Applying Machine Learning Toward an Automatic Classification of It

Article *in* Literary and Linguistic Computing · April 2001

DOI: 10.1093/llc/16.1.45 · Source: OAI

1 author:

Richard Evans
University of Wolverhampton
48 PUBLICATIONS **844** CITATIONS

Applying Machine Learning Toward an Automatic Classification of *It*

Richard Evans,

School of Humanities, Languages and Social Sciences,

University of Wolverhampton,

Stafford Street,

Wolverhampton,

WV1 1SB.

UK.

Abstract

In the majority of cases, the pronoun *it* illustrates nominal anaphora, tending to refer back to another noun phrase in the text. However, in a significant minority of cases, the pronoun is used in exceptional ways that fail to demonstrate strict nominal anaphora. The identification of these uses of *it* is important in all fields where pronoun resolution has an impact. Following a survey of previous treatments of the pronoun *it* in the literature, some features of instances of *it* are proposed that can be used in a novel memory-based learning method to automatically classify those instances. On evaluating the method, it is found that the implemented system performs comparably well with respect to a rule-based system and with an extended training set it is expected that the accuracy of the system will improve, offering greater coverage than rule-based methods.

Applying Machine Learning Toward an Automatic Classification of *It*

1. Introduction

Surveying the uses of *it* in English, both as described in the literature (see Section 2.1) and by examination of natural language data from the SUSANNE (Sampson, 1995) and BNC (Burnard, 1995) corpora, allows one to identify seven different profiles for this pronoun. They are spelt out with examples below:

1. Nominal anaphoric, such as 'Do not sweep *the dust*$_i$ when dry, you will only recirculate *it*$_i$,' in which the pronoun takes its reference from another nominal expression in the text.

2. Clause anaphoric, such as '*One day in 1970, fifty thousand women marched down fifth Avenue in New York*$_i$. *It*$_i$ is said to have been the biggest women's gathering since suffrage days,' in which the pronoun can be interpreted with reference to a preceding clause in the text.

3. Proaction, such as 'Mays *walloped four home runs in a span of nine innings*$_i$. Incidentally, only two did *it*$_i$ before a home audience,' in which *it* combines with *do* to form a unit that takes its interpretation from a preceding verb phrase in the text.

4. Cataphoric, such as 'when *it*$_i$ fell, *the glass*$_i$ broke,' in which the pronoun is coreferential with a following nominal expression in the text. In light of discourse theories such as Veins Theory (Cristea et al., 1998), there is some controversy over the notion of 'cataphora' (Dan Cristea, personal communication) and perhaps the term 'initial mention *it*' should be used instead. It is clear that resolution of cataphoric expressions cannot be accomplished on the instant that they present themselves to a language user. However, it is also clear that the resolution of such

expressions by machine increases the amount of useful information available from a text. For instance, from the illustrative sentence above, it is informative to derive both the propositions that *the glass*$_i$ *fell* and *the glass*$_i$ *broke*. In his thesis, (Tanaka, 1999), argues convincingly in favour of the existence of cataphora.

5. Discourse topic, such as 'Always use a tool for the job it was designed to do. Always use tools correctly. If *it* feels very awkward, stop,' in which an interpretation for the pronoun must be derived by some non-trivial interpretation of the subject matter of the discourse.

6. Pleonastic, such as '*It* is worth having more than one size or a good-quality set with interchangeable bits,' in which there is no interpretation for the pronoun. It is non-referential, used due to certain requirements of the grammar.

7. Idiomatic / Strereotypic, such as 'I take *it* you're going now,' where again, the pronoun is non-referential, but used in certain fixed expressions in the language.

Cases 1 - 3 can be regarded as anaphoric, which is to say that the interpretation for the pronoun is explicit and may be obtained by backward search through the text. Case 4 suggests a forward search strategy. No explicit interpretation is available for the remaining cases.

In this paper, Section 2 surveys previous work in the description and recognition of some of the seven uses of *it*. The classification of *it* with respect to these seven uses is motivated in Section 3. In Section 4, a machine learning method for classifying *it* is proposed and similar work on the application of machine learning to a different task in NLP is also reviewed. In Section 5, the construction of a training file for use in applying and evaluating the machine learning approach is described. Section 6 presents evaluation and comparison of the machine learning classification method

with respect to a rule-based method. Finally, in Section 7, findings are discussed and ideas for future work presented.

2. Previous Work

2.1 A Grammatical Description of the Phenomenon

Use of the pronoun *it* receives coverage in most serious surveys of English grammar, including (Quirk et al., 1985), (Sinclair et al., 1995) and (Swan, 1995). In addition to its more common anaphoric profile, (Sinclair et al., 1995) note that *it* is used in describing places and situations as in 1, commenting about time and the weather, as in 2, referring to whole situations or facts, as in 3, making requests or passing on instructions, as in 4, and commenting on actions, activities or experiences, as in 5, etc.

1. *It*'s lonely here.
2. *It* had been raining all day.
3. *It* was agreed that the transaction be kept secret.
4. Is *it* okay if we sit here?
5. Gretchen found *it* difficult to speak.

(Quirk et al., 1985) also present a number of idiomatic and miscellaneous uses of the pronoun. With respect to previous work, the reader notes the diverse terminology used in order to describe non-nominal and non-anaphoric *it*. For example, it has been referred to as *pleonastic*, *expletive*, *structural*, *dummy*, and *non-anaphoric*, to name just a few. (Morgan, 1968) explores implications for transformational accounts of English grammar, especially with respect to D-structure, of the occurrence of pleonastic *it*.

2.2 Automatic Identification of Pleonastic It

Automatic identification of pleonastic *it* is a challenge that has been addressed in a number of papers on anaphora resolution (Lappin and Leass, 1994) and (Denber, 1998). Researchers such as (Baldwin, 1997), (Kennedy and Boguraev, 1996), and (Marilyn Walker, personal communication to Ruslan Mitkov) also claim to have tackled the issue, but at the present time details of those implementations are sparse. The methods reported in the former papers are based on simple pattern recognition. To illustrate, one pattern set out in (Lappin and Leass, 1994) is stated as:

It + (any form of the verb 'to *be*') + (a cognitive verb (past tense)) + *that*

This pattern would identify pleonastic *it* in constructs like '*it* is thought *that* …'. However, if one assumes that a strict, uninterrupted sequence must match, it is clear that the example above would fail to identify a construct such as '*it* was never *thought* by the committee *that* …' and its coverage is seen to be relatively limited. If a more general application of the pattern were allowed, (Lappin and Leass, 1998) do not suggest any general constraints on the intervening material that may be permitted to lie between the triggering elements of the patterns. An additional problem is that application of such patterns relies on explicit knowledge of elements such as 'cognitive verbs' or 'weather adjectives' or 'time expressions'. It is not clear, given the creativity of language users, that lists of this kind would ever be complete, though the availability of ontologies such as WordNet (Fellbaum, 1998) would help to increase the generality of such lists. In any case, the necessary triggering elements are often complex constituents that are difficult to identify automatically, as in 'it was the last hogmanay of the second millennium and raining as usual when …'.

Neither of the methods described in the work of (Lappin and Leass, 1994) or (Denber, 1998) was evaluated. It is suspected, in the light of work by (Paice and Husk,

1987) that applying the patterns in a less constrained manner in order to extend their coverage, would entail a significant number of false positive identifications.

That problem was avoided in the system proposed by (Paice and Husk, 1987). There, a number of patterns were proposed, based on data from the LOB corpus and prior grammatical descriptions of the use of *it*. The approach differed from those of (Lappin and Leass, 1994) and (Denber, 1998) because constraints were applied during the pattern matching process. To illustrate, one pattern identified *it* as being non-referential if it occurred in the sequence '***it*** *…that*.' This rule is prevented from over-applying by setting some constraints on the material permitted to lie between ***it*** and *that*. For example, no more than 25 words may lie between them and there are limits on the appearance of punctuation symbols. Another constraint was that *structural* (here referred to as pleonastic) uses of *it* are never immediately preceded by a 'prepositional' word. Appendix 3 of (Paice and Husk, 1987) lists the set of thirty prepositional words considered. It can be viewed as a subset of English prepositions that includes items like *beside*, *to*, and *upon* but excludes ambiguous items like *for*, and *so*. The method posits patterns that are general enough to minimise the system's reliance on large lists of trigger words, although some small lists are still used.

(Paice and Husk, 1987) reported 93.9% accuracy for the binary classification of *it* as structural or not. This level is especially impressive, given that the system did not incorporate the use of a part of speech tagger in order to make the detection of verbs and other items of interest more accurate. The system was implemented for comparison with the machine learning approach presented here, and was found to perform less accurately on the data used in this evaluation. More information can be found in Section 6.

In general, it should be acknowledged that it was the work surveyed in this section, rather than that in Section 2.1 that had the greater impact on the method proposed in Section 4. In addition, observations of patterns found in the corpus used for training in this method, supported by the observations of (Paice and Husk, 1987) in their test data, allowed the formulation of some useful features for recognition of pleonastic *it*. It was felt that reliance on surveys such as (Quirk et al., 1987) would lead to the formulation of a large number of triggers, each providing coverage of a small number of infrequent patterns. Unless great care is taken, implementing and applying a large number of triggers carries the significant danger of over-application and poor coverage.

3. Motivation

Coreference resolution has been found to be crucial in the fields of information extraction (Chinchor and Hirschman, 1997), machine translation (Peral et al., 1999), and automatic summarization (Harabagiu and Maiorano, 1999). The resolution of pronouns to nominal expressions constitutes an important component in that process. Disambiguation of the profile of an occurrence of *it* will increase the accuracy of all systems undertaking pronoun resolution. It is easy to infer that if a non-nominal usage of *it* is resolved to a noun phrase, then the mention counts of entities in texts or particular sentences will be skewed - affecting automatic information retrieval and summarization tasks - and automatic translations will be adversely affected.

The scale of the problem is quite large. (Lappin and Leass, 1994) noted that 8% of all the pronouns occurring in the texts upon which they tested their RAP system were pleonastic. In the present work, it was found that almost one third of the uses of *it* in the training set were not examples of nominal anaphora. In evaluating anaphora

resolution systems, pronouns failing to exhibit nominal anaphora are usually excluded from the test data. (Ge et al., 1998) removed such expressions by hand. The point is that for full automation of an anaphora resolution system, it must incorporate a recognition component for pronouns that are not anaphoric to nominal expressions, (Orasan et al., 2000).

As will have been noted from Section 2.2, previous work has only addressed identification of pleonastic *it*. As will be seen in section 5, pleonastic uses account for 83% of the total range of non-nominal or non-anaphoric uses in the data used here, leaving 17% unidentified. The system presented in this paper is intended to identify *all* such uses of *it*, not just the pleonastic subset.

The motivation for extending the coverage of a system in this way derives from the fact that although most current work on pronoun resolution has been concerned with finding noun phrase antecedents for pronouns, it is envisaged that this focus will be relaxed to consider resolution to other types of antecedent, as in the thesis proposal of (Byron, 1999). This would allow consideration of pronominal reference to events and propositions, a concern in information extraction (Chinchor, 1997). It is possible to foresee that the automatic classification of *it* and other pronouns would trigger different resolution strategies such as forward search for cataphoric pronouns and the consideration of verbal and clausal antecedents in other cases.

4. A Machine Learning Approach

4.1 Machine Learning Applied to a Similar Task

Machine learning methods were used by (Litman, 1996) in order to make a classification of cue phrases that contribute either discourse structural or semantic information to texts. In that paper, the author used the machine learning methods C

4.5 and CGRENDEL in order to derive classification procedures from human annotated training data. She found that the decision tree classification method obtained by C 4.5 and the sets of 'if-then' rules obtained by CGRENDEL both outperformed human-derived sets of classification rules. As in this paper, Litman used ten-fold cross-validation over the human annotated training data to evaluate and compare the methods.

4.2 Memory Based Learning

The machine learning method proposed in this work was executed using Tilburg University's Memory Based Learner (TiMBL), (Daelemans, 1999). For learning, the k-nearest neighbour method was used, which is a simple memory-based learning algorithm, available in the TiMBL package. That method uses a pre-classified training set to classify new instances. Each instance in the training set is represented by a vector of feature values that has been explicitly classified. When a new vector of feature values is presented to TiMBL, a distance metric is computed between the new vector and the set of vectors held in the training set. The k nearest vectors are determined using the metric. The new vector is then classified based on the most frequent classification of the k nearest neighbours. In the current system, for optimal results, 15 nearest neighbours were considered and gain ratio, (Quinlan, 1993), was used as the distance metric.

4.3 Features Used for Classification of ***It***

In order to classify instances of *it* according to the types presented in Section 1, 35 features were proposed and a method for obtaining the values of those features for instances of *it* was implemented.

In the first step, the text containing instances to be classified is analysed using Conexor's FDG-Parser (Tapanainen and Jarvinen, 1997). This program returns information on the part of speech and morphological lemma of words in a text, as well as returning the dependency relations between those words. The dependency information allows the identification of complex constituents in a text. For example, complex noun phrases can be identified by transitively grouping together all the words dependent on a noun head. Additional software was implemented in order to perform this. It is supposed that the return of word lemmas by the parser allows a greater degree of generality in the training instances.

In the current approach, 35 features are used. For the purpose of description, it is suitable to regard them as broadly belonging to one of six different classes, detailed below.

1. Positional information describing the position of the instance in terms of word position in a sentence and sentence position in a paragraph. It is expected that non-anaphoric pronouns may appear in initial positions because they require no preceding antecedents.

2. Features describing the number of elements suggestive of the pronoun's class in the surrounding text. For example, pleonastic pronouns rarely appear immediately after a prepositional word, as noted by (Paice and Husk, 1987) and complementisers or adjectives often follow pleonastic instances. In anaphora resolution, researchers such as (Mitkov, 1998) have noted that an anaphoric pronoun's antecedent is often present in the same paragraph as the pronoun. It is therefore expected that nominal anaphors will follow prior noun phrases more frequently on average than is the case for non-nominal anaphors and non-anaphoric pronouns.

3. Lemmas of preceding material such as verbs and following material such as verbs or adjectives in the same sentence as the instance. It can be seen that incorporating the lemmas of such elements into the feature value vectors, and associating this information with instances, reduces the system's requirement for external lists of trigger elements such as 'weather adjectives' or 'cognitive verbs'.

4. The parts of speech of eight tokens, four words prior to and four words after the instance.

5. Pleonastic uses of *it* are noted to be associated with certain sequences of elements following the instance in the same sentence as the pronoun. The sequences used presently are 'adjective + noun phrase,' as in constructions like 'It was *obvious the book* would fall,' and 'complementiser + noun phrase' as in constructions like 'It was obvious *that the book* would fall.'

6. Proximity of following elements such as complementisers, *-ing* forms of verbs, and prepositions, expressed in tokens.

The entire set of features can be regarded as a synthesis of information derived from corpus data and noted by the researchers whose surveys are reviewed in Section 2 to be useful for identifying non-nominal and non-anaphoric uses of *it*.

5. Building A Training File

The machine learning method used for classification of *it* requires training data. Here, the training data is derived from 77 texts taken from the SUSANNE and BNC corpora. To ensure broad coverage, those texts belonged to numerous genres including politics, science, fiction, and journalism. The corpus contains 368830 words with 3171 examples of *it*. Of these examples, 67.93% were nominal anaphoric, 0.82% were clause anaphoric, 0.06% were proaction uses, 0.09% were cataphoric, 2.08% were

discourse topic mentions, 26.77% were pleonastic, and 2.24% were used in idiomatic/stereotypic constructions. It is notable here that almost one third of the uses of *it* are not examples of nominal anaphora.

A program was written in order to extract all occurrences of *it* from the corpus and to assign to each the vector of feature values described in Section 4. The human annotator is then presented with the paragraph in which the instance appears and is prompted to classify the instance. The vectors, together with their manual classification are written to the training file.

6. Evaluation and Comparison

The memory based learning method was tested using ten-fold cross-validation over the training corpus described in Section 5. Its performance is shown in Table 1. In order to give some context to the performance results reported here, a system was also implemented to execute the method proposed in (Paice and Husk, 1987). A more detailed description of that implementation is detailed in (Evans, 2000). Below, it is referred to as 'Rule Based.' That method was tested on the data used to develop the memory-based learning approach. Direct comparison is slightly difficult, given that the two methods are intended to perform slightly different tasks. In the case of the method proposed by (Paice and Husk, 1987), the goal is to identify pleonastic uses of *it*, whereas the machine learning system is intended to make a classification of the pronoun into one of seven classes.

When evaluating a system designed to make multiple classifications, it is not possible to use the measures of precision and recall to provide an overview of the system's performance. The statistics: true positives; false positives; and false negatives can only be derived from a particular standpoint such as the 'classification

of cataphoric pronouns,' in which assignment of a cataphoric pronoun to the cataphoric class is a true positive, assignment to another class is a false negative and assignment of non-cataphoric pronouns to the cataphoric class are false positives. Since the traditional measures of precision and recall are constructed in terms of these statistics, as presented in (Manning and Schuetze, 1999), those measures will not be applicable to determining the efficacy of the current classification system. For this task, it is necessary to use more general measures and here, *7-ary classification accuracy* is proposed. It is defined as the ratio of the number of pronouns assigned to the correct class, called *true 7-ary classifications*, to the total number of pronouns classified.

It may be the case that the user is concerned only with the separation of NP anaphoric pronouns from non-NP anaphoric pronouns, a crucial task in traditional pronominal anaphora resolution. Performance on this task can be evaluated using a measure that will be called *binary classification accuracy*. It will be defined as the ratio of *true binary classifications* to the total number of pronouns classified. *True binary classifications* is defined as the sum of the number of non-NP anaphoric pronouns assigned to *any* of the six non-NP anaphoric classes and the number of NP anaphoric pronouns assigned to the NP anaphoric class. Thus the classification model has been reduced from seven classes to two. A cataphoric pronoun assigned to the pleonastic class would be a *true binary classification* but a NP anaphoric pronoun assigned to the proaction class would not be. Below, *binary classification accuracy* is expressed as a percentage.

Table 1 Performance of the machine learning and rule-based classification methods

	Machine Learning	Rule-Based

#True 7-ary classifications	2261	2233
#True binary classifications	2333	2290
7-ary classification accuracy %	69.27	68.41
Binary classification accuracy %	71.48	70.16

The 'rule-based' system is intended only to make a binary classification of instances. The classification of *it* as either NP anaphoric or pleonastic will therefore be considered, enabling a more suitable comparison between the two systems. For each of these individual classification tasks, it is possible to measure system performance in terms of precision and recall. For correctly classifying instances as examples of nominal anaphora, the machine learning system has 67.94% precision and 89.14% recall. The rule-based system has 66.47% precision and 89.19% recall. For the task of classifying *it* as pleonastic, the measures are 73.38% precision and 69.25% recall for the machine learning method and 72.68% precision and 66.03% recall for the rule-based one. The difference in performance between the systems is thus marginal, but it can be said that the machine learning system compares favourably with the rule-based one.

The general efficacy of the machine learning method was assessed by deriving precision and recall in terms of each of the seven classes proposed in Section 1. These measures appear in Table 2.

Table 2 Measures assessing the general efficacy of the machine learning method

Pronoun Usage	Precision %	Recall %
Nominal anaphoric	67.94	89.14
Clause anaphoric	0	0
Proaction	0	0
Cataphoric	0	0

Discourse topic	25	1.51
Pleonastic	73.38	69.25
Idiomatic/stereotypic	33.33	0.7

Broadly speaking, there are two factors explaining the very poor performance of the machine learning method in classifying all but the pleonastic and NP anaphoric uses of *it*.

1. The features assigned to instances in the training sets are most appropriate for classification of pleonastic instances. Given that the vectors convey information obtained from the paragraph in which the instance appears, it is not reasonable to suppose that such local information will be sufficient to identify pronominal mentions of the discourse topic, for example. In respect of clause referential uses, at the present time, no clause identification programs have been incorporated into the system. For this reason, there is no way to express indicative information such as the presence and location of suitable clause 'antecedents' in the text using features.

2. The training data is insufficient. Clausal, discourse topic and cataphoric mentions are used infrequently. Only 45 proaction mentions and 10 cataphoric mentions appear in all the training data used in this method. Such a small number of instances require that those uses are consistently set in highly regular expressions if they are to be identified using memory based learning. Even so, in the case of cataphoric uses, ten-fold cross-validation may leave just nine instances in each training set. Given that classification is being made on the basis of 15 nearest neighbours, a correct classification from TiMBL is still rather unlikely.

7. Discussion and Future Work

In this paper an automatic classification system for the pronoun *it* has been proposed. The system itself is based on memory-based learning and was found to compare well with a rule-based classification system. It is presently being used as a component of the anaphora resolution system, MARS (Orasan et al., 2000), where it has been found to make a useful contribution. The system can be criticised on numerous grounds and was found to be totally ineffective in the identification of clause anaphoric, proaction and cataphoric uses of *it*. Future research is envisaged to proceed as follows.

It is proposed that identification of clause anaphoric uses will require the implementation of new features and a system capable of identifying clauses. In the absence of specialized software for clause identification, it may be possible to derive some suitable heuristics on the basis of dependency information between noun heads and verbs, information made accessible by the parsing software.

In the case of proaction uses, it is believed that the system's failure to make the relevant classification arises from sparse data. The task should be straightforward, given that in most cases, proaction uses involve the appearance of *it* as object of the verb 'to *do*' in a regular and consistent way.

Accurate identification of cataphoric uses may require the formulation of additional features, but these must follow an examination of many more cases occurring in natural language. (Tanaka, 1999) suggests that the use of cataphoric pronouns tends to be identified with a relatively small number of patterns, the major problem lying in the complexity of those patterns and the identification of particular multi-word units.

In general, improving the performance of the system relies on increasing the size of the training data by a significant factor. The addition of new features to account for cataphoric or clause anaphoric uses of *it* will count for nothing unless the number of training instances illustrating these phenomena is increased. Of course, memory-based

learning is vulnerable to the possibility of learning exceptions from huge data sets, but that problem seems very distant for the current application.

Another area of concern for us is the validity of the training data. At the present time, that data was manually annotated by one person. This state of affairs is not regarded as acceptable and so it is proposed that the data should be re-classified by at least one other person and the classifications checked against the original. Disagreements should then be resolved by inter-annotator discussion and research. It is suspected that in spite of the definitions of terms used to describe the uses of *it*, inter-annotator disagreement may be high, due to the number of ambiguous cases occurring in the corpus used for training. For example, a sentence such as 'The man who normally tends the garden on Wednesdays decided *it* should be done on Tuesdays' seems to be a proaction use, but in fact no antecedent is explicit in the sentence. 'The man who normally tends the garden on Wednesday decided *tends the garden* should be done on Tuesdays' does not seem correct. For this reason, an instance like this was manually classified as a discourse topic usage because some processing and derivation from the text is required to produce a suitable argument for the verb 'to *do*'. Due to problems for annotators such as these, it is proposed that a slightly modified classification model is required. However, the addition of many more fine-grained classes would not be particularly useful, as that would exacerbate the problem of sparse training data.

The current system seems to provide a promising basis for future work. The expansion of the training file should offset many of the problems found so far. At the very least, more corpus data would allow the formulation of new, more effective classification features. Of course, *it* is not the only pronoun that demonstrates non-nominal anaphora. It is believed that the current system could be used as the basis of a

system to classify other pronouns. However, it is proposed that each different pronoun will require a different set of classification features and the notion of combining all features and all pronominal instances into a single training file should be treated with caution in order to avoid potentially inconsistent or conflicting training data.

Just as *it* is not the only pronoun, English is not the only language. It will be a matter of great interest to combine a survey of the grammatical characteristics of pronouns in other languages with the application of parallel features and ideas in order to obtain a machine learning classification system for those pronouns.

8. References

Baldwin, B. (1997) *CogNIAC: high precision coreference with limited knowledge and linguistic resources, Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, pp. 38-45, Madrid, 1997.

Burnard, L. (1995) *Users Reference Guide British National Corpus Version 1.0,* Oxford University Computing Services, UK.

Byron, D.K. (1999) *Resolving Pronominal Reference to Abstract Entities: Thesis Project Proposal.* URCS Technical Report #714, Rochester.

Chinchor, N. (1997) *MUC-7 Information Extraction Task Definition,* http://www.muc.saic.com/proceedings/ie_task.pdf.

Chinchor, N. and Hirschman, L. (1997) *MUC-7 Coreference Task Definition,* http://www.muc.saic.com/proceedings/co_task.pdf.

Cristea, D., Ide, N. and Romary, L. (1998) *Veins Theory – A Model of Global Discourse Cohesion and Coherence, Proceedings of Coling/ACL,* Montreal, 1998.

Daelemans, W. (1999) *TiMBL: Tilburg University Memory Based Learner version 2 Reference Guide,* ILK Technical Report – ILK 99-01, Tilburg University, The Netherlands.

Denber, M. (1998) *Automatic Resolution of Anaphora in English*, Imaging Science Division, Eastman Kodak Co.

Evans, R. (2000) *A Comparison of Rule-Based and Machine Learning Methods for Identifying Non-nominal* It, *Proceedings of Natural Language Processing – NLP 2000*, pp. 233-241, Patras, 2000.

Fellbaum, C. (Ed.) (1998) *WordNet An Electronic Lexical Database*, MIT Press, Massachusetts, US.

Ge, N., Hale, J. and Charniak, E. (1998) *A statistical approach to anaphora resolution, Proceedings of the Sixth Workshop on Very Large Corpora*, Coling-ACL, Montreal, 1998.

Harabagiu, S. M. and Maiorano, S. J. (1999) *Knowledge-Lean Coreference Resolution and its Relation to Textual Cohesion and Coherence, Proceedings of the Workshop The Relation of Discourse / Dialogue Structure and Reference,* ACL, Maryland, 1999.

Kennedy, C. and Boguraev, B. (1996) *Anaphora for everyone: pronominal anaphora resolution without a parser, Proceedings of the 16th International Conference on Computational Linguistics (COLING'96),* pp. 113-118. Copenhagen, 1996.

Lappin, S. and Leass, H. J. (1994) *An Algorithm for Pronominal Anaphora Resolution, Computational Linguistics*, 20.4.

Litman, D. J. (1996) *Cue Phrase Classification Using Machine Learning, Journal of Artificial Intelligence Research*, 4, pp. 53-94.

Manning, C. D. and Schuetze, H. (1999) *Foundations of Statistical Natural Language Processing*, MIT Press, Massachusetts, US.

Mitkov, R. (1998) *Robust pronoun resolution with limited knowledge, Proceeding of the 18th International Conference on Computational Linguistics (Coling/ACL),* pp. 869-875, Montreal, 1998.

Morgan, J. L. (1968) *Some Strange Aspects of It, Papers from the Fourth Regional Meeting. Chicago Linguistic Society*, Chicago, 1968.

Orasan, C., Evans, R. and Mitkov, R. (2000) *Enhancing Preference-Based Anaphora Resolution with Genetic Algorithms, Proceedings of Natural Language Processing – NLP 2000*, pp. 185-195, Patras, 2000.

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

Peral, J., Saiz-Noeda, M., Ferrandez, A. and Palomar, M. (1999) *Anaphora Resolution and generation in a multilingual system. An Interlingua mechanism, Proceedings of VEXTAL*, Venice, 1999.

Quinlan, J. R. (1993) *C 4.5: Programs for Machine Learning*, Morgan Kaufmann, US.

Quirk, R., et al. (1985) *A Compehensive Grammar of the English Language*, Longman, UK.

Sampson, G. (1995) *English for the Computer: The SUSANNE Corpus and analytic scheme*, Oxford University Press, Oxford.

Sinclair, J. et al. (1995) *English Grammar*, Harper Collins Publishers, UK.

Swan, M. (1995) *Practical English Usage*, Oxford University Press, Oxford.

Tanaka, I. (1999) *The value of an annotated corpus in the investigation of anaphoric pronouns, with particular reference to backwards anaphora in English, PhD. Thesis*, Lancaster University, Lancaster.

Tapanainen, P. and Jarvinen, T. (1997) A Non-Projective Dependency Parser, Proceedings of the 5th Annual Conference of Applied Natural Language Processing, pp. 64-71, Washington, D.C., 1997.

	Machine Learning	Rule-Based
#True 7-ary classifications	2261	2233
#True binary classifications	2333	2290
7-ary classification accuracy %	69.27	68.41
Binary classification accuracy %	71.48	70.16

Pronoun Usage	Precision %	Recall %
Nominal anaphoric	67.94	89.14
Clause anaphoric	0	0
Proaction	0	0
Cataphoric	0	0
Discourse topic	25	1.51
Pleonastic	73.38	69.25
Idiomatic/stereotypic	33.33	0.7

Table Captions

Table 1 Performance of the machine learning and rule-based classification methods

Table 2 Measures assessing the general efficacy of the machine learning method