Towards more comprehensive evaluation in anaphora resolution

Ruslan Mitkov

School of Humanities, Languages and Social Studies
University of Wolverhampton
Stafford Street
Wolverhampton WV1 1SB
United Kingdom
Email R.Mitkov@wlv.ac.uk

Abstract

The paper presents a package of evaluation tasks for anaphora resolution. We argue that these newly added tasks which have been carried out on Mitkov's (1998) knowledge-poor, robust approach, provide a better picture of the performance of an anaphora resolution system. The paper also outlines future work on the development of a "consistent" evaluation environment for anaphora resolution.

1. Introduction

The last few years have seen the emergence of a number of new projects on anaphora resolution, due to its importance in key NLP applications such as natural language interfaces, machine translation, automatic abstracting and information extraction. In particular, the recent search for practical robust, corpus-based approaches has produced promising solutions (Baldwin 1997; Cardie and Wagstaff 1999; Ge et al. 1998; Kameyama 1997; Mitkov 1996; 1998).

Against the background of growing interest in the field, it seems that still insufficient attention has been paid to the evaluation of the systems developed. Even though the number of works reporting extensively on evaluation in anaphora resolution is increasing (Azzam et al. 1998; Baldwin 1997; Cardie & Wagstaff 1999; Gaizauskas & Humphreys 1996; Lappin & Leass 1994; Mitkov 1998, 2000; Mitkov & Stys 1997; Tetrault 1999; Walker 1989), the forms of evaluation that have been proposed are not sufficient or perspicuous.

As in any other NLP task, evaluation is of crucial importance to anaphora resolution. The MUC (Message Understanding Conference) initiatives suggested the measures "recall" and "precision" be used for evaluating the performance of anaphora (coreference) resolution systems. These measures have proved to be useful indicators and they have already been used in some of the above mentioned works (Aone and Bennett 1995; Baldwin 1997, Gaizauskas & Humphreys 1996).[1]

It is felt, however, that evaluation in anaphora resolution needs further attention. Measuring the success rate of an anaphora resolution system in terms of "recall" and "precision" is an important step in assessing the efficiency of anaphora resolution approaches, but as we shall point out further in the paper, they may not serve as distinct measures for robust systems. In addition, it appears that they alone cannot provide a comprehensive overall assessment of an approach. In order to see how much a certain approach is "worth", it would be necessary to assess it against other "benchmarks", e.g. against other existing or baseline models. It also makes sense to evaluate the performance on anaphors which do not point to sole candidates for antecedents and which cannot be disambiguated on the basis of gender and number agreement alone (see the notion of *critical success rate*, section 3.1.3). Finally, a comparison with other similar or well-known approaches would be indicative of where the approach stands in the state of play of anaphora resolution.

Furthermore, the evaluation would be more revealing, if in addition to evaluating a specific approach as a whole, we break down the evaluation process by looking at the different components involved. In the case of factor-based anaphora resolution, we propose methods for evaluation of each individual factor employed in the resolution process. Such evaluation would provide important insights as to how the overall performance of factor-based systems could be improved (e.g. through changing the weights/scores of the factors). In this work we propose the notion of *decision power* of anaphora resolution factors which can play an important role in preferential architectures.

The paper is structured as follows. Section 2 briefly outlines the approach which we use as a testbed for our evaluation. Section 3 elaborates on the evaluation tasks and measures that we have taken on board. Section 4 discusses issues relevant to the evaluation in anaphora resolution, while section 5 proposes the idea of developing an evaluation environment.

2. Evaluation: using our robust, knowledge-poor anaphora resolution system as testbed

The approach which we used as a testbed for the evaluating methodology described was our robust, knowledge-poor approach to pronoun resolution (Mitkov 1998) which will be referred to as "the robust approach".

2.1. The robust approach: a brief outline

Our robust approach works as follows: it takes as an input the output of a text processed by a part-of-speech tagger, identifies the noun phrases which precede the anaphor within a distance of 2 sentences, checks them for gender and number agreement with the anaphor and then applies the so-called antecedent indicators to the remaining candidates by assigning a positive or negative score (-1, 0, 1 or 2). The noun phrase with the highest aggregate score is proposed as antecedent. Some indicators give NPs a bonus and are therefore called *boosting indicators* (e.g.

[1] However, see the comments in section 3.1.1 on some confusing definitions of these measures.

first noun phrases, *lexical reiteration*, *section heading*, *collocation pattern preference*, *immediate reference*, *referential distance*, *term preference*), whereas others penalise certain NPs and are referred to as *impeding indicators* (*indefiniteness*, *non-prepositional noun phrase*). Most of the indicators are genre-independent and related to coherence phenomena (such as salience and distance) or to structural matches, whereas others are genre-specific (term preference). For instance, *indefiniteness* considers indefinite noun phrases preceding the anaphor to have lesser chances of being the antecedent than definite ones and therefore, penalises the former by the negative score of –1. Also, *first noun phrases* in previous sentences/clauses are deemed good candidates for antecedents and score 1. For more details on the indicators and their patterns see (Mitkov 1998).

3. Evaluation: towards a more comprehensive framework of evaluation tasks

In the search for more comprehensive evaluation methodology, we have carried out evaluation tasks related to (i) the evaluation of the performance of the anaphora resolution system as a whole and (ii) the evaluation of separate components of the anaphora resolution algorithm.

3.1. Evaluation of the overall performance of the anaphora resolution system

3.1.1. "Traditional" evaluation: success rate

We evaluated our approach in terms of its resolution *success rate* expressed as the ratio [number of successfully resolved anaphors] / [number of all anaphors]. We prefer only to use the term success rate and not 'recall' or precision' for two reasons. First, there appears to be some terminological confusion as to what exactly recall and precision in anaphora resolution are. Aone and Bennett (1995) define recall as the ratio [number of correctly resolved anaphors] / [number of all anaphors identified by the system], and precision - as the ratio [number of correctly resolved anaphors] / [number of anaphors attempted to be resolved]. On the other hand, Baldwin (1997) defines recall[2] as the ratio [number of correctly resolved anaphors] / [number of all anaphors] and precision – as the ratio [number of correctly resolved anaphors] / [number of anaphors attempted to be resolved]. Note that Aone and Bennet count only those anaphors identified by the program, whereas Baldwin looks at all anaphors.[3] Secondly, if we adopt Baldwin's definition, robust approaches as ours (which propose an antecedent for each pronominal anaphor) would not be able to distinguish between recall and precision since both measures would be equal to the success rate.

The evaluation in English (we evaluated the approach for other languages as well - see below) included texts from different technical manuals (Minolta Photocopier, Portable Style Writer (PSW), Alba Twin Speed Video Recorder, Seagate Medalist Hard Drive, Haynes Car Manual, Sony Video Recorder) which contained a total of 223 anaphoric pronouns. The robust approach resolved 200 anaphors correctly which gives a success rate of 89.7%. The success rates were different for each of the technical manuals (Table 1) which shows that even for texts belonging to the same genre, results may differ. Therefore, for "more definitive" figures very large test data containing thousands of anaphors are needed.[4]

Manual	Success rate	Number of anaphoric pronouns
Minolta Photocopier	95.8	48
Portable Style Writer (PSW)	83.8	54
Alba Twin Speed Recorder	100.0	13
Seagate Medalist Hard Drive	77.8	18
Haynes Car Manual	80.0	50
Sony Video Recorder	90.6	40
All manuals	89.7	223

Table 1: Success rate(s) of the robust approach on different manuals

3.1.2. Evaluation against Baseline models

In order to evaluate the effectiveness of the approach and to explore whether or by how much it is superior to the baseline models for anaphora resolution, we also tested the sample texts on (i) a Baseline Model which checks agreement in number and gender and, where more than one candidate remains, picks out as antecedent the most recent subject matching the gender and number of the anaphor and (ii) a Baseline Model which selects as antecedent the most recent noun phrase that matches the gender and number of the anaphor. The evaluation results suggest a recall of 31.6% and a precision of 48.6 % for the first baseline model and a success rate (which is in this case the same as precision and recall) of 65.9% for the second (Table 2).

Approach	Success rate in %	Number of anaphoric pronouns	Comments
Robust approach	89.7	223	
Baseline Most Recent	65.9	223	
Baseline Subject	48.6/31.6	223	48.6 precision, 31.6 recall

Table 2: Comparison of the success rates of the robust approach and two baseline models

3.1.3. Critical success rate

We propose the measure *critical success rate* which applies only to those "tough" anaphors which still have more than one candidate for antecedent after gender and

[2] Baldwin's definition is in line with that proposed by Gaizauskas and Humphreys (1996).

[3] See also relevant comments at the end of section 4.

[4] Or more appropriate comprehensive sampling procedure.

number filters. More formally, if T is the set of anaphors in an evaluation corpus which do not have sole candidates for antecedents and whose antecedents cannot be identified on the basis of gender and number only, and if S ($S \subset T$) is the set of the anaphors which are resolved successfully, and if s = card S, t = card T, then

$$Critical\ success\ rate = \frac{s}{t}$$

The figure corresponding to critical success rate would normally be lower than the overall success rate. This is so because if f ($f \geq 0$) is the number of anaphors which have their antecedents as sole candidates or identified only after gender and number agreement checks, then the total number of anaphors in the evaluation corpus is t+f and the success rate would be (s+f)/(t+f) which is equal to or greater than s/t.

$$Critical\ success\ rate = \frac{s}{t} \leq \frac{s+f}{t+f} = Success\ rate$$

This measure would be an important criterion for evaluating the efficiency of a factor-based anaphora resolution system in the "critical cases" where agreement constraints alone cannot point to the antecedent.[5] It is logical to assume that good anaphora resolution approaches should have high critical success rates which are close to the overall success rates. In fact, it is really the critical success rate that matters: high critical success rate naturally implies high overall success rate.

In our case, the critical success rate exclusively accounts for the performance of the antecedent indicators since it is associated with anaphors whose antecedents can be tracked down only with the help of the antecedent indicators.

We measured the critical success rate as 82% on the basis of the Portable Style Writer manual (Table 3). This figure[6] (note that the success rate on these texts is 83.8%) and the significantly lower success rates of the Baseline Most Recent and of the Baseline Subject undoubtedly demonstrate the efficiency of the antecedent indicators.

3.1.4. Comparison to similar approaches: comparative evaluation of Breck Baldwin's CogNIAC

We felt it appropriate to extend the evaluation of the robust approach by comparing it to Breck Baldwin's CogNIAC approach (Baldwin 1997). The reason for this is that both our approach and Breck Baldwin's approach share common principles (both are knowledge-poor and use a POS tagger to provide the input) and therefore a comparison would be appropriate.

CogNIAC operates in two modes: "high precision" mode in which, if the 6 basic rules on which the approach is based do not resolve the pronoun, then it is left unresolved, and "resolve all" mode which includes two additional rules. Since our approach is robust and returns an antecedent for each pronoun, in order to make the comparison as fair as possible, we used CogNIAC's "resolve all" version by simulating it manually on the texts from the Portable Style Writer.

CogNIAC successfully resolved the pronouns in 75% of the cases. This result is compatible with the results described in (Baldwin 1997).

3.1.5. Comparison to "classical" approaches: Hobbs' Naive Algorithm

We have also started comparative evaluation of Jerry Hobbs' naive algorithm (Hobbs 1976) on the basis of the same texts used for the comparative evaluation of Baldwin's approach (StyleWriter 1994). The preliminary results obtained suggest a success rate in the range of 71%.

The above results (Table 3) show that on this small set of data from the genre of technical manuals, the robust approach performs better than Baldwin's or Hobbs' approaches. These results, however, cannot be generalised for other genres or unrestricted texts and for a more accurate picture, further extensive tests are necessary.

Approach	Success rate in %	Critical success rate	Number of anaphoric pronouns
Robust approach PSW	83.8	82	54
Baldwin's CogNIAC	75		54
Hobb's naïve algorithm	71		54

Table 3: Comparative evaluation and critical success rate based on the PSW corpus

3.2. Evaluation of separate components of the anaphora resolution algorithm: antecedent indicators in focus

We believe that the evaluation of each antecedent indicator is very important because (i) it gives us an idea of the relative importance of each indicator and (ii) it provides a basis upon which we can fine-tune the indicator scores and thus attain an overall improvement to the approach. As will be seen below, the importance should not be reduced to the confidence of a certain indicator in rare situations, but should be also regarded as the useful contribution of factors on a more frequent basis.

We propose the notion *decision power* as a measure of the influence of each factor (in our case indicator) on the final decision, its ability to "impose" its preference in line with, or contrary to the preference of the remaining indicators. The decision power (DP_K) of a boosting indicator K is defined in the following way:

$$DP_K = \frac{SI_K}{A_K}$$

where SI_K is the number of successful antecedent identifications (resolutions) when this indicator is applied and A_K is the number of applications of this indicator. For the penalising indicators *prepositional noun phrase* and *indefiniteness* this figure is calculated as

[5] Factor-based systems typically employ a number of factors after gender and number checks.

[6] This figure was obtained on a comparatively small set of data; recent unpublished tests have confirmed the good critical success rates of the approach.

$$DP_K = \frac{UI_K}{A_K}$$

where UA_K is the number of unsuccessful antecedent identifications and A_K the number of applications of this indicator. The immediate reference emerges as the most "influential" indicator (1), followed by prepositional noun phrase (0.922), collocation (0.909), section heading (0.619), lexical reiteration (0.585), first NP (0.493), term preference (0.357) and referential distance (0.344) (Table 4). The relatively low figures for the majority of (seemingly very useful) indicators should not be regarded as a surprise: firstly, we should bear in mind that in most cases a candidate is picked (or rejected) as an antecedent on the basis of applying a number of different indicators and secondly, that most anaphors have a relatively high number of candidates for antecedent.

Indicator	Decision power	Comments
Immediate reference	1	Very decision-powerful, points always to the correct candidate
Prepositional noun phrase	0.922	Very decision-powerful and discriminating
Collocation	0.909	Very decision-powerful and discriminating
Section heading	0.619	Fairly decision-powerful, but alone cannot impose the antecedent
Lexical reiteration	0.585	Sufficiently decision-powerful
First NP	0.493	Averagely decision-powerful
Term preference	0.357	Not sufficiently decision-powerful
Referential distance	0.344	Not sufficiently decision-powerful

Table 4: Decision power values for the antecedent indicators

Another way of measuring the importance of a specific factor (in our case indicator) would be to evaluate the approach with this indicator (factor) "switched off"[7]. We call this measure *indispensability* since it shows how vital, indispensable the presence of specific factor is. Indispensability (Ind_K) for a given indicator K is defined as

$$Ind_K = \frac{SR - SR_{-K}}{SR}$$

where SR_{-K} is the success rate obtained when the indicator K is excluded, and SR is the success rate (with all the indicators on). In other words, indispensability is a measure for the non-absolute, relative contribution of this indicator to the "collective efforts" of all indicators: this measure shows how much the approach would lose out if the specific indicator were removed. It should be noted that being indispensable does not mean decision-powerful, confident and vice-versa. For instance, we found that *referential distance* has the highest value for indispensability, whereas this factor is among the least 'confident' ones. One possible explanation comes from the fact that indicators such as *immediate reference* and *collocation pattern preference* are applied relatively seldom and even though they impose their decision very strongly towards the correct antecedent, they do not score very highly as indispensable factors given their infrequent intervention. Finally, due to the complicated interactions of all indicators, there is no direct correlation between these two measures.

4. Discussion

In the previous sections we proposed different methods for assessing the performance of an anaphora resolution approach. Traditionally, evaluation is done by calculating the success rate (or recall and precision) on the basis of different test samples. These samples should be sufficiently representative. What has emerged is that the evaluation has to cover not hundreds of anaphors but many thousands: we have already seen that even in the same genre, results may differ if the samples are not large enough (Table 1).[8] Theoretically speaking, the success rate, the recall or precision figures could be regarded as definitive only if the approach were tested on all naturally occurring texts, which of course is an unrealistic task. Nevertheless, this consideration highlights the advantages of carrying out the evaluation task automatically. Automatic evaluation requires a large corpus with annotated coreferential links, against which the output of the anaphora resolution systems is to be matched. We have already started working on the development of coreferentially annotated corpora, with a view to using them in the evaluation process (Mitkov et al. 1999).

Evaluation should also provide information as to how effective an approach is, by comparing it with typical baseline models. Even if the success rate obtained from a number of test samples were high, would an approach be worth developing if it were only minimally more successful than a baseline approach? We believe that evaluation against baseline models is an imperative task in justifying the usefulness of the approach developed: unless the approach demonstrates clear superiority over baseline models, it may not be worthwhile developing it at all. The new measure *critical success rate* that we proposed should provide an equally useful insight.

In addition, evaluation helps us discover what the new approach brings to the current state of play of the field. Therefore, the approach should be compared with other similar methods (if available) and with other well known ("classical") methods (e.g. Hobbs' algorithm). However, we should point out that it is difficult to reach definite conclusions even if we compare "anaphora resolution proper" on the basis of the same data since due to the different pre-processing tools used, the added error may vary.

[7] Similar techniques have been used in (Lappin & Leass, 1994).

[8] An alternative way would be to employ comprehensive sampling procedures.

This raises the question of how we can evaluate the accuracy of the anaphora resolution algorithm independently of any pre-analysis. It makes sense to measure the efficiency of the anaphora resolution algorithm in an "ideal environment" and to identify the error rate due to the pre-processing tools (e.g. POS tagger, parser etc.). This is problematic too, because unless we do it by hand assuming that the analysed input to the anaphora resolver is 100% correct, the final results will always depend on the reliability of pre-processing.

Another purpose of the evaluation is to help the developers improve their system by looking at the relative importance/impact of each factor. Also, error diagnosis proves to be particularly important since it singles out the typical failure cases - it is through error diagnosis that we looked at the failure cases of the first version of the robust approach and came back with ideas for improvement (Mitkov & Stys, 1997).

Finally, it is worth mentioning that so far we have looked at the evaluation of the performance of the resolution program. This performance depends to a certain extent on accuracy of pre-processing results but also, on the accuracy of identifying pleonastic pronouns. Before attempting to resolve anaphora in English, the program should recognise and remove all occurrences of pleonastic (non-anaphoric) *it* such as in expressions like *it is recommended that, it is raining* etc. The recognition of pleonastic pronouns is therefore a task intrinsic to the process of anaphora resolution and its evaluation has to be addressed as well. The evaluation of the performance of the module for recognising pleonastic pronouns can be done either on its own or as a combined measure with the success rate (or recall and precision) of the anaphora resolution algorithm, for instance.

5. A way forward

In order to develop a "fair", consistent and accurate evaluation environment, and to address some of the problems identified above, we are developing a set of pre-processing tools which will enable us to test all anaphora resolution approaches (algorithms) sharing common principles (e.g. POS tagger, NP extractor, parser). This is a time-consuming task, given that we may have to re-implement most of the algorithms (for instance, we are re-implementing J. Hobb's algorithm as well as Lappin and Leass' one). This will give us a better picture as to the advantages and disadvantages of the different approaches: for a fairer comparison it is important that all compared approaches use the same pre-processing tools. Developing our own evaluation environment (and even re-implementing some of the key algorithms) also alleviates the formidable difficulties associated with obtaining the codes of the original programs.

One reservation must be stated, however: the comparison will still be approximate for approaches which do not share the same pre-processing philosophy, for instance the comparison between an approach which uses a full parser and an approach which relies on a POS tagger and NP extractor.

Finally, given the rarity of correferentially annotated corpora to be used for automatic evaluation, one of the priorities of our research on evaluation in anaphora resolution is to develop such corpora: we have already started producing such a corpus in the genre of technical manuals and will soon move to other genres such as newspaper articles and research papers.

6. Conclusion

The paper argues that in the absence of a wide range of "universal" benchmarks which could serve as evaluation samples for the different anaphora resolution approaches, the evaluation task would be more complete if (in addition to the sufficient number of random sample tests for measuring the recall and precision, or, for robust approaches, simply the success rate), the following additional evaluation tasks were carried out:

- comparative evaluation of baseline models
- comparative evaluation of other similar methods
- comparative evaluation of other well-known methods
- evaluation of the performance on anaphors which cannot be identified on the basis of gender and number agreement only (critical success rate)

The paper presents the results obtained from evaluating the author's robust pronoun resolution approach, proposes new measures relevant to evaluation in anaphora resolution and maintains that an anaphora resolution approach/system is only worth developing if it demonstrates clear superiority over baseline models.

7. References

Aiwa. 1996. Operating Instructions for the Aiwa Compact Disc Stereo System. Aiwa Corporation.

Alba. 1995. Alba Twin Speed Video Recorder Instruction Manual.

Aone, Chinatsu & Scot W. Bennett. 1995. "Evaluating automated and manual acquisition of anaphora resolution rules". Proceedings of ACL'95, 122-129.

Azzam, Saliha, Kevin Humphreys & Robert Gaizauskas. 1998. "Coreference resolution in a multilingual information extraction". Proceedings of the Workshop on Linguistic Coreference. Granada, Spain.

Baldwin, Breck. 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, 38-45, Madrid, Spain.

Brennan, S., M. Fridman and C. Pollard. 1987. A centering approach to pronouns. Proceedings of the 25th Annual Meeting of the ACL (ACL'87), 155-162. Stanford, CA, USA.

Cardie, Claire & Kiri Wagstaff. 1999. "Noun phrase coreference as clustering". Proceedings of the 1999 Joint SIGDAT conference on Empirical Methods in NLP and Very Large Corpora (ACL'99). 82-89. University of Maryland, USA.

Dagan, Ido. John Justeson, Shalom Lappin, Herbert Leass & Amnon Ribak. 1995. Syntax and lexical statistics in anaphora resolution. Applied Artificial Intelligence, 9.

Gaizauskas, Robert & Kevin Humphreys. 1996. Quantitative evaluation of coreference algorithms in an information extraction system. Paper presented at the Discourse Anaphora and Anaphor Resolution Colloquium (DAARC), Lancaster, UK

Ge, Niyu, John Hale & Eugene Charniak. 1998. "A statistical approach to anaphora resolution". Proceedings of the Workshop on Very Large Corpora, 161-170. Montreal, Canada.

Haynes. 1994. Haynes car manual.
Hobbs, Jerry R. 1978 "Resolving pronoun references". Lingua, 44, 339-352.
Internet Manual. 1994. Translation of Internet Manual Internet i okolice: Przewodnik po swiatowych sieciach komputerowych. Tracy LaQuey, Jeanne C. Ryer Translated by Monika Zielinska, BIZNET Poland.
Java Manual. 1998. Jezyk Java. Clico, Krakow.
Kameyama, Megumi. 1997. "Recognizing referential links: an information extraction perspective" Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution, 46-53. Madrid, Spain.
Lappin, Shalom & Herbert Leass. 1994. "An algorithm for pronominal anaphora resolution". Computational Linguistics, 20(4), 535-561.
Minolta. 1994. Minolta Operator's Manual for Photocopier EP5325. Technical Manual Minolta Camera Co., Ltd., Business Equipment Division 3-13, 2-Chome, Azuchi, - Machi, Chuo-Ku, Osaka 541, Japan.
Mitkov, Ruslan. 1996. "Pronoun resolution: the practical alternative". Paper presented at the Discourse Anaphora and Anaphor Resolution Colloquium (DAARC'96), Lancaster, UK.
Mitkov, Ruslan. 1997. "Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches" Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution, 14-21. Madrid, Spain.
Mitkov, Ruslan. 1998. "Robust pronoun resolution with limited knowledge". Proceedings of the 18.th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference, 869-875. Montreal, Canada.
Mitkov, Ruslan. 2000. "Multilingual anaphora resolution". Machine Translation. (forthcoming)
Mitkov, Ruslan & Malgorzata Stys. 1997. "Robust reference resolution with limited knowledge: high precision genre-specific approach for English and Polish". Proceedings of the International Conference "Recent Advances in Natural Language Proceeding" (RANLP'97), 74-81. Tzigov Chark, Bulgaria.
Mitkov, Ruslan, Constantin Orasan & Richard Evans. 1999. "The importance of annotated corpora for Natural Language Processing: the case of anaphora resolution and clause splitting". Proceedings of the TALN'99 workshop on Corpora and NLP, 60-69. Cargèse, France.
Seagate. 1997. Seagate Medalist Hard Drive Installation Guide.
Sony. 1992. Video cassette recorder. Operating Instructions. Sony Corporation.
Stylewriter 1994. Portable StyleWriter. User's guide. Apple Computers.
Tetreault, Joel R. 1999. "Analysis of Syntax-Based Pronoun Resolution Methods". Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '99), 602-605. Maryland, USA.
Walker, Marylin. 1989. "Evaluating discourse processing algorithms". Proceedings of the 27th Annual Meeting of the ACL (ACL'97. Vancouver, Canada.