

The task of the system described in the following is to identify and filter out nonreferential instances of *it*, like the first and second one in the example. By preventing these instances from triggering the search for an antecedent, the precision of a coreference resolution system is improved.

Up to the present, coreference resolution has mostly been done on written text. In this domain, the detection of nonreferential *it* has by now become a standard preprocessing step (e.g. Ng & Cardie (2002)). In the few works that exist on coreference resolution in spoken language, on the other hand, the problem could be ignored, because almost none of these aimed at developing a system that could handle unrestricted input. Eckert & Strube (2000) focus on an unimplemented algorithm for determining the type of antecedent (mostly NP vs. non-NP), given an anaphorical pronoun or demonstrative. The system of Byron (2002) is implemented, but deals mainly with how referents for already identified discourse-deictic anaphors can be created. Finally, Strube & Müller (2003) describe an implemented system for resolving 3rd person pronouns in spoken dialog, but they also exclude nonreferential *it* from consideration. In contrast, the present work is part of a project to develop a coreference resolution system that, in its final implementation, can handle *unrestricted* multi-party dialog. In such a system, no *a priori* knowledge is available about whether an instance of *it* is referential or not.

The remainder of this paper is structured as follows: Section 2 describes the current state of the art for the detection of nonreferential *it* in written text. Section 3 describes our corpus of transcribed spoken dialog. It also reports on the annotation that we performed in order to collect training and test data for our machine learning experiments. The annotation also offered interesting insights into how reliably humans can identify nonreferential *it* in spoken language, a question that,

to our knowledge, has not been adressed before. Section 4 describes the setup and results of our machine learning experiments, Section 5 contains conclusion and future work.

2 Detecting Nonreferential *It* In Text

Nonreferential *it* is a rather frequent phenomenon in written text, though it still only constitutes a minority of all instances of *it*. Evans (2001) reports that his corpus of approx. 370.000 words from the SUSANNE corpus and the BNC contains 3.171 examples of *it*, approx. 29% of which are nonreferential. Dimitrov et al. (2002) work on the ACE corpus and give the following figures: the newspaper part of the corpus (ca. 61.000 words) contains 381 instances of *it*, with 20.7% being nonreferential, and the news wire part (ca. 66.000 words) contains 425 instances of *it*, 16.5% of which are nonreferential. Boyd et al. (2005) use a 350.000 word corpus from a variety of genres. They count 2.337 instances of *it*, 646 of which (28%) are nonreferential. Finally, Clemente et al. (2004) report that in the GENIA corpus of medical abstracts the percentage of nonreferential *it* is as high as 44% of all instances of *it*. This may be due to the fact that abstracts tend to contain more stereotypical formulations.

It is worth noting here that in all of the above studies the referential-nonreferential decision implicitly seems to have been made by the author(s). To our knowledge, no study provides figures regarding the reliability of this classification.

Paice & Husk (1987) is the first corpus-based study on the detection of nonreferential *it* in written text. From examples drawn from a part of the LOB corpus (technical section), Paice & Husk (1987) create rather complex pattern-based rules (like ***SUBJECT VERB** it **STATUS** to **TASK***), and apply them to an unseen part of the corpus. They report a final success rate of 92.2% on the test corpus. Nowadays, most current coreference resolution systems for written text include some means for the detection of nonreferential *it*. However, evaluation figures for this task are not always given. As the detection of nonreferential *it* is supposed to be a *filtering* condition (as opposed to a *selection* condition), high precision is normally considered to be more important than high recall. A *false negative*, i.e. a nonreferential *it* that is not detected, can still be filtered out later when resolution fails, while a *false positive*, i.e. a referential *it* that is wrongly removed, is simply lost and will necessarily harm overall recall. Another point worth mentioning is that mere classification accuracy (percent correct) is not an appropriate evaluation measure for the detection of nonreferential *it*. Accuracy will always be biased in favor of predicting the majority class *referential* which, as the above figures show, can amount to over 80%.

The majority of works on detecting nonreferential *it* in written text uses some variant of the partly syntactic and partly lexical tests described by Lappin & Leass (1994), the first work about computational pronoun resolution to address the potential benefit of detecting nonreferential *it*. Lappin & Leass (1994) mainly supply a short list of modal adjectives and cognitive verbs, as well as seven syntactic patterns like *It is **Cogv-ed** that **S***. Like many works that treat the detection of nonreferential *it* only as one of several steps of the coreference resolution process, Lappin & Leass (1994) do not give any figures about the performance of this filtering method.

Dimitrov et al. (2002) modify and extend the approach of Lappin & Leass (1994) in several respects. They extend the list of modal adjectives to 86 (original: 15), and that of cognitive verbs to 22 (original: seven). They also increase the coverage of the syntactic patterns, mainly by allowing for optional adverbs at certain positions. Dimitrov et al. (2002) report performance figures for each of their syntactic patterns individually. The first thing to note is that 41.3% of the instances of nonreferential *it* in their corpus do not comply with any of the patterns they use, so even if each pattern worked perfectly, the maximum recall to be reached with this method would be 58.7%. The actual recall is 37.7%. Dimitrov et al. (2002) do not give any precision figures. One interesting detail is that the pattern involving the passive cognitive verb construction accounts for only three instances in the entire corpus, of which only one is found.

Evans (2001) employs memory-based machine learning. He represents instances of *it* as vectors of 35 features. These features encode, among other things, information about the parts of speech and lemmata of words in the context of *it* (obtained automatically). Other features encode the presence or absence of, resp. the distance to, certain element sequences indicative of pleonastic *it*, such as complementizers or present participles. Some features explicitly reference structural properties of

the text, like position of the *it* in its sentence, and position of the sentence in its paragraph. Sentence boundaries are also used to limit the search space for certain distance features. Evans (2001) reports a precision of 73.38% and a recall of 69.25%.

Clemente et al. (2004) work on the GENIA corpus of medical abstracts. They assume perfect preprocessing by using the manually assigned POS tags from the corpus. The features are very similar to those used by Evans (2001). Using an SVM machine learning approach, Clemente et al. (2004) obtain an accuracy of 95.5% (majority base line: approx. 56%). They do not report any precision or recall figures. Clemente et al. (2004) also perform an analysis of the relative importance of features in various settings. It turns out that features pertaining to the distance or number of complementizers following the *it* are consistently among the most important.

Finally, Boyd et al. (2005) also use a machine learning approach. They use 25 features, most of which represent syntactic patterns like *it* VERB ADJ *that*. These features are numeric, having as their value the distance from a given instance of *it* to the end of the match, if any. Pattern matching is limited to sentences, sentence breaks being identified by punctuation. Other features encode the (simplified) POS tags that surround a given instance of *it*. Like in the system of Clemente et al. (2004), all POS tag information is obtained from the corpus, so no (error-prone) automatic tagging is performed. Boyd et al. (2005) obtain a precision of 82% and a recall of 71% using a memory-based machine learning approach, and a similar precision but much lower recall (42%) using a decision tree classifier.

In summary, the best approaches for detecting nonreferential *it* in written text already work reasonably well, yielding an F-measure of over 70% (Evans, 2001; Boyd et al., 2005). This can at least partly be explained by the fact that many instances are drawn from texts coming from rather stereotypical domains, like e.g. news wire text or scientific abstracts. Also, some make the rather unrealistic assumption of perfect POS information, and even those who do *not* make this assumption take advantage of the fact that automatic POS tagging is generally very good for these types of text. This is especially true in the case of complementizers (like *that*) which have been shown to be highly indicative of extraposition constructions. Structural properties of the context of *it*, including sentence boundaries and position within sentence or paragraph, are also used frequently, either as numerical features in their own right, or as means to limit the search space for pattern matching.

3 Nonreferential *It* in Spoken Dialog

Spontaneous speech differs considerably from written text in at least two respects that are relevant for the task described in this paper: it is less structured and more noisy than written text, and it contains significantly more instances of *it*, including some types of nonreferential *it* not found in written text.

3.1 The ICSI Meeting Corpus

The ICSI Meeting Corpus (Janin et al., 2003) is a collection of 75 manually transcribed group discussions of about one hour each, involving 3 to 13 speakers. It features a semiautomatically generated segmentation in which the corpus developers tried to track the flow of the dialog by inserting segment starts approximately whenever a person started talking. Each of the resulting *segments* is associated with a single speaker and contains start and end time information. The transcription contains manually added punctuation, and it also explicitly records disfluencies and speech repairs by marking both *interruption points* and *word fragments* (Heeman & Allen, 1999). Consider the following example:

ME010: Yeah. Yeah. No, no. There was a whole co- There was a little contract signed. It was - Yeah. (Bed017)

Note, however, that the extent of the *reparandum* (i.e. the words that are replaced by following words) is not part of the transcription.

3.2 Annotation of *It*

We performed an annotation with two external annotators. We chose annotators outside the project in order to exclude the possibility that our own preconceived ideas influence the classification. The purpose of the annotation was twofold: Primarily, we wanted to collect training and test data for our machine learning experiments. At the same time, however, we wanted to investigate how reliably this kind of annotation could be done. The annotators were asked to label instances of *it* in five ICSI Meeting Corpus dialogs[1] as belonging

[1] Bed017, Bmr001, Bns003, Bro004, and Bro005

to one of the classes **normal**, **vague**, **discarded**, **extrapos it**, **prop-it**, or **other**.[2] The idea behind using this five-fold classification (as opposed to a binary one) was that we wanted to be able to investigate the inter-annotator reliability for each of the sub-types individually (cf. below). The first two classes are sub-types of referential *it*: **Normal** applies to the normal, anaphoric use of *it*. **Vague** *it* (Eckert & Strube, 2000) is a form of *it* which is frequent in spoken language, but rare in written text. It covers instances of *it* which are indeed referential, but whose referent is not an identifiable linguistic string in the context of the pronoun. A frequent (but not the only) type of **vague** *it* is the one referring to the current discourse topic, like in the following example:

ME011: [...] [M]y vision of **it** is you know each of us will have our little P_D_A in front of us *Pause* and so the acoustics - uh you might want to try to match the acoustics. (Bmr001)

Note that we treat vague *it* as referential here even though, in the context of a coreference resolution preprocessing filter, it would make sense to treat it as nonreferential since it does not have an antecedent that it can be linked to. However, we follow Evans (2001) in assuming that the information that is required to classify an instance of *it* as a mention of the discourse topic is far beyond the local information that can reasonably be represented for an instance of *it*.

The classes **discarded**, **extrapos it** and **prop-it** are sub-types of nonreferential *it*. The first two types have already been shown in the example in Section 1. The class **prop-it**[3] was included to cover cases like the following:

FE004: So **it** seems like a lot of - some of the issues are the same. [...] (Bed017)

The annotators received instructions including descriptions and examples for all categories, and a decision tree diagram. The diagram told them e.g. to use *wh*-question formation as a test to distinguish **extrapos it** and **prop-it** on the one hand from **normal** and **vague** on the other. The criterion for distinguishing between the latter two phenomena was to use **normal** if an antecedent could be identified, and **vague** otherwise. For **normal** pronouns, the annotators were also asked to indicate the antecedent. The annotators were also told to tag as **extrapos it** only those cases in which an extraposed element (*to*-infinitive, *ing*-form or *that*-clause with or without complementizer) was available, and to use **prop-it** otherwise. The annotators individually performed the annotation of the five dialogs. The results of this initial annotation were analysed and problems and ambiguities in the annotation scheme were identified and corrected. The annotators then individually performed the actual annotation again. The results reported in the following are from this second annotation.

We then examined the inter-annotator reliability of the annotation by calculating the κ score (Carletta, 1996). The figures are given in Table 1. The category **other** contains all cases in which one of the minor categories was selected. Each table cell contains the percentage agreement and the κ value *for the respective category*. The final column contains the overall κ for the entire annotation.

The table clearly shows that the classification of *it* in spoken dialog appears to be by no means trivial: With one exception, κ for the category **normal** is below .67, the threshold which is normally regarded as allowing tentative conclusions (Krippendorff, 1980). The κ for the nonreferential sub-categories **extrapos it** and **prop-it** is also very variable, the figures for the former being on average slightly better than those for the latter, but still mostly below that threshold. In view of these results, it would be interesting to see similar annotation experiments on written texts. However, a study of the types of confusions that occur showed that quite a few of the disagreements arise from confusions of sub-categories belonging to the same super-category, i.e. referential resp. nonreferential. That means that a decision on the level of granularity that is needed for the current work can be done more reliably.

The data used in the machine learning experiments described in Section 4 is a *gold standard* variant that the annotators agreed upon after the annotation was complete. The distribution of the five classes in the *gold standard* data is as follows: **normal**: 588, **vague**: 48, **discarded**: 222, **extrapos it**: 71, and **prop-it**: 88.

[2] The actual tag set was larger, including categories like **idiom** which, however, the annotators turned out to use extremely rarely only. These values are therefore conflated in the category **other** in the following.

[3] Quirk et al. (1991)

	normal	vague	discarded	extrapos it	prop-it	other	κ
Bed017	81.8% / .65	36.4% / .33	94.7% / .94	30.8% / .27	63.8% / .54	44.4% / .42	.62
Bmr001	88.5% / .69	23.5% / .21	93.6% / .92	50.0% / .48	40.0% / .33	0.0% / -.01	.63
Bns003	81.9% / .59	22.2% / .18	80.5% / .75	58.8% / .55	27.6% / .21	33.3% / .32	.55
Bro004	84.0% / .65	0.0% / -.05	89.9% / .86	75.9% / .75	62.5% / .59	0.0% / -.01	.65
Bro005	78.6% / .57	0.0% / -.03	88.0% / .84	60.0% / .58	44.0% / .36	25.0% / .23	.58

Table 1: Classification of *it* by two annotators in a corpus subset.

4 Automatic Classification

4.1 Training and Test Data Generation

4.1.1 Segmentation

We extracted all instances of *it* and the segments (i.e. speaker units) they occurred in. This produced a total of 1.017 instances, 62.5% of which were referential. Each instance was labelled as *ref* or *nonref* accordingly. Since a single segment does not adequately reflect the context of the *it*, we used the segments' time information to join segments to larger units. We adopted the concept and definition of *spurt* (Shriberg et al., 2001), i.e. a sequence of speech not interrupted by any pause longer than $500ms$, and joined segments with time distances below this threshold. For each instance of *it*, features were generated mainly on the basis of this spurt.

4.1.2 Preprocessing

For each spurt, we performed the following preprocessing steps: First, we removed all single dashes (i.e. *interruption points*), non-lexicalised filled pauses (like *em* and *eh*), and all word fragments. This affected only the *string* representation of the spurt (used for pattern matching later), so the information that a certain spurt position was associated with e.g. an interruption point or a filled pause was not lost.

We then ran a simple algorithm to detect direct repetitions of 1 to up to 6 words, where removed tokens were skipped. If a repetition was found, each token in the *first* occurrence was tagged as *discarded*. Finally, we also temporarily removed potential discourse markers by matching each spurt against a short list of expressions like *actually*, *you know*, *I mean*, but also *so* and *sort of*. This was done rather agressively and without taking any context into account. The rationale for doing this was that while discourse markers do indeed convey important information to the discourse, they are not relevant for *the task at hand* and can thus be considered as noise that can be removed in order to make the (syntactic and lexical) patterns associated with nonreferential *it* stand out more clearly. For each spurt thus processed, POS tags were obtained automatically with the Stanford tagger (Toutanova et al., 2003). Although this tagger is trained on written text, we used it without any retraining.

4.1.3 Feature Generation

One question we had to address was which information from the transcription we wanted to use. One can assume that using information like sentence breaks or interruption points should be expected to help in the classification task at hand. On the other hand, we did not want our system to be *dependent* on this type of human-added information. Thus, we decided to do several setups which made use of this information to various degrees. Different setups differed with respect to the following options:

-use_eos_information: This option controls the effect of explicit end-of-sentence information in the transcribed data. If this option is active, this information is used in two ways: Spurt strings are trimmed in such a way that they do not cross sentence boundaries. Also, the search space for distance features is limited to the current sentence.

-use_interruption_points: This option controls the effect of explicit interruption points. If this option is active, this information is used in a similar way as sentence boundary information.

All of the features described in the following were obtained fully automatically. That means that errors in the shallow feature generation methods could propagate into the model that was learned from the data. The advantage of this approach is, however, that training and test data are *homogeneous*. A model trained on partly erroneous data is supposed to be more robust against similarly noisy testing data.

The first group of features consists of 21 surface syntactic patterns capturing the left and right context of *it*. Each pattern is represented by a binary feature which has either the value *match* or *nomatch*. This type of pattern matching is done

for two reasons: To get a simplified symbolic representation of the syntactic context of *it*, and to extract the other elements (nouns, verbs) from its predicative context. The patterns are matched using shallow (regular-expression based) methods only.

The second group of features contains lexical information about the predicative context of *it*. It includes the verb that *it* is the grammatical subject resp. object of (if any). Further features are the nouns that serve as the direct object (if *it* is subject), and the noun resp. adjective complement in cases where *it* appears in a copula construction. All these features are extracted from the patterns described above, and then lemmatized.

The third group of features captures the wider context of *it* through distance (in tokens) to words of certain grammatical categories, like next complementizer, next *it*, etc.

The fourth group of features contains the following: *oblique* is a binary feature encoding whether the *it* is preceeded by a preposition. *in_seemlist* is a feature that encodes whether or not the verb that *it* is the subject of appears in the list *seem, appear, look, mean, happen, sound* (from Dimitrov et al. (2002)). *discarded* is a binary feature that encodes whether the *it* has been tagged as discarded during preprocessing. The features are listed in Table 2. Features of the first group are only given as examples.

4.2 Machine Learning Experiment

We then applied machine learning in order to build an automatic classifier for detecting nonreferential instances of *it*, given a vector of features as described above. We used JRip, the WEKA[4] reimplementation of Ripper (Cohen, 1995). All following figures were obtained by means of ten-fold cross-validation. Table 3 contains all results discussed in what follows.

In a first experiment, we did not use either of the two options described above, so that no information about interruption points or sentence boundaries was available during training or testing. With this setting, the classifier achieved a recall of 55.1%, a precision of 71.9% and a resulting F-measure of 62.4% for the detection of the class *nonreferential*. The overall classification accuracy was 75.1%.

The advantage of using a machine learning system that produces human-readable models is that it allows direct introspection of which of the features were used, and to which effect. It turned out that the *discarded* feature is very successful. The model produced a rule that used this feature and correctly identified 83 instances of nonreferential *it*, while it produced no false positives. Similarly, the *seem_list* feature alone was able to correctly identify 22 instances, producing nine false positives. The following is an example of a more complex rule involving distance features, which is also very successful (37 true positives, 16 false positives):

```
dist_to_next_to <= 8 and
dist_to_next_adj <= 4
  ==> class = nonref (53.0/16.0)
```

This rule captures the common pattern for extraposition constructions like *It is important to do that.*

The following rule makes use of the feature encoding the distance to the next complementizer (14 true positives, five false positives):

```
obj_verb = null and
dist_to_next_comp <= 5
  ==> nonref (19.0/5.0)
```

The fact that these rules with these conditions were learned show that the features found to be most important for the detection of nonreferential *it* in written text (cf. Section 2) are also highly relevant for performing that task for spoken language.

We then ran a second experiment in which we used sentence boundary information to restrict the scope of both the pattern matching features and the distance-related features. We expected this to improve the performance of the model, as patterns should apply less generously (and thus more accurately), which could be expected to result in an increase in precision. However, the second experiment yielded a recall of 57.7%, a precision of only 70.1% and an F-measure of 63.3% for the detection of this class. The overall accuracy was 74.9%. The system produced a mere five rules (compared to seven before). The model produced the identical rule using the *discarded*-feature. The same applies to the *seem_list* feature, with the difference that both precision and recall of this rule were altered: The rule now produced 23 true positives and six false positives. The slightly higher recall of the model using the sentence boundary information is mainly due to a better coverage of the rule using the features encoding the distance to the next to-infinitive and the next adjective: it now produced

[4] http://www.cs.waikato.ac.nz/ ml/

	Syntactic Patterns	
1.	INF_it	*do it*
10.	it_BE_adj	*it was easy*
11.	it_BE_obj	*it's a simple question*
13.	it_MOD-VERBS_INF_obj	*it'll take some more time*
20.	it_VERBS_TO-INF	*it seems to be*
	Lexical Features	
22.	noun_comp	noun complement (in copula construction)
23.	adj_comp	adjective complement (in copula construction)
24.	subj_verb	verb that *it* is the subject of
25.	prep	preposition before indirect object
26.	ind_obj	indirect object of verb that *it* is subject of
27.	obj	direct object of verb that *it* is subject of
28.	obj_verb	verb that *it* is object of
	Distance Features (in tokens)	
29.	dist_to_next_adj	distance to next adjective
30.	dist_to_next_comp	distance to next complementizer (*that,if,whether*)
31.	dist_to_next_it	distance to next *it*
32.	dist_to_next_nominal	distance to next nominal
33.	dist_to_next_to	distance to next to-infinitive
34.	dist_to_previous_comp	distance to previous complementizer
35.	dist_to_previous_nominal	distance to previous nominal
	Other Features	
36.	oblique	whether *it* follows a preposition
37.	seem_list	whether subj_verb is *seem, appear, look, mean, happen, sound*
38.	discarded	whether *it* has been marked as discarded (i.e. in a repetition)

Table 2: Our Features (selection)

57 true positives and only 30 false positives.

We then wanted to compare the contribution of the sentence breaks to that of the interruption points. We ran another experiment, using only the latter and leaving everything else unaltered. This time, the overall performance of the classifier improved considerably: recall was 60.9%, precision 80.0%, F-measure 69.2%, and the overall accuracy was 79.6%. The resulting model was rather complicated, including seven complex rules. The increase in recall is mainly due to the following rule, which is not easily interpreted:[5]

```
it_s = match and
dist_to_next_nominal >=21 and
dist_to_next_adj >=500 and
subj_verb = null
  ==> nonref (116.0/31.0)
```

The considerable improvement (in particular in precision) brought about by the interruption points, and the comparatively small impact of sentence boundary information, might be explainable in several ways. For instance, although sentence boundary information allows to limit both the search space for distance features and the scope of pattern matching, due to the shallow nature of preprocessing, what is *between* two sentence breaks is by no means a well-formed sentence. In that respect, it seems plausible to assume that smaller units (as delimited by interruption points) may be beneficial for precision as they give rise to fewer spurious matches. It must also be noted that interruption points do not mark *arbitrary* breaks in the flow of speech, but that they can signal important information (cf. Heeman & Allen (1999)).

[5]The value 500 is used as a MAX_VALUE to indicate that no match was found.

5 Conclusion and Future Work

This paper presented a machine learning system for the automatic detection of nonreferential *it* in spoken dialog. Given the fact that our feature extraction methods are only very shallow, the results we obtained are satisfying. On the one hand, the good results that we obtained when utilizing information about interruption points (P:80.0% / R:60.9% / F:69.2%) show the feasibility of detecting nonreferential *it* in spoken multi-party dialog. To our knowledge, this task has not been tackled before. On the other hand, the still fairly good results obtained by only using automatically determined features (P:71.9% / R:55.1% / F:62.4%) show that a practically usable filtering component for nonreferential *it* can be created even with rather simple means.

All experiments yielded classifiers that are *conservative* in the sense that their precision is considerably higher than their recall. This makes them particularly well-suited as *filter* components.

For the coreference resolution system that this

	P	R	F	% Correct
None	71.9 %	55.1 %	62.4 %	75.1 %
Sentence Breaks	70.1 %	57.7 %	63.3 %	74.9 %
Interruption Points	80.0 %	60.9 %	69.2 %	79.6 %
Both	74.2 %	60.4 %	66.6 %	77.3 %

Table 3: Results of Automatic Classification Using Various Information Sources

work is part of, only the fully automatic variant is an option. Therefore, future work must try to improve its recall without harming its precision (too much). One way to do that could be to improve the recognition (i.e. correct POS tagging) of grammatical function words (in particular complementizers like *that*) which have been shown to be important indicators for constructions with nonreferential *it*. Other points of future work include the refinement of the syntactic pattern features and the lexical features. E.g., the values (i.e. mostly nouns, verbs, and adjectives) of the lexical features, which have been almost entirely ignored by both classifiers, could be generalized by mapping them to common WordNet superclasses.

Acknowledgements

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) in the context of the project DIANA-Summ (STR 545/2-1), and by the Klaus Tschira Foundation (KTF), Heidelberg, Germany. We thank our annotators Irina Schenk and Violeta Sabutyte, and the three anonymous reviewers for their helpful comments.

References

Boyd, A., W. Gegg-Harrison & D. Byron (2005). Identifying non-referential *it*: a machine learning approach incorporating linguistically motivated patterns. In *Proceedings of the ACL Workshop on Feature Selection for Machine Learning in NLP, Ann Arbor, MI, June 2005*, pp. 40–47.

Byron, D. K. (2002). Resolving pronominal reference to abstract entities. In *Proc. of ACL-02*, pp. 80–87.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Clemente, J. C., K. Torisawa & K. Satou (2004). Improving the identification of non-anaphoric *it* using Support Vector Machines. In *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications,* Geneva, Switzerland.

Cohen, W. W. (1995). Fast effective rule induction. In *Proc. of the 12th International Conference on Machine Learning*, pp. 115–123.

Dimitrov, M., K. Bontcheva, H. Cunningham & D. Maynard (2002). A light-weight approach to coreference resolution for named entities in text. In *Proc. DAARC2*.

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

Evans, R. (2001). Applying machine learning toward an automatic classification of *It*. *Literary and Linguistic Computing*, 16(1):45 – 57.

Heeman, P. & J. Allen (1999). Speech repairs, intonational phrases, and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):527–571.

Janin, A., D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke & C. Wooters (2003). The ICSI Meeting Corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Hong Kong, pp. 364–367.

Krippendorff, K. (1980). *Content Analysis: An introduction to its methodology.* Beverly Hills, CA: Sage Publications.

Lappin, S. & H. J. Leass (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Ng, V. & C. Cardie (2002). Improving machine learning approaches to coreference resolution. In *Proc. of ACL-02*, pp. 104–111.

Paice, C. D. & G. D. Husk (1987). Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun 'it'. *Computer Speech and Language*, 2:109–132.

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik (1991). *A Comprehensive Grammar of the English Language.* London, UK: Longman.

Shriberg, E., A. Stolcke & D. Baron (2001). Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH '01),* Aalborg, Denmark, 3–7 September 2001, Vol. 2, pp. 1359–1362.

Strube, M. & C. Müller (2003). A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* Sapporo, Japan, 7–12 July 2003, pp. 168–175.

Toutanova, K., D. Klein & C. D. Manning (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 03*, pp. 252–259.