Refined Salience Weighting and Error Analysis in Anaphora Resolution

Richard Evans
School of Humanities, Languages, and Social Sciences
University of Wolverhampton
r.j.evans@wlv.ac.uk

Abstract

In this paper, the behaviour of an existing pronominal anaphora resolution system is modified so that different types of pronoun are treated in different ways. Weights are derived using a genetic algorithm for the outcomes of tests applied by this branching algorithm. Detailed evaluation and error analysis is undertaken. Proposals for future research are put forward.

1 Introduction

In this paper we propose modifications to the behaviour of MARS, a pronominal anaphora resolution system that processes texts in English. Based on the algorithm presented in (Mitkov, 1998), the most comprehensive description of the implemented system appears in (Mitkov et al., 2002).

The algorithm used by MARS is presented in Section 2. The system tests noun phrase (NP) candidates preceding a pronoun in a text and applies salience weights to them on the basis of the outcomes of these tests. The same set of tests is applied in each case, and the same salience weights are applied by them regardless of the number of potential antecedents under consideration and the form of the pronoun being processed.

In its original statement, the weights applied by the tests were selected by hand, on an intuitive basis, reflecting observations taken from a small corpus. Initially, it was assumed that all the tests (referred to as indicators) could be implemented perfectly, and NP gender and number information could be obtained accurately to allow the implementation of agreement constraints. In practical systems, this is not necessarily the case and it may be that the originally proposed weights are not appropriate for a practical system. In this paper, a new corpus-based weight selection procedure is tested, a procedure based on the application of a genetic algorithm (GA) to an algorithm that varies in its behaviour across different types of anaphoric pronoun. The typology for pronouns used in this paper is based on their morphological form and the cardinality of the sets of their potential antecedents, information that is easily obtained.

Rather than constrain the anaphora resolution algorithm to operate in a uniform fashion for all pronouns, we hypothesise that performance will improve if we allow it to behave differently when resolving different types of pronoun. The possibility that the importance and weighting of indicators varies according to the type of pronoun being resolved is thus accommodated by the proposed system.

In this paper, Section 2 presents the algorithm followed by MARS in detail. Section 3 makes the proposal for refined salience weighting with respect to MARS's indicators and describes the method used in order to find the most effective set of weights to be applied by the anaphora resolution system. This section also contains evaluation data and discusses the findings of this undertaking. Section 4 presents error analysis. In Section 5 related work is discussed and in Section 6, conclusions are drawn and directions for future research are proposed.

2 Anaphora Resolution Algorithm Followed by MARS

For every pronoun in the text that belongs to the set {*he*, *her*, *him his*, *it*, *its*, *she*, *their*, *them*, *they*}:

1. Classify the pronoun
2. For those pronouns ascertained by the classification process to demonstrate nominal anaphora:
   (a) Extract NPs prior to the pronoun from the sentence in which the pronoun appears, from the two sentences prior to this, and from the most recent section heading. These extracted NPs form a set of candidates for the antecedent of the pronoun in question.
   (b) Ascertain for each candidate whether or not it agrees with the pronoun in terms of number, gender, and syntactic constraints. NPs that agree with the pronoun form the set of competing candidates that is expected to contain the antecedent of the pronoun.
   (c) For the set of competing candidates:
      - A set of 14 tests for salience (indicators) are applied to the NPs in the set. Each NP evokes a single outcome from each indicator and the algorithm assigns it the relevant salience weight associated with that outcome. To illustrate, one indicator assumes that definite NPs tend to be more salient than indefinite ones. Thus, when testing a NP with this indicator, one of two outcomes are evoked, one for the case in which the NP is definite and one for the case in which the NP is indefinite. Once an outcome has been evoked, the NP is assigned a salience weight. In its original form, the algorithm assigns indefinite NPs a negative weight of -1, whereas definite NPs are assigned a neutral weight of 0. Full details of the entire set of 14 indicators can be found by study of (Mitkov, 1998) and (Mitkov et al., 2002).

   (d) Select the one most highly weighted NP in the sets of competing candidates as the antecedent of each nominally anaphoric pronoun. Where numerous candidates have the highest weight, the most recent one is selected as the antecedent.

This algorithm is used as the basis of the branching algorithm described in Section 3. As is noted in Section 4, steps 1 and 2b are not performed with 100% accuracy by MARS .

3 Refined Salience Weighting

As with many other systems for anaphora resolution, the algorithm described in Section 2 operates in a similar fashion over all the pronouns that it is presented with. For this reason, it will hereafter be referred to as the *uniform algorithm*. In step 2c, the invocation of an indicator outcome leads to the application of the same salience weight to candidate NPs, regardless of the type of pronoun being resolved or the numbers of competing candidates under consideration as antecedent of that pronoun. In the original work, (Mitkov, 1998), these weights were decided in accordance with corpus observation and certain tenets of centering theory on an intuitive manual basis.

It is hypothesised that the success rate of the implemented system can be increased by having the algorithm branch and execute slightly different behaviour according to a pronoun typology derived from form[1] and the cardinality of the automatically generated sets of competing candidates. Differences in MARS's behaviour will be implemented by applying different weights for particular indicator outcomes for different types of pronoun. For this reason, the algorithm presented in this work will be referred to as the *branching algorithm*. The typology results in classes of pronouns such as "*it* with 3 candidates," "*it* with 5 candidates," or "*she* with 9 candidates" that can be treated independently.

3.1 Search Using a Genetic Algorithm

It has been shown that genetic algorithms can be used to automatically obtain less intuitive

[1] Here, different pronouns such as *it*, *he*, or *she* are regarded as different forms.

but more optimal weights from annotated data (Mitkov et al., 2002). Genetic algorithms (GA) are a well known parameter optimisation method (Holland, 1975). In the case of anaphora resolution, the algorithm is regarded as a function which returns a level of fitness in the evaluation phase. Salience weights are regarded as the parameters of this function. The goal is to search for salience weights which, when applied, lead to the maximisation of fitness. This search is performed by generating sets of parameters (salience weights), evaluating their application, and then adjusting the most successful of them and repeating this process for a specified number of generations. Here, the GA first generates a population of sets of random salience weights. MARS applies these weights and its success rate is obtained. In each case, success rate represents the fitness of a set of weights. Roulette wheel selection, biased by fitness, sends a proportion of sets for further processing. Crossover and mutation operators are applied to selected sets in order to generate a new population. The algorithm iterates for 300 generations.

GA were used to improve the performance of the uniform algorithm in the work described in (Orăsan et al., 2000). It was found that this application of GA leads to performance gains when applied over individual texts, but cross-evaluation results were poor (Mitkov et al., 2002). Overall, the originally proposed salience weights outperformed the GA-derived ones. Further, the application of the GA to the uniform algorithm is not effective when different texts are combined. No improvement results from this type of search when processing combinations of multiple texts.

In this paper, a GA is used to search for salience weights that lead to increased performance when the branching algorithm is used to process a corpus consisting of different texts (see Section 3.2). Here, the algorithm is broken into different parts, each one dealing with a different type of pronoun. The GA searches for optimal weights in each of the different parts. The GA operates in the same way for *each branch* of the algorithm as it did for the uniform algorithm described earlier.

It may be found, for example, that the best score given to a candidate of “it with 5 candidates” for one outcome of *definiteness* is x whereas for candidates of “it with 9 candidates” it is y. Section 3.3 shows that search of this type does yield salience weights that lead to increased performance.

3.2 The Data Set

The corpus used here consists of 9 texts, 8 from the domain of computer hardware and software technical manuals (247,401 words), and one text consisting of material from the annual report of human rights organisation *Amnesty International* (15,767 words). Examination of the *Num* rows in Table 1 reveals the distribution of pronouns by type in this corpus. MARS is designed to resolve nominal anaphoric pronouns of the forms listed in the table. 2684 of the pronouns that appear here demonstrate nominal anaphora.

3.3 Findings

Table 1 includes an evaluation of the anaphora resolution system. The column *Form* lists all the different forms of pronoun processed by both the uniform and branching instantiations of MARS. The major column *#Candidates* provides information on the pronouns, arranged by virtue of the numbers of candidates that the algorithms must select an antecedent from. In each case, the row *Num* shows the number of such pronouns present in the corpus, the row *Base* shows how many of those pronouns MARS is able to resolve correctly when the original uniform weighting scheme presented in (Mitkov et al., 2002) is applied, the row *Branch* shows the number of pronouns resolved successfully in each case when the salience weights used by MARS are those derived by the GA applied under the branching algorithm. The row *Total* and the column *All* sum the numbers presented in the previously described columns and rows.

Conflating the classes of pronouns with 0 and 1 competing candidates hides the fact that 169 (5.96%) of the pronouns have no competing candidates for selection by the algorithm. Further, and more significantly as discussed in Section 4.3, 592 (20.87%) of the pronouns have no valid antecedents present in their sets of competing candidates. 152 of these pronouns do not demonstrate nominal anaphora. These instances are presented to the system as a result of pronoun classification

Form		#Candidates												
		0-1	2	3	4	5	6	7	8	9	10	11-15	16+	All
He	Num	9	0	3	6	7	11	8	8	2	7	27	7	95
	Base	3	0	0	3	4	7	5	4	0	5	16	4	51
	Branch	3	0	1	5	5	10	6	6	1	6	20	4	68
Her	Num	2	4	3	0	4	3	4	5	3	3	2	0	33
	Base	0	2	0	0	1	0	0	2	0	0	1	0	6
	Branch	0	3	0	0	2	2	1	3	2	1	1	0	15
Him	Num	1	0	0	0	0	3	2	3	3	4	4	0	20
	Base	1	0	0	0	0	0	0	2	1	2	3	0	9
	Branch	1	0	0	0	0	1	0	2	2	3	3	0	12
His	Num	6	3	5	7	4	5	6	8	12	4	23	0	83
	Base	0	2	4	4	3	4	4	5	7	1	12	0	46
	Branch	0	2	4	6	3	4	4	7	11	3	19	0	63
It	Num	226	192	184	172	169	145	133	96	87	75	177	57	1713
	Base	91	126	112	101	77	74	73	50	48	41	91	22	906
	Branch	91	128	117	103	86	77	74	63	56	48	108	36	987
Its	Num	54	27	19	15	13	15	11	25	9	6	24	9	227
	Base	26	22	9	8	7	8	7	15	4	2	18	3	129
	Branch	26	24	13	12	10	10	8	19	7	5	19	4	157
She	Num	0	0	0	0	0	0	0	0	2	0	0	0	2
	Base	0	0	0	0	0	0	0	0	0	0	0	0	0
	Branch	0	0	0	0	0	0	0	0	1	0	0	0	1
Their	Num	47	18	17	14	12	9	8	5	5	4	3	0	142
	Base	22	13	12	6	7	7	4	1	2	1	1	0	76
	Branch	22	14	13	10	10	8	6	4	3	2	1	0	93
Them	Num	74	52	44	35	19	13	9	7	3	0	0	0	256
	Base	45	36	28	21	11	5	6	4	1	0	0	0	157
	Branch	45	40	34	27	15	10	8	6	2	0	0	0	187
They	Num	80	67	39	26	23	14	7	0	4	2	3	0	265
	Base	51	46	26	17	17	6	6	0	2	0	1	0	172
	Branch	51	50	32	19	20	10	6	0	3	1	1	0	193
Total	Num	499	363	314	275	251	218	188	157	130	105	263	73	2836
	Base	239	247	191	160	127	111	105	83	65	52	143	29	1552
	Branch	239	261	214	182	151	132	113	110	88	69	172	44	1708

Table 1: Evaluation of the uniform algorithm with original salience weights (*Base*) and the branching algorithm with GA-derived salience weights (*Branch*)

and preprocessing errors.

As mentioned in Section 3.2, the corpus contains 2684 pronouns that demonstrate nominal anaphora. Examination of the intersection of *Total* and *All* in Table 1 reveals that overall, MARS's success rate[2] improves from 57.82% (1552/2684) to 63.64% (1708/2684) when the GA-derived weights under the branching algorithm are used.

Experiments conducted so far showed that the application of the GA to the uniform algorithm when processing this combination of multiple texts did not lead to any improvement. Indeed, the level of performance obtained by the original statement of the algorithm was not reached in these experiments.

[2]Success rate is defined as the ratio of the number of nominal anaphoric pronouns resolved successfully to the number of nominal anaphoric pronouns in the corpus.

In one preliminary experiment, all instances of the pronoun *it* (regardless of the cardinality of their sets of competing candidates) were extracted from the data set and divided into a training and testing portion. Using the GA to derive optimal salience weights from the training portion and applying them when processing the testing portion did not lead to an improvement in performance. This shows that refined weighting can give higher accuracy over particular annotated texts, but not over all texts in general.

When applied, the GA normally generates a large number of sets of salience weights that share best fitness over the training data. So far, only a handful of them have been tried over the test data. Perhaps, in time, different sets will be found that lead to improved performance over both training and testing data. However, for practical purposes this would not be so useful because then an automatic method to separate such solutions out from the rest would have to be developed.

4 Error Analysis and Discussion

In this Section, a detailed examination of the shortcomings of the implemented system is undertaken and findings are discussed.

4.1 Agreement Constraints

One significant cause of errors is the unreliability of the system's method for enforcing gender and number agreement constraints between pronouns and antecedents. Section 4.3 shows the extent of the problem. The difficulties in applying such constraints are discussed in (Barlow, 1998). A survey carried out in (Barbu et al., 2002 Forthcoming) presents a typology of difficulties in enforcing agreement with plural pronouns. Work presented in (Orăsan and Evans, 2001b) sets out a method for determining the animacy of common NPs, whereas (Mikheev et al., 1998) tackles named entity (including person name) recognition. Gender agreement is applied very loosely by MARS. Person names appearing in a proper name list are assigned the relevant gender whilst all other NPs are allowed to agree with any type of pronoun until they are resolved to a particular one. At that point, their gender is set to match that of the pronoun to which they were resolved. Number agreement is enforced more strictly by MARS because this information is normally available with a fair degree of accuracy from the syntactic parser used by the system (Tapanainen and Järvinen, 1997).

4.2 Pronoun Classification

Many of the pronouns being processed by MARS do not actually demonstrate nominal anaphora. This occurs in spite of MARS's use of the classification method for the pronoun *it* proposed in (Evans, 2001). This classification system has not been evaluated over all the texts present in the corpus used in this work, but it obtained an accuracy of 85.54% over the 8 computer technical manuals used in this corpus. It is unable to classify any other form of pronoun than *it*. The method is quite accurate in classifying nominal anaphoric and pleonastic instances but is poor at classifying other types of non-nominal-anaphoric pronouns. Problematic cases are discussed in Section 4.3.

Class	Proportion
Parse error	36.15%
Distant antecedent	15.71%
Discourse topic	15.20%
Annotation error	11.49%
Pleonastic	10.64%
Plural disagreement	6.42%
Idiomatic	6.42%
Proaction	4.90%
Singular disagreement	3.72%
Inferred antecedent	2.70%
Cataphoric	2.36%
Discontinuous antecedent	1.52%
Gender disagreement	00.67%
Verb anaphora	00.34%
Text formatting	00.17%
Ambiguous (Multiple Classes)	18.41%

Table 2: Classes of pronouns for which no antecedents can be chosen.

4.3 Missing Antecedents

It has been noted that 592 pronouns have no antecedents available for selection from their sets of competing candidates. The reasons for this have been investigated and Table 2 presents statistics obtained regarding the proportions of this number that are accounted for by different phenomena.

1. *Parse error* refers to those cases in which a failure in syntactic analysis means that MARS is unable to extract a valid antecedent and incorporate it into the set of competing candidates.

2. *Distant antecedent* refers to those pronouns whose antecedents are located in a different paragraph to the pronoun, or else more distant than 2 sentences in the text prior to the pronoun.

3. *Discourse Topic* refers to pronouns whose interpretation relies upon non-trivial interpretation of the preceding discourse. Such cases may involve the description of some activity followed by a construction such as "*It*'s an easy way to..."

4. *Annotation error* refers to those cases in which an annotator has erroneously ne-

glected to mark any correct antecedent for a pronoun even though they are available.

5. *Pleonastic pronouns* do not demonstrate nominal anaphora and appear in such sentences as "*it*'s impossible to know what will happen next".

6. *Plural disagreement* refers to those plural pronouns whose antecedents do not have plural nouns as heads and are identified as singular at the syntactic analysis stage. This can occur when a singular company name such as *Alphavision*, generic NPs such as *any person* or quantified NPs such as *one range of square roots* is later referred to using *they*.

7. Neither *Idiomatic* nor *Proaction* pronouns demonstrate nominal anaphora. The former are those pronouns used in idioms and stereotypic constructions in the language such as "when *it* comes to..." whereas the latter combine with *do* to point to previously mentioned activities, rather than discourse referents.

8. *Singular disagreement* refers to pronouns whose antecedents do not have singular noun heads and are identified as plural at the syntactic analysis stage. Such instances can occur when the name of a program such as *ears*, or *Microeyes* is referred to.

9. *Inferred antecedent* refers to pronouns that take their interpretation from non-explicit entities in the text. The end-products of some process may not appear explicitly in the text, but may later be referred to pronominally. An example is illustrated in "when you have finished typing, press save, and store *it* in a suitable location" in which some document understood to have arisen from the typing process is referred to using *it*.

10. The antecedents of *Cataphoric* pronouns cannot be located by an algorithm for anaphora resolution. The cases reported here are of genuine cataphora in which the initial mention of the entity is made using a pronoun.

11. *Discontinuous antecedent* refers to pronouns whose antecedents are available in the text, but they are built up from numerous smaller NPs that must be extracted and conjoined.

12. *Gender disagreement* refers to cases in which a pronoun of one gender is resolved to a particular NP, and the anaphora resolver then sets the gender of that NP to match the pronoun. Later, a pronoun of a different gender points back at the same NP but is incompatible with it. This occurs in the example "the user of the machine should recall *his* or *her* own settings."

13. *Verb anaphora* refers to those pronouns whose antecedents are verb phrases and by definition do not demonstrate nominal anaphora. The presentation of these pronouns, as well as those described in items 3, 5, 7, 9, and 10 of this section, to the resolution system results from errors in Step 1 of the algorithm.

14. *Text formatting* refers to those pronouns for which no antecedent is available due to erroneous editing and deletion in the source text.

15. *Ambiguous* pronouns are those that may belong to more than one of these classes.

It can be concluded from this investigation that almost one third of the pronouns whose competing candidates contain no antecedents can be resolved by an anaphora resolution system. They remain beyond the scope of MARS's current instantiation but correction of annotation errors, modification of the enforcement of agreement constraints, and extension of the algorithm's search scope are all initial steps that can be taken in order to improve system performance. Further development of pronoun classification software would also help to reduce the number of non-nominal-anaphoric pronouns that the system is confronted with.

4.4 Discussion: Sparse Data

The typology used by the branching algorithm results in classes that are too small to allow effective training and testing over each of its branches.

For example, there are just 2 examples in the corpus of the pronoun *Him* with 7 candidates. If 100 pronouns is viewed as a minimum size for a training sample and 100 for testing, this means that a corpus 100 times the size of the current one (containing almost 283,600 pronouns for the purposes of this hypothetical example) may need to be annotated. Such resources are simply not available at present.

If pronouns are simply considered by type, the data set does contain enough examples of some pronouns to allow more conclusive training and testing. For instance, as described in Section 3.3 there are enough examples of *it* to warrant training and testing using the GA. However, as was found in (Mitkov et al., 2002), cross-evaluation shows that performance on the testing data is worse than that when the default weights are used. It must be concluded firstly that not enough representative data is available. Unfortunately, the more data that is available, the less likely it is that the GA will find values to obtain increased performance.

5 Related Work

Tests for the salience of candidate NPs in anaphora and coreference resolution have been used successfully in numerous works, including (Mitkov, 1998), (Lappin and Leass, 1994), the system based on Lappin and Leass's algorithm presented in (Kennedy and Boguraev, 1996), (Palomar et al., 2000) and (Nasukawa, 1994). It must be said that the precise details of the ways in which these tests are used to select an antecedent vary from case to case.

Genetic algorithms have been applied to methods that use salience tests, but the success of such applications has been mixed (Byron and Allen, 1999), (Orăsan et al., 2000). A pilot study, presented in (Orăsan and Evans, 2001a), suggested that GAs are more effective in optimising MARS than neural network or memory based learning methods.

Alternative ways to use salience tests in anaphora and coreference resolution have been employed in related work. (Cardie and Wagstaff, 1999) used weighted features to compute a distance metric between pairs of potentially coreferential NPs in a coreference resolution system. (Denber, 1998) presents an algorithm in which antecedents are selected for pronouns by applying a set of salience tests and checking for which NP the greatest number of tests are positive. The system presented by (Ge et al., 1998), uses empirical corpus-based methods to assess the likelihood that a NP found manifesting a particular salience property (such as subjecthood) or found as the argument of a particular verb, is the antecedent of a given pronoun.

(Harabagiu and Maiorano, 2000) present an automatic statistical corpus-based method for testing heuristics before deciding which ones should be incorporated into an anaphora resolution system.

6 Conclusions

For anaphora resolution algorithms such as those presented in (Mitkov, 1998) and (Kennedy and Boguraev, 1996), in which salience tests assign numerical weights to candidate NPs, it is clear that performance can be improved over particular annotated corpora by use of GA. It is encouraging to find that the GA did obtain weights that led to improved performance by MARS over the combination of multiple texts present in the corpus used here. However, the study, supported by work carried out in (Mitkov et al., 2002) suggests that the weights derived by the GA over-fit the corpora from which they were derived, limiting their general applicability.

The error analysis carried out in this paper revealed that further progress is possible with the system presented here, but initially, developments in areas such as the accurate enforcement of number and gender agreement constraints, pronoun classification, the search scope of the algorithm, and validation of annotated data will be more beneficial than developments in the process of selecting antecedents from sets of competing candidates.

It is also clear that there is an urgent need for more annotated data in this field. The fact that in order to obtain useful quantities of each of the types of pronoun referred to in this work, the corpus used here must be extended by perhaps 100 times, demonstrates the true extent of this demand.

It has been noted that although a number of sets of salience weights derived by the GA from training data has not led to improved performance over testing data, this does not mean that none of these sets may lead to improvements. In future, it will be interesting to investigate selection methods for GA solutions in order to produce generally useful results.

References

Cătălina Barbu, Richard Evans, and Ruslan Mitkov. 2002 Forthcoming. A corpus based investigation of morphological disagreement in anaphoric relations. In *Proceedings of LREC-2002*, Las Palmas de Gran Canaria, Spain, May.

Michael Barlow. 1998. Feature mismatches in anaphora resolution. In Paul Baker, Andrew Hardie, Tony McEnery, and Anna Siewierska, editors, *Proceedings of Discourse Anaphora and Reference Resolution Conference (DAARC-98)*, pages 34 – 41, Lancaster, UK. Lancaster University.

Donna Byron and James Allen. 1999. Applying genetic algorithms to pronoun resolution. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*.

Claire Cardie and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. In *Proceedings of the 1999 conference on Empirical Methods in NLP and Very Large Corpora (ACL'99)*, pages 82 – 89, Maryland, USA, 20 June - 26 June.

Michel Denber. 1998. Automatic resolution of anaphora in English. Technical report, Eastman Kodak Co., Imaging Science Division.

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

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora (COLING-ACL'98)*, pages 161 – 170, Montreal, Canada, August.

Sanda Harabagiu and Steven Maiorano. 2000. Multilingual coreference resolution. In *Proceedings of ANLP-NAACL 2000*, May.

John H. Holland. 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press.

Christopher Kennedy and Branimir Boguraev. 1996. Anaphora for everyone: pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 113 – 118, Copenhagen, Denmark.

Shalom Lappin and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535 – 561.

Andrei Mikheev, Claire Grover, and Marc Moens. 1998. Description of the LTG system used for MUC-7. In *Proceedings of MUC-7*, Fairfax, USA. TIPSTER.

Ruslan Mitkov, Richard Evans, and Constantin Orăsan. 2002. A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In *Proceedings of CICLing 2002*, pages 168 – 186, Mexico City, Mexico, February 17 - 23. Springer.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98/ACL'98)*, Montreal, Canada.

Tetsuya Nasukawa. 1994. Robust method of pronoun resolution usinf full-text information. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, Kyoto, Japan, August.

Constantin Orăsan and Richard Evans, 2001a. *Enterprise Information Systems II*, pages 69 – 73. Kluwer Academic Publishers.

Constantin Orăsan and Richard Evans. 2001b. Learning to identify animate references. In *Proceedings of the Workshop on Computational Natural Language Learning (CoNLL-2001)*, pages 129 – 136, Toulouse, France, July 6 - 7.

Constantin Orăsan, Richard Evans, and Ruslan Mitkov. 2000. Enhancing preference-based anaphora resolution with genetic algorithms. In Dimitris N. Christodoulakis, editor, *Natural Language Processing - NLP 2000*, number 1835 in Lecture Notes in Artificial Intelligence, pages 185 – 195, Patras, Greece, June. Springer.

Manuel Palomar, Maximiliano Saiz-Noeda, Rafael Munoz, Armando Suarez Cueto, and Patricio Martinez-Barco. 2000. Phora: A system to solve the anaphora in spanish. In *Proceedings of DAARC-2000*, pages 206 – 211, Lancaster, UK., November.

Pasi Tapanainen and Timo Järvinen. 1997. Improving anaphora resolution by identifying animate entities in texts. In *Proceedings of the 5th Conference of Applied Natural Language Processing (ANLP'97)*, pages 64 – 71, Washington D.C., US, 31 March – 3 April.