# Improving the Identification of Non-Anaphoric *it* using Support Vector Machines

**José Carlos Clemente Litrán**[*]
clemente@jaist.ac.jp

**Kenji Satou**[*]
ken@jaist.ac.jp

**Kentaro Torisawa**[†]
torisawa@jaist.ac.jp

[*]Graduate School of Knowledge Science
[†]Graduate School of Information Science
Japan Advanced Institute of Science and Technology (JAIST)
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan

## Abstract

Identification of non-anaphoric use of the pronoun *it* is crucial to achieve full anaphora resolution. Nevertheless, this problem has been either ignored or considered too simple to deserve a deeper study. In this paper we present a machine-learning approach using Support Vector Machines. We collected several instances of both anaphoric and non-anaphoric *it* from the GENIA corpus, together with syntactic information about the context. We show how by using a limited amount of knowledge our approach can achieve better accuracy than previous methods. We also analyze the relevance of features used to predict non-anaphoric uses.

## 1 Introduction

Recent years have seen an increasing interest in the process of anaphora resolution. The first step in automatic anaphora resolution is to identify the candidate anaphors whose antecedents we should find out. The identification of anaphoric words is, at least in English, not a trivial task. In pronoun resolution, specifically in the case of the impersonal pronoun *it*, to distinguish anaphoric and non-anaphoric instances has proven to be a very challenging problem. Consider the following examples:

(i) **It** still remains to be shown which molecular events lead to...

(ii) Studies of the ring chromosome that has XIST DNA but does not transcribe **it** show that its AR allele is..

Example (i) is non-anaphoric, while example (ii) is anaphoric. Nevertheless, most anaphora identification systems would incorrectly predict the first example as anaphoric and the second one as non-anaphoric! In example (i), the pronoun *it* in the idiomatic construction *It still remains* does not point to any previously mentioned entity in the discourse. The fact that this construction in which the pronoun appears is often associated with anaphoric uses (*The protein does not suffer major changes after this process, and it remains stable...*), can prompt this misclassification. In example (ii), on the other hand, the pronoun *it* is followed by the verb *show* and the complementiser *that*, which commonly identify non-anaphoric uses (*It has been shown that...*).

The importance of accurately classifying anaphoric and non-anaphoric cases is clear since the resolution of antecedents for a given anaphora relies on the assumption that the anaphora does have an antecedent. Filtering out non-anaphoric instances is therefore basic in order to reduce sources of error and improve overall accuracy of anaphora resolution systems.

Also, it should be noticed that most approaches on anaphora resolution (and, by extension, on identification of non-anaphoric instances) are tested against corpora like Penn Treebank, LOB or SUSANNE, which are mainly composed of newswire stories and journal articles. In this paper, we used the GENIA corpus (Ohta et al., 2002), which is entirely constructed of scientific abstracts. The percentage of non-anaphoric uses in this corpus turned out to be bigger than for previous studies, therefore conceding more importance to the correct identification of such cases.

The first approaches to this problem were based on basic pattern-matching techniques, as in Paice and Husk (1987), Lappin and Leass (1994) and Denber (1998). Even recent work on anaphora resolution still relies on this kind of approach to identify non-anaphoric uses (Castaño et al., 2002). As it will be seen in section 2.2, these pattern-based methods present certain drawbacks. In order to overcome them, Evans (2001) presented a machine-learning approach which performed slightly better than the best of the pattern-based ones.

In this paper, we present an approach based on support vector machines (SVM) (Cortes and Vapnik, 1995), a machine-learning technique that has shown very solid results in many areas of NLP. The data set we use in our experiments is constructed from the GENIA corpus, a collection of biological and medical documents extracted from the National Library of Medicine's Medline database. From this corpus, we will extract all instances of the pronoun *it*, together with syntactic information about its context. A set of vectors encoding this information will then be used to build a training and test set, which in turn will serve as input for the SVM. The results presented in section 3 will show how our approach outperforms previous methods relying on less extracted information from the data set.

The structure of the rest of the paper is as follows. Section 2 gives a more detailed explanation about previous research work and SVM. Section 3 presents details about the data set, our approach and discussion of results. Finally, section 4 summarizes our work.

# 2 Background

## 2.1 Support Vector Machines

Support Vector Machines (SVM) is a machine-learning technique based on statistical learning theory. The use of SVM for binary classification problems follows two steps. First, a set of vectors is constructed, each representing an instance of the input data (in our case, each vector will represent lexical and syntactic information extracted from the context of the pronoun *it*). This set of vectors is mapped into a feature space, usually with a higher dimension, either through a linear or a non-linear transformation. Second, a linear division (a hyperplane) that maximizes the margin between the two classes in the feature space is calculated.

SVM has many interesting properties: it is theoretically well-founded and shows a very good performance in practice, having been successfully applied in many NLP problems.

## 2.2 Previous Work on Identification of Non-Anaphoric *it*

We can classify methods for the identification of non-anaphoric *it* as either pattern-matching based or machine-learning based. Among the first ones, Paice and Husk (1987) have the highest performance with an accuracy of 93.9%. In this work a set of rules is constructed manually, and then every instance of the pronoun *it* is matched against the rule to decide if the use is anaphoric or not. Rules usually follow a similar scheme: mark left and right delimiters of the anaphoric/non-anaphoric use, and indicate the presence of certain kind of "special" word. For instance, the rule to detect non-anaphoric uses of the form *it...to*:

> an "it...to" construct: is delimited on the left by "it"; is delimited on the right by "to"; contains a task status word.

In this case, *task status word* is a list of certain kind of adjectives (such as *good, bad, hard, important...*) and nouns indicating purpose (*aim, goal, objective...*) which, if present inbetween the indicated delimiters, would trigger the rule marking the instance as non-anaphoric. In a similar way, there are lists for *state of knowledge words* (*certain, clear, known, doubtful...*), prepositions (*among, at, before, below...*) and idioms (expressions containing *it* which are treated as non-referential: *"on the face of it", "as it turned out", "as we know it"...*).

Pattern-matching methods achieve a reasonable accuracy despite the scarce amount of information with which they work. A collection of slightly over a hundred words combined with a smart selection of a few rules results in an accurate prediction of anaphoric and non-anaphoric uses of *it*.

Nevertheless, approaches based on patterns have certain drawbacks. The use of lists of words is troublesome in two senses: first, it is doubtful that such lists can ever be complete. In Lappin and Leass (1994), for instance, they make use of 15 adjectives and 5 verbs, which seems clearly insufficient. Denber (1998) recommends a list of meteorological verbs and adjectives, plus the use of adjectives marked in WordNet with the attributes *state, condition, quality* or *quantity* and verbs marked by the *engcg* tagger (Samuelson and Voutilainen, 1997) as cognitive. A second, more complex problem, arises with words that sometimes indicate the presence of non-anaphoric *it*, while other times they do not. In Paice and Husk (1987), for instance such "borderline" words are manually included or excluded from the lists when convenient in order to optimize the performance.

Evans (2001) used machine-learning method, with results slightly better than those of the most accurate pattern-matching approach. In this work, 35 syntactic features of the context of the pronoun are extracted with a commercial parser, including part-of-speech tags, morphological lemmas and dependency relations between certain words. The extracted information serves then as input to a memory-based learning (MBL) algorithm. MBL is based on the idea that learning relies on the reuse of previous experience rather than on extracted abstractions and generalizations.

While this approach has the best prediction accuracy so far and seems more scalable since it does not rely on lists of words or fixed constructions (as these are automatically "learned"), it also has some problems. Attributes like dependency relations between words need partial parsing of the sentence. Recent advances in anaphora resolution show how knowledge-poor approaches based exclusively on surface analysis can achieve as good results as those making use of extensive knowledge (see Mitkov (1998)). In our method, we just make use of POS tagging, a NP chunker and a morphological processor to extract lemmas of certain words.

# 3 Experiments and Discussion

## 3.1 Dataset

Our data set was constructed from instances of the pronoun *it* found in the GENIA corpus (version 3.02p), a collection of Medline abstracts selected from the search results with the keywords *Human, Blood Cells* and *Transcription Factors.* Differences in topic and writing style with other corpora make identification of non-anaphoric *it* more relevant. While in the SUSANNE and BNC corpora the percentage of non-anaphoric uses does not reach 30% of the total number of instances, in GENIA this percentage rises to nearly 44% of the cases.

For every instance of the pronoun, we constructed a vector containing information about its context. This contextual information was chosen as a subset of the attributes presented by Evans (2001), concretely those attributes that could be extracted without the need of a parser: POS tags, lemma of verbs and noun phrase chunks. A complete list of these attributes can be found in Table 1. The GENIA corpus is distributed with POS tagging. Lemma of verbs is obtained using *morpha* (Minnen et al., 2001), a morphological processing software based on a set of morphological generalisations together with a list of exceptions for irregular forms. Noun phrases were extracted using *BaseNP* (Ramshaw and Marcus, 1995), which makes use of heuristic transformational rules to bracket base noun phrase structures. A program written in Perl collected all the attributes together with the manually annotated class of every instance (i.e. anaphoric or non-anaphoric) to construct a file of 532 vectors, each representing the extracted 21 features.

| Attribute | Description |
|---|---|
| i | Line number |
| ii | Position in the line |
| iii-vi | POS tag of previous 4 words |
| vii-x | POS tag of next 4 words |
| xi | Distance to next compl. |
| xii | Distance to next gerund |
| xiii | Distance to next prep. |
| xiv | Lemma of previous verb |
| xv | Lemma of next verb |
| xvi | Adjective + NP sequence |
| xvii | Compl. + NP sequence |
| xviii | Number of previous prep. |
| xix | Number of following compl. |
| xx | Number of following adj. |
| xxi | Number of previous NP |

Table 1: Attributes

### 3.2 Experiment Setup and Performance Criteria

All experiments were performed using 7-fold cross validation. In order to compare performance, we implemented a pattern-based approach (Paice and Husk, 1987), and a MBL approach using the package *Timbl* (Daelemans et al., 2002). For our SVM-based approach, we used *LibSVM* (Chang and Lin, 2001).

For every approach, we calculated correctly predicted positive cases (true positives, $P$), correctly predicted negative cases (true negatives, $N$), uncorrectly predicted positive cases (false positives, $p$) and uncorrectly predicted negative cases (false negatives, $n$). From these values, we obtained the prediction accuracy of each method, percentage of correctly predicted cases (see equation 1). In order to avoid certain problems related to prediction accuracy (an unbalanced data set with too many positives or negatives can have a strong influence on the results), we also used Matthew's correlation coefficient (*Mcc*) as a measure criterion (see equation 2).

$$Acc = \frac{P+N}{P+N+p+n} * 100 \tag{1}$$

$$Mcc = \frac{PN - pn}{\sqrt{(P+n)(P+p)(N+n)(N+p)}} \tag{2}$$

Results for the first set of experiments are summarized in Table 2. For each approach we show the best results obtained with different parameter settings. In our experiments with MBL, the best results were obtained with the IB2 algorithm, incremental edited memory-based learning (Aha et al., 1991), with parameter *k=5* and modified value difference metric. For SVM we tried four different kernels (linear, polynomial, RBF and sigmoid), with linear one achieving the best performance.

| | Patterns | MBL | SVM |
|---|---|---|---|
| Accuracy | 90.789 | 92.295 | 92.717 |
| Mcc | 0.8174 | 0.8434 | 0.85646 |

Table 2: Prediction results

With the models trained by the memory-based learning and SVM, we also studied the relative importance of each attribute to predict instances. Results are summarized in Tables 3, 4 and 5. All results were normalized in the range (0,1) for a clearer comparison.

| Attribute | Chi-square |
|---|---|
| Distance to following compl. | 1 |
| Compl. + NP sequence | 0.8977 |
| Number of following compl. | 0.8116 |
| Lemma of previous verb | 0.7695 |
| Lemma of next verb | 0.6457 |
| Position in line | 0.5213 |
| POS - 1 | 0.5033 |
| POS + 2 | 0.4396 |
| POS + 3 | 0.4020 |
| POS - 2 | 0.3045 |

Table 3: Weight of most relevant attributes with MBL ($\chi^2$)

| Attribute | Gain Ratio |
|---|---|
| Compl. + NP sequence | 1 |
| Number of following compl. | 0.6810 |
| Distance to following compl. | 0.5255 |
| Lemma of next verb | 0.2313 |
| Lemma of previous verb | 0.2125 |
| POS - 1 | 0.1911 |
| POS + 2 | 0.1506 |
| POS + 3 | 0.1305 |
| Position in line | 0.1282 |
| POS + 1 | 0.1017 |

Table 4: Weight of most relevant attributes with MBL (GR)

| Attribute | Support |
|---|---|
| Number of following compl. | 1 |
| Compl. + NP sequence | 0.9879 |
| Lemma of next verb | 0.5862 |
| POS - 1 | 0.5397 |
| POS - 2 | 0.4863 |
| Lemma of previous verb | 0.4639 |
| POS + 4 | 0.4251 |
| POS + 2 | 0.3838 |
| POS - 4 | 0.3810 |
| POS + 3 | 0.3517 |

Table 5: Weight of most relevant attributes with SVM

### 3.3 Discussion

Results in section 3.2 clearly show how machine-learning based methods (MBL and SVM) outperform pattern-matching based ones (1.506% and 1.928% respectively). Although in Evans (2001) MBL and pattern methods performed similarly, our results show more clearly that approaches based on rules and sets of words do not adapt so well to different data sets. In addition, SVM prediction of non-anaphoric uses of *it* in the GENIA corpus is slightly more accurate than predictions obtained with MBL (0.422%). This difference should be taken with care though, since our data set is relatively small (532 instances, of which SVM correctly predicted 493, while MBL solved 490). The analysis of results for each fold shows that SVM outperformed MBL in 4 out of the 7 folds, and the variance was also slightly smaller with SVM (0.062 versus 0.090). We are therefore considering to conduct experiments with a bigger data set in order to confirm more clearly the improvement of performance with SVM.

The study of relevance of the attributes (which has not been addressed by previous studies) provides a more important insight in the nature of the problem. Tables 3 and 4 show normalized results for MBL according to $\chi^2$ and Gain Ratio measures. In both cases, the three attributes related to complementizers (*Distance to following complementizer, Complementizer + NP sequence* and *Number of following complementizers*) are the most relevant ones, although in different orders. The attributes *Lemma of previous verb* and *Lemma of next verb* follow in importance. These results are coherent with most studies on identification of non-anaphoric *it*. As noted by Paice and Husk (1987), structure of context surrounding non-anaphoric instances of *it* tends to be more stereotyped than in anaphoric instances, and is therefore easier to detect. The presence of a complementizer close to the pronoun often triggers non-anaphoric predictions, which makes attributes related to complementizers more relevant. Verbs are also a strong indicator of non-anaphoric uses: expressions such as *It remains unclear...*, *It is known...*, *It has been proved...*, etc. can also be correctly classified as non-anaphoric considering the verb following the noun. A preceeding verb usually indicates anaphoric instances.

Results with SVMs (see Table 5) are slightly different than those obtained with MBL. In this case *Number of following complementizers* and *Complementizer + NP sequence* are the two most relevant attributes, but *Distance to the following complementizer* is not even among the 10 most important. This could be (arguably) attributed to the fact that this attribute and the previous ones related to complementizers convey similar information, and SVM is able to construct an accurate prediction model by just using the information encoded in the first two arguments. Attributes related to verbs are also relevant, as in the case of MBL. Also, information about the POS tag of the word appearing before the pronoun shows slightly more importance than it does in MBL. It was also noticed that the attribute *Position in line* has nearly no relevance with SVM, while with MBL its relevance is above the average.

## 4 Conclusions

In this paper we have presented a SVM approach to identificate non-anaphoric uses of the pronoun *it*. Using a limited set of features representing lexical and syntactic information of the context, our approach achieves better accuracy than previous methods without the need of hand-made rules or a parser. The analysis of the relevance of features, which had not been addressed previously, gives a more complete description of the factors that influence non-anaphoric uses of *it*. Attributes related to complementizers and verbs showed to have the strongest impact on the final results, followed by information on the POS tag preceeding and following the pronoun.

## 5 Acknowledgements

We would like to thank Tho Hoam Pham, Dr. Jun'Ichi Kazama and Professor Tu Bao Ho for their comments on this work. We would also like to thank three anonymouse reviewers for their suggestions.

This work was supported by Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

## References

David W. Aha, Dennis Kibler, and Marc K. Albert. 1991. Instance-Based Learning Algorithms. *Machine Learning*, 6:37–66.

José Castaño, Jason Zhang, and James Pustejovsky. 2002. Anaphora Resolution in Biomedical Literature. In *International Symposium on Reference Resolution*, Alicante, Spain.

Chih-Chung Chang and Chi-Jen Lin, 2001. *LIBSVM: a Library for Support Vector Machines.* Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning*, 20(3):273–297.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2002. Tilburg Memory Based Learner, version 4.3, Reference Guide. Technical report, ILK Technical Report 02-10.

Michel Denber. 1998. Automatic Resolution of Anaphora in English. Technical report, Eastman Kodak Co.

Richard Evans. 2001. Applying Machine Learning Toward an Automatic Classification of *It*. *Literary and Linguistic Computing*, 16(1):45–57.

Shalom Lappin and Herbert Leass. 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535–561.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied Morphological Processing of English. *Natural Language Engineering*, 7(3):207–223.

Ruslan Mitkov. 1998. Robust Pronoun Resolution with Limited Knowledge. In *Proceedings of 17th International Conference on Computational Linguistics*, pages 869–875, Montreal, Canada.

Tomoko Ohta, Yuka Tateisi, Jin-Dong Kim, and Jun'ichi Tsujii. 2002. The GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. In *Proceedings of the Human Language Technology Conference*, San Diego CA, USA.

C. D. Paice and G. D. Husk. 1987. Towards an Automatic Recognition of Anaphoric Features in English Text: the Impersonal Pronoun 'it'. *Computer Speech and Language*, 2:109–132.

Lance Ramshaw and Mitchell Marcus. 1995. Text Chunking Using Transformation-Based Learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*, pages 82–94, MIT, USA.

Christer Samuelson and Arto Voutilainen. 1997. Comparing a Linguistic and a Stochastic Tagger. In *Proceedings of the ACL/EACL Joint Conference*, pages 246–253, Madrid, Spain.