Anaphora Resolution: To What Extent Does It Help NLP Applications?

Ruslan Mitkov, Richard Evans, Constantin Orăsan, Le An Ha,
and Viktor Pekar

University of Wolverhampton
Research Group in Computational Linguistics
Research Institute in Information and Language Processing
Stafford Street, Wolverhampton, WV1 1SB, United Kingdom
r.mitkov@wlv.ac.uk

Abstract. Papers discussing anaphora resolution algorithms or systems usually focus on the intrinsic evaluation of the algorithm/system and not on the issue of *extrinsic evaluation*. In the context of anaphora resolution, extrinsic evaluation concerns the impact of an anaphora resolution module on a larger NLP system of which it is part. In this paper we explore the extent to which the well-known anaphora resolution system MARS [1] can improve the performance of three NLP applications: text summarisation, term extraction and text categorisation. On the basis of the results so far we conclude that the deployment of anaphora resolution has a positive albeit limited impact.

1 Introduction

Papers discussing anaphora resolution algorithms or systems[1] usually describe the work of the algorithm or the system. In the majority of cases, they also report evaluation results related to its performance. This type of evaluation is known as *intrinsic* evaluation and accounts for the performance of the algorithm/system which is measured in terms of metrics such as recall, precision or success rate [3,1]. On the other hand, *extrinsic evaluation* in the context of anaphora resolution concerns the impact of an anaphora resolution module on a larger NLP system of which it is part. In this paper we address the issue of extrinsic evaluation in anaphora resolution and explore for the first time the extent to which our anaphora resolution system MARS [1] can improve the performance of three NLP applications: text summarisation, term extraction and text categorisation.

Section 2 of this paper will introduce Mitkov's original knowledge-poor algorithm, whereas section 3 will discuss its fully automatic implementations: the early version (hereafter referred to as MARS02) and the recent version (MARS06). Section 4 will outline the evaluation data used in our experiments and Section 5 will report the evaluation results when deploying MARS in three

[1] For definition of the distinction between anaphora resolution algorithms and anaphora resolution systems, see [2].

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 179–190, 2007.

different NLP applications: text summarisation, term extraction and text categorisation. Section 6 will provide a discussion of the evaluation results and finally section 7 will offer concluding remarks.

2 Brief Overview of Mitkov's Original Knowledge-Poor Pronoun Resolution Algorithm

MARS is based on Mitkov's [4,5] robust, knowledge-poor approach to pronoun resolution which was motivated by the pressing need in the 1990s for anaphora resolution algorithms operating robustly in real-world, knowledge-poorer environments in order to meet the demands of practical NLP systems. The first version of the algorithm was reported in [4] as an inexpensive, fast and yet reliable alternative to the labour-intensive and time-consuming construction of a knowledge-based system.[2] This project was also an example of how anaphors can be resolved quite successfully (at least in a specific genre, namely computer/technical manuals) without any sophisticated linguistic knowledge, even without parsing. In addition, the evaluation showed that the basic set of factors employed (referred to as 'indicators', see below) can work well not only for English, but also for other languages.

Mitkov's approach relies on a list of preferences known as *antecedent indicators.* The approach operates as follows: it works from the output of a text processed by a part-of-speech tagger and an NP extractor, identifies noun phrases which precede the anaphor within a distance of 2 sentences[3], checks them for gender and number agreement with the anaphor and then applies the indicators to the remaining candidates by assigning them a positive or negative score (2, 1, 0 or -1). The noun phrase[4] with the highest composite score is proposed as antecedent.

The antecedent indicators are applied to all NPs which have passed the gender and number filter.[5] These indicators can act in either a *boosting* or an *impeding* capacity. The boosting indicators apply a positive score to an NP, reflecting a positive likelihood that it is the antecedent of the current pronoun. In contrast, the impeding ones apply a negative score to an NP, reflecting a lack of confidence that it is the antecedent of the current pronoun. Most of the indicators are genre-independent and related to coherence phenomena (such as salience and distance)

[2] The approach has become better known through a later updated publication [5].

[3] Subsequent versions of the approach have used search scopes of different lengths (e.g. 2, 3 or 4 sentences), but the original algorithm only considered two sentences prior to the sentence containing the anaphor. The NP patterns are limited to the identification of base NPs and do not include complex or embedded phrases.

[4] The approach only handles pronominal anaphors whose antecedents are noun phrases.

[5] The approach takes into consideration the fact that in English there are certain collective nouns which do not agree in number with their antecedents (e.g. *government, team, parliament* etc. These entities and can be referred to by plural pronouns; equally some plural nouns such as *data* can be referred to by *it*) and are thus exempted from the agreement test.

or to structural matches, whereas others are genre-specific.[6] The boosting and impeding indicators are described in detail in [5]. The work presented in [1] provides some additional detail on the indicators used by the algorithm.

The aforementioned antecedent indicators are preferences and not absolute factors. There might be cases where one or more of the antecedent indicators do not 'point' to the correct antecedent. For instance, in the sentence 'Insert the cassette into the VCR making sure it is turned on', the indicator *prepositional noun phrases* would penalise the correct antecedent. When all preferences (antecedent indicators) are taken into account, however, the right antecedent is still likely to be tracked down - in the above example, the *prepositional noun phrases* heuristic stands a good chance of being overturned by the *collocation match* heuristics since the collocation *The VCR is turned on* is likely to appear previously in the text, as it is a typical construction in video technical manuals.

The antecedent indicators have proved to be reasonably efficient in identifying the correct antecedent and the results show that for the genre of technical manuals they may be no less accurate than syntax- and centering-based methods (see [5]). The approach described is not dependent on any theories or assumptions; in particular, it does not operate on the assumption that the subject of the previous utterance is the highest-ranking candidate for the backward-looking center - an approach which can sometimes lead to incorrect results.[7]

3 Outline of MARS

Mitkov's algorithm was enhanced and developed into the fully-automatic pronoun resolution system referred to as MARS.[8] The initial, as well as the current implementations of MARS, which both employ the FDG shallow parser [6] as their main pre-processing tool, are based on a revised version of the original algorithm.

3.1 Early Version of MARS

The initial implementation of MARS [1] followed Mitkov's original approach closely, the main differences being (i) the addition of three new indicators and (ii) a change in the way some of the indicators were implemented or computed due to the available pre-processing tools. Later, MARS also incorporated a program for automatically recognising instances of anaphoric or pleonastic pronouns [7] and intra-sentential syntax filters. This early version of MARS is referred to as MARS02 in the evaluation below.

[6] Typical of the genre of user guides.

[7] For instance, subject-favouring methods or methods relying heavily on syntactic parallelism would incorrectly propose *the utility* as the antecedent of *it* in the sentence 'The utility shows you *the LIST file* on your terminal for a format similar to that in which it will be printed' as it would prefer the subject as the most salient candidate. The *indicating verbs* preference of Mitkov's approach, however, would prefer the correct antecedent *the LIST file*.

[8] MARS stands for Mitkov's Anaphora Resolution System.

The system operates in five phases. In *phase 1*, the text to be processed is parsed syntactically, using Conexor's FDG Parser [6] which returns the parts of speech, morphological lemmas, syntactic functions, grammatical number, and dependency relations between tokens in the text, facilitating complex noun phrase (NP) extraction.

In *phase 2*, anaphoric pronouns are identified and non-anaphoric and non-nominal instances of it are filtered using the machine learning method described in [7].

In *phase 3*, for each pronoun identified as anaphoric, candidate NPs are extracted from the heading of the section in which the pronoun appears, and from the current and preceding two sentences within the paragraph under consideration. Once identified, these candidates are subjected to further morphological and syntactic tests. Extracted candidates are expected to obey a number of constraints if they are to enter the *set of competing candidates*, i.e. the candidates that are to be considered further. Competing candidates are required to agree with the pronoun in terms of number and gender, as was the case in the original algorithm. They must also obey syntactic constraints [1].

In *phase 4*, 14 preferential and impeding factors are applied to the sets of competing candidates. On application, each factor applies a numerical score to each candidate, reflecting the extent of the system's confidence about whether the candidate is the antecedent of the current pronoun. In the implemented system, certain practical issues led to the weights assigned by indicators being computed in a different way from that described in the original algorithm. The full details of these differences are beyond the scope of the current paper, but they are described in detail in [1]. In addition, three new indicators were added, one of which (*syntactic parallelism*) exploits new, previously unavailable features of the pre-processing software.

Finally, in *phase 5*, the candidate with the highest composite score is selected as the antecedent of the pronoun. Ties are resolved by selecting the most recent highest scoring candidate.

3.2 Recent Version of MARS

The more recent version of MARS incorporates several advancements over the system described in [1]. These improvements introduce the inclusion of more precise and strict number and gender agreement, and the addition of one indicator which employs the modelling of selectional restrictions. This recent version is referred to as MARS06 in the evaluation below.

MARS was improved to cater for several frequent causes of apparent number disagreement. These consist of (i) collective nouns, (ii) NPs whose gender is under-specified, (iii) quantified nouns/indefinite pronouns, and (iv) organisation names. These cases were handled by a combination of gazetteers, the integration of an animacy recognition module [8], and named entity recognition [9]. Patterns were used to identify the occurrence of quantified NPs in the parsed text. MARS's recognition of the gender of NP candidates has also been improved. In addition to gazetteers, a NER system is used to recognise person names and the system for NP animacy recognition is deployed.

Following work such as that described in [10], a new salience indicator was implemented. The selectional preference indicator processes a pronoun, exploits functional dependency information provided by the FDG Parser, and extracts a pair consisting of the verb on which the pronoun depends and the functional role of the pronoun. The selectional preference for the verb is then modeled by means of a distributional approach and used to obtain the likelihood for each candidate NP that it is a potential argument of the verb given that it has the same role as the pronoun. The selectional preference model is used to determine the most likely candidate in the set, and this candidate is awarded a boosting score of +1.

4 Evaluation Data

In this paper, a corpus of newspaper articles published in *New Scientist* was used. There were several reasons for selecting texts from this magazine. First, texts were required which contain a relatively high number of pronouns and are ideally not too different from texts from the technical domain, for which MARS was initially designed. We decided against using technical documents for two reasons. Firstly, we wanted to see how well MARS performs on texts from a different domain and secondly, technical documents are rather long and unsuitable for some types of manual annotation (e.g. coreferential links and automatic summarisation annotation, as explained later in this section). By contrast, the texts from New Scientist were preferred because they were short enough to be manually annotated and were suitable for all the extrinsic evaluation tasks performed.

Fifty-five texts from New Scientist distributed in the BNC were included in our corpus. These texts contained almost 1,200 third person pronouns and over 48,000 words. Before selecting the texts, a filter was applied to ensure that very short (under 2 kilobytes including the SGML annotation) and very long (over 15 kilobytes including the SGML annotation) texts were not included in the corpus. The reason for filtering out texts that are too short is that such texts could not be used in automatic summarisation (see extrinsic evaluation, section 5.1), whereas texts which were too long cannot be reliably annotated. The texts from New Scientist also proved appropriate for the other two extrinsic evaluation tasks investigated in this paper. They are scientific texts that contain a relatively large number of terms and are therefore appropriate for the application of automatic term extraction methods. Further, as they address different clearly identifiable topics, it is feasible for them to be categorised by automatic means.

All the texts in the corpus were annotated with several layers of annotation using PALinkA, a multipurpose annotation tool [11]. First, the texts were annotated for coreferential links using the methodology for NP coreference described in [12]. Once all the markables were identified by the annotators, the head of each one was manually marked in order to facilitate evaluation. Six files, accounting for about 10% of the corpus, were annotated by two annotators in order to assess the inter-annotator agreement. Using the method described in [13], sets of coreference chains were derived from each annotated file and each pair of elements in a chain was used to produce an exhaustive set of the coreferential links annotated in a document. The set of coreference links derived from the annotations of one annotator is then considered to be the key while the set derived from the annotations made by the other annotator is considered to be the response. On the annotated data used in the current study, evaluation revealed an average F-score of 0.6601 between the two annotators.

In order to be able to run the extrinsic evaluation with regard to text summarisation, the corpus was also annotated with information about important sentences. The methodology applied to annotate the important sentences in a text is the one described in [14] and entailed identification of 15% of the text as essential and of a further 15% of the text as important. In this way, it was possible to evaluate automatic summaries at two different compression rates: 15% and 30%.

In order to evaluate the effect of MARS on automatic term extraction, a reader with good general knowledge was asked to read the same texts annotated with coreference information, and identify terms appearing in each text. These sets of terms served as the gold standard in the evaluation experiment on the basis of which precision, recall and F-measure are computed.

To evaluate automatic classification, each text was annotated with a relevant label derived from the New Scientist web site: "health", "earth", "fundamentals", "being human", "living world", "opinion" and "sex and cloning". The texts from the BNC were not labeled for their original category and so they had to be assigned manually by our annotators. Texts for which none of these labels seemed wholly suitable were assigned the category "other". The 55 texts annotated with coreference information and important sentences were not sufficient to train and test a classifier, and so a further 120 texts from New Scientist were selected and annotated with information about their class. Given that annotation for coreference and summarisation is difficult and time consuming, these 120 were not also annotated with this information.

5 Evaluation

Papers discussing anaphora resolution usually describe the work of the algorithm or the system. In the majority of cases, they also report evaluation results related to the performance of the algorithm/system which is known as intrinsic evaluation and which accounts for its performance. In this paper we shall not discuss the performance of MARS in terms of intrinsic evaluation and for the first time in the literature, we shall seek to focus on the extrinsic evaluation with a view to establishing the extent to which the deployment of our anaphora resolution system, MARS [1], can improve the performance of various NLP applications. For details on the intrinsic evaluation of MARS, the reader is referred to [1] and [15] where the success rate of MARS is reported to range from 45% to 65% depending on the evaluation data. As pointed out in both papers, the performance of fully automatic anaphora resolution systems is markedly inferior to the performance of algorithms which benefit from post-edited input, i.e., from "perfect" pre-processing. Despite the comparatively low figures reported, MARS still fares as one of the best performing systems operating from the input of a shallow parser. Over the test data described in the present study, MARS02 performs with an average success rate of 0.4663, whereas MARS06 has an average success rate of 0.4947.

5.1 Summarisation

We evaluate the potential usefulness of anaphora resolution (as performed by MARS) in term-based summarisation which operates on the premise that it is possible to determine the importance of a sentence on the basis of the words it contains. The most common way of achieving this is to weight all the words in a text and calculate the score of a sentence by adding the weights of the words from it. In this way, a summary can be produced by extracting the sentences

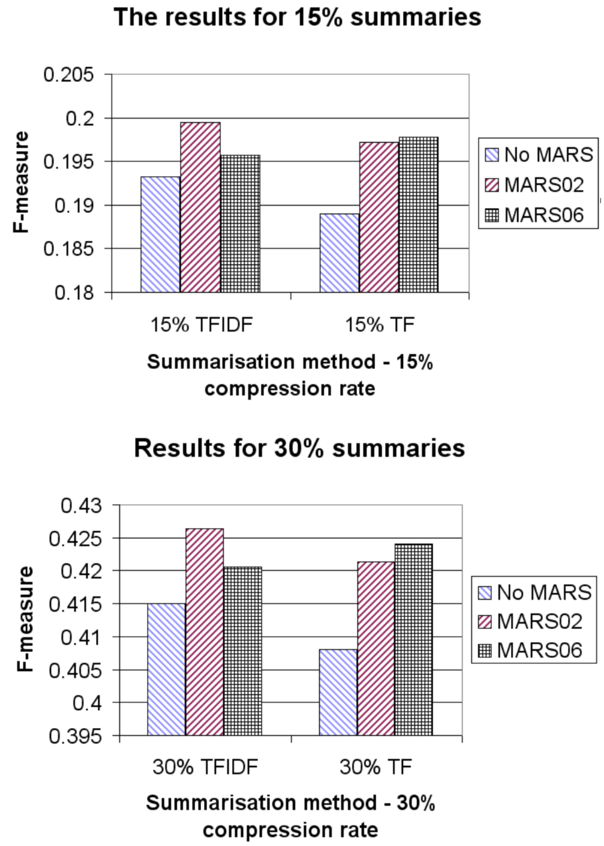

Fig. 1. The results of the automatic summarisation evaluation

with the highest scores until the desired length is reached. In order to calculate the importance of sentences several statistical measures can be applied [16], but the majority of them require that at least the frequency of the word in the document is known. For this reason, words which are referred to by pronouns do not have their weight correctly calculated. For the purpose of this extrinsic evaluation, we integrated MARS into a term-based summariser in an attempt to produce more accurate word weightings and as a result to improve the quality of the summaries produced.

In this paper two term weighting methods are investigated: term frequency and TF*IDF. The corpus used in this evaluation is the one described in Section 4 and the evaluation measures are precision, recall and f-measure. As explained earlier the corpus is annotated for 15% and 30% extracts, so the evaluation was performed for both compression rates. Figure 1 presents the results of the evaluation.

For both compression rates the value of F-measure increases when an anaphora resolution method is used by the summarisation method, but this increase is not statistically significant according to the paired t-test at the .05 level. For term frequency the results are better when the recent version of MARS is used, whereas for TF*IDF the best results are obtained by the early version of MARS.

5.2 Term Extraction

To examine the extent to which the employment of MARS could improve the performance of an automatic term extraction method, we compare the performances of a term extraction engine when run on various versions of a text: the original one and the ones in which pronouns are replaced by the antecedents proposed by MARS02 and MARS06. The term extraction method used is based on a hybrid approach which combines statistical and lexical-syntactic filters similar to [17] and [18]. First n-grams satisfying the POS pattern [AN]*NP?[AN]*N are collected, and then their TF*IDF scores are calculated. Candidates with a frequency count greater than one and TF*IDF score greater than 0.4[9] are selected. The set of New Scientist texts from the BNC is used as the document collection in the calculation of TF*IDF.

In our experiment, the term extraction engine extracts terms from three versions of a text: the original text, the text processed by MARS02, and the one processed by MARS06.[10] For each version of a text, precision, recall, and F-measure are calculated using the gold standard described in Section 4. The average F-measures are shown in Figure 2.

For both versions of MARS, there are improvements in the performance of the term extraction engine (measured using F-measure) although the improvements are not statistically significant, according to the paired t-test. MARS06 does not seem to boost the performance of term extraction over MARS02. MARS02 improves both precision and recall, whereas the main improvement engendered by MARS06 is in terms of recall. In 41% of the texts processed by MARS02, there

[9] These thresholds were determined empirically.

[10] By “processed” we mean that pronouns in the text have been replaced by the antecedents proposed by MARS.

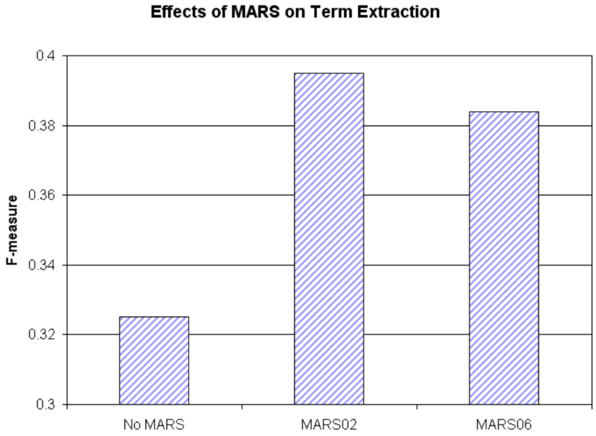

Fig. 2. The effects of two versions of MARS on automatic term extraction

is improvement in F-measures. Declining F-measures are observed in 33% of the texts and there is no change of the F-measure in the rest of the texts (26%).

5.3 Text Classification

In this experiment we examined the influence of anaphora resolution on the quality of automatic text classification. We experimented with four different text classification methods: k nearest neighbours (kNN), Naïve Bayes (NB), Rocchio, Maximum Entropy (MaxEnt), and Support Vector Machines(SVM).[11] We assess the quality of classification models that are learned from documents, in which pronouns are substituted for the noun phrases recognised as their antecedents by the MARS system. The model is then tested on documents where pronouns have been similarly replaced for antecedent noun phrases. Three different models are included in the experiment. The first two are learned from a document collection on which prior anaphora resolution has been applied: one using the early version of MARS (MARS02) and another that uses its new version incorporating proposed improvements (MARS06). The third model is induced from the same document collection without first performing anaphora resolution on it.

During the evaluation, the document collection described in Section 4 was randomly split into ten parts, one part to be used for testing and the rest for training of the model. Each model was evaluated in terms of F-measure, averaged over ten such runs. Figure 3 describes the results of the experiments.

The results show that the use of either version of MARS consistently yields improved classification effectiveness in comparison with the baseline model. MARS02 achieved the best results on two classification methods (kNN and Rocchio), while MARS06 was the best on the other three (NB, MaxEnt, and SVM).

[11] We used the implementations of these methods distributed with the Rainbow text classification toolkit [19].

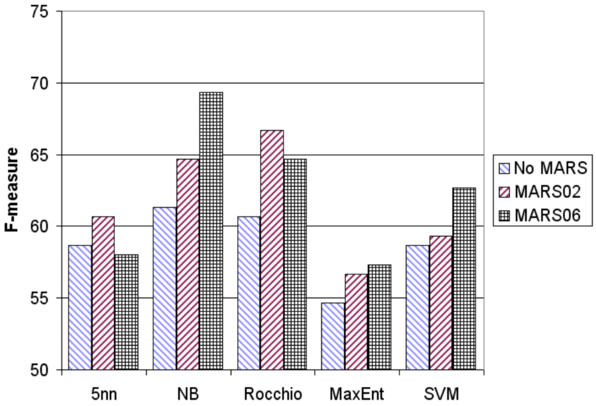

Fig. 3. The effect of anaphora resolution on the accuracy of text categorisation

Although consistent, the improvement on the baseline is not considerable (max. 8% with MARS06, using NB); in none of the compared pairs could statistical significance according to an independent sample t-test be established to the .05 level.

The experiment did not show any considerable differences in the performance of MARS02 and MARS06. MARS06 showed the greatest improvement of 4.66% (on NB), but was worse than MARS02 by 2.66% (on kNN). None of these differences is statistically significant according to the independent sample t-test.

6 Discussion

We aimed to establish the extent to which a task such as anaphora resolution could be useful in other NLP applications. From the results reported above, it is obvious that deployment of MARS has different impacts on the performance of each of the applications experimented with: text summarisation, term extraction and text classification. By and large the deployment of MARS has a positive but at the same time limited impact. In summarisation the deployment of MARS on the evaluation data results in improvement of the F-measure, although this is not significant. Term extraction also benefits from incorporation of MARS as a preprocessing module, but again, the improvement is not statistically significant. In both applications there are different cases where each version of MARS has a more favourable impact. In text categorisation, MARS provides a statistically significant improvement to one of the classification methods (kNN), and statistically insignificant improvement to two other classification methods. However, with respect to the MaxEnt method, performance deteriorates when MARS is employed.

One observation worth making on the basis of the experiments conducted is that the slight improvement in performance of MARS06 as opposed to MARS02 does not necessarily result in improvement of the performance of the application in which it is employed. It would be interesting to see whether a dramatic improvement in performance of the resolution of anaphors would lead to a marked improvement of the NLP application that exploits it.

To this end, our next step is to establish whether there is any threshold to be achieved in order for anaphora resolution to be considered beneficial in that it almost always enhances the performance of the above applications, possibly bringing a statistically significant improvement. The results of preliminary studies in the area of text summarisation applied to a collection of scientific texts have already been established [20].

7 Conclusion

This paper covers for the first time, to the best of our knowledge, the issue of extrinsic evaluation in the context of anaphora resolution for more than one NLP application. In particular, we explore the extent to which our well-known anaphora resolution system, MARS, can improve the performance of three NLP applications (text summarisation, term extraction and text categorisation). On the basis of the results so far we conclude that the deployment of anaphora resolution has a positive albeit limited impact.

References

1. Mitkov, R., Evans, R., Orasan, C.: A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In: Proceedings of CICLing-2002, Mexico City, Mexico (February 2002) 168 – 186
2. Mitkov, R.: Anaphora resolution. Longman (2002)
3. Lappin, S., Leass, H.J.: An algorithm for pronominal anaphora resolution. Computational Linguistics **20**(4) (1994) 535 – 562
4. Mitkov, R.: Pronoun resolution: the practical alternative. In: Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium (DAARC), Lancaster, UK (1996)
5. Mitkov, R.: Robust pronoun resolution with limited knowledge. In: Proceedings of the 18th International Conference on Computational Linguistics (COLING'98/ACL'98), Montreal, Quebec, Canada (August 10 - 14 1998) 867 – 875
6. Tapanainen, P., Järvinen, T.: A non-projective dependency parser. In: Proceedings of the 5th Conference of Applied Natural Language Processing, Washington D.C., USA (March 31 - April 3 1997) 64 – 71
7. Evans, R.: Applying machine learning toward an automatic classification of *It*. Literary and Linguistic Computing **16**(1) (2001) 45 – 57
8. Orăsan, C., Evans, R.: Learning to identify animate references. In Daelemans, W., Zajac, R., eds.: Proceedings of CoNLL-2001, Toulouse, France (July, 6 – 7 2001) 129–136
9. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: Proceedings of ACL02. (2002)
10. Muñoz, R., Saiz-Noeda, M., Montoyo, A.: Semantic information in anaphora resolution. In: Proceedings of PorTAL 2002. (2002) 63 – 70

11. Orăsan, C.: PALinkA: a highly customizable tool for discourse annotation. In: Proceedings of the 4th SIGdial Workshop on Discourse and Dialog, Sapporo, Japan (July, 5 -6 2003) 39 – 43
12. Hasler, L., Orăsan, C., Naumann, K.: NPs for Events: Experiments in Coreference Annotation. In: Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation (LREC2006), Genoa, Italy (24 – 26 May 2006) 1167 – 1172
13. Vilain, M., Burger, J., Aberdeen, J., Connoly, D., Hirschman, L.: A model-theoretic coreference scoring scheme. In: Proceedings of the 6th Message Understanding Conference (MUC-6), San Francisco, California, USA (1995) 45–52
14. Hasler, L., Orăsan, C., Mitkov, R.: Building better corpora for summarisation. In: Proceedings of Corpus Linguistics 2003, Lancaster, UK (March, 28 – 31 2003) 309 – 319
15. Mitkov, R., Hallett, C.: Comparing pronoun resolution algorithms. Journal of Computational Intelligence (forthcoming)
16. Orăsan, C., Pekar, V., Hasler, L.: A comparison of summarisation methods based on term specificity estimation. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC2004), Lisbon, Portugal (May 26 – 28 2004) 1037 – 1041
17. Justeson, J.S., Katz, S.L.: Technical terminology: some linguistic properties and an algorithm for identification in text. Journal of Natural Language Engineering **3**(2) (1996) 259–289
18. Hulth, A.: Reducing false positives by expert combination in automatic keyword indexing. In: Proceedings of RANLP 2003, Borovetz, Bulgaria (September 2003) 197–203
19. McCallum, A.K.: Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow (1996)
20. Orăsan, C.: Comparative evaluation of modular automatic summarisation systems using CAST. PhD thesis, University of Wolverhampton (2006)