A New, Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method

Ruslan Mitkov, Richard Evans, and Constantin Orasan

School of Humanities, Languages and Social Sciences,
University of Wolverhampton,
Stafford Street,
Wolverhampton,
WV1 1SB.
UK.
{r.mitkov, r.j.evans, c.orasan}@wlv.ac.uk
http://www.wlv.ac.uk/sles/compling/

Abstract. This paper describes a new, advanced and completely revamped version of Mitkov's knowledge-poor approach to pronoun resolution [21]. In contrast to most anaphora resolution approaches, the new system, referred to as MARS, operates in fully automatic mode. It benefits from purpose-built programs for identifying occurrences of non-nominal anaphora (including pleonastic pronouns) and for recognition of animacy, and employs genetic algorithms to achieve optimal performance. The paper features extensive evaluation and discusses important evaluation issues in anaphora resolution.

1 The original approach

Mitkov's approach to anaphora resolution [21] avoids complex syntactic, semantic and discourse analysis relying on a list of preferences known as antecedent indicators. The approach operates as follows: it works on texts first processed by a part-of-speech tagger and a noun phrase (NP) extractor, locates NPs which precede the anaphor within a distance of 2 sentences, checks them for gender and number agreement with the anaphor and then applies indicators to the remaining candidates that assign positive or negative scores to them (-1, 0, 1 or 2). The NP with the highest composite score is proposed as antecedent[1].

The antecedent indicators[2] can act either in a boosting or 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

[1] The approach only handles pronominal anaphors whose antecedents are NPs.

[2] The original indicators are named *First NPs* (FNP), *Indefinite NPs* (INDEF), *Indicating Verbs* (IV), *Lexical Reiteration* (REI), *Section Heading Preference* (SH), *Collocation Match* (CM), *Prepositional Noun Phrases* (PNP), *Immediate Reference* (IR), *Sequential Instructions* (SI), *Referential Distance* (RD), and *Term Preference* (TP).

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) or to structural matches, whereas others are genre-specific[3]. For a complete and detailed description see [21]. As an illustration, the indicator, *Immediate Reference* (IR) acts in a genre-specific manner and predicts that an NP appearing in a construction of the form "...(You) V_1 NP ... *con* (you) V_2 it (*con* (you) V_3 it)", where *con* ϵ {*and*/*or*/*before*/*after*...} will be the antecedent of a given pronoun. This preference is highly genre-specific and occurs frequently in imperative constructions such as "To turn on the printer, press *the Power button* and hold *it* down for a moment" or "Unwrap *the paper*, form *it* and align *it*, then load *it* into the drawer." This indicator, together with *collocation match* and *prepositional noun phrases* was most successful in pointing to the correct antecedent[4] of a given pronoun. In fact, initial results showed that the NP awarded a score by *immediate reference* always emerged as the correct antecedent.

The evaluation of Mitkov's knowledge-poor approach which was carried out by running the algorithm on post-edited outputs from the POS tagger and NP extractor, showed a success rate of 89.7% on a collection of texts, including the user guide referred to in Section 3 as PSW.

2 MARS: a re-implemented and improved fully automatic version

Our project addresses the most crucial type of anaphora to NLP applications - that of identity-of-reference nominal anaphora, which can be regarded as the class of single-document identity coreference. This most frequently occurring class of anaphora has been researched and covered most extensively, and is the best understood within the field[5]. The current implementation of MARS is limited to pronoun resolution.

2.1 Fully automatic anaphora resolution

MARS is a re-implemented version of Mitkov's robust, knowledge-poor approach which uses the FDG-parser [30] as its main pre-processing tool. MARS operates

[3] Typical of the genre of user guides.

[4] The confidence is computed in terms of decision power, which is a measure of the influence of each 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 values partially served as a guide in proposing the numerical scores for each indicator. For a definition of this measure see [22].

[5] *Nominal anaphora* arises when a referring expression - pronoun, definite noun phrase, or proper name, has a non-pronominal noun phrase as antecedent. MARS does not handle identity-of-sense anaphora where the anaphor and the antecedent do not correspond to the same referent in the real world but to ones of a similar description as in the example "The man$_i$ who gave his$_i$ paycheck$_j$ to his$_i$ wife was wiser than the man$_k$ who gave it$_j$ to his$_k$ mistress."

in a fully automatic mode, in contrast to the vast majority of approaches which rely on some kind of pre-editing of the text which is fed to the anaphora resolution algorithm[6] or which have only been manually simulated. As an illustration, Hobbs's naïve approach [17] was not implemented in its original version. In [7], [8], [1], and [19] pleonastic pronouns are removed manually[7], whereas in [21] and [12] the outputs of the PoS tagger and the NP extractor/partial parser are post-edited in a similar way to [20] where the output of the Slot Unification Grammar parser is corrected manually. Finally, [13] and [31] make use of annotated corpora and thus those approaches do not perform any pre-processing.

The development of MARS and also the re-implementation of fully automatic versions of Baldwin's as well as Kennedy and Boguraev's approaches for comparative purposes in another project [2], showed that fully automatic anaphora resolution is more difficult than previous work has suggested[8]. In the real-world fully automatic resolution must deal with a number of hard pre-processing problems such as morphological analysis/POS tagging, named entity recognition, NP gender identification, unknown word recognition, NP extraction, parsing, identification of pleonastic pronouns, selectional constraints, etc. Each one of these tasks introduces error and thus contributes to a reduction of the success rate of the anaphora resolution system; the accuracy of tasks such as robust parsing and identification of pleonastic pronouns is way below 100%[9]. For instance, many errors will be caused by the failure of systems to recognise pleonastic pronouns - and their consequent attempt to resolve them as anaphors.

2.2 Differences between MARS and the original approach

The initial implementation of MARS followed Mitkov's original approach more closely, the main differences being (i) the addition of three new indicators and (ii) a change in the way some of the indicators are implemented or computed due to the available pre-processing tools. In its most recent version, however, MARS uses a program for automatically recognising instances of non-nominal

[6] This statement refers to anaphora resolution systems and not to the coreference resolution systems implemented for MUC-6 and MUC-7.

[7] In addition, [8] undertook additional pre-editing such as removing sentences for which the parser failed to produce a reasonable parse, cases where the antecedent was not an NP etc.; [19] manually removed 30 occurrences of pleonastic pronouns (which could not be recognised by their pleonastic recogniser) as well as 6 occurrences of *it* which referred to a VP or prepositional constituent.

[8] By fully automatic anaphora resolution we mean that there is no human intervention at any stage: such intervention is sometimes large-scale such as manual simulation of the approach and sometimes smaller-scale as in the cases where the evaluation samples are stripped of pleonastic pronouns or anaphors referring to constituents other than NPs.

[9] The best accuracy reported in robust parsing of unrestricted texts is around the 86% mark; the accuracy of identification of non-nominal pronouns is under the 80% mark though [27] reported 92% for identification of pleonastic *it*.

pronominal anaphors and pleonastic pronouns[10], it incorporates two new syntax filters, and a program for automatic gender identification. Each of these new components is described in sections 2.2.1-2.2.4 below.

2.2.1 New Indicators

The three new indicators that were included in MARS are:

Boost Pronoun (BP): As NPs, pronouns are permitted to enter the sets of competing candidates for other pronouns. The motivation for considering pronominal candidates is twofold. Firstly, pronominalised forms represent additional mentions of entities and therefore increase their topicality. Secondly, the NP corresponding to an antecedent may be beyond the range of the algorithm, explicitly appearing only prior to the two sentences preceding the one in which the pronoun appears. Pronoun candidates may thus serve as a stepping-stone between the current pronoun and its more distant nominal antecedent. Of course, it is not helpful in any application for the system to report that the antecedent of a pronoun *it* is another pronoun *it*. When a pronoun is selected as the antecedent, the system has access to that pronoun's own antecedent in a fully transitive fashion so that a NP is always returned as the antecedent of a pronoun, even when this is accessed via one or more pronouns. Given that pronominal mentions of entities may reflect the salience of their antecedents, pronouns are awarded a bonus of +1.

Syntactic Parallelism (SP): The pre-processing software (FDG-Parser) used by MARS also provides the syntactic role of the NP complements of the verbs. This indicator increases the chances that a NP with the same syntactic role as the current pronoun will be its antecedent by awarding it a boosting score of +1.

Frequent Candidates (FC): This indicator was motivated by our observations during annotation of coreference that texts frequently contain a narrow "spine" of references, with perhaps less than three entities being referred to most frequently by pronouns throughout the course of the document. This indicator awards a boosting score (+1) to the three NPs that occur most frequently in the sets of competing candidates of all pronouns in the text (for a definition of 'set of competing candidates' see Section 2.3).

Five of the original indicators are computed in a different manner by MARS. In the case of the indicator *lexical reiteration*, in addition to counting the number of explicit occurrences of an NP, MARS also counted pronouns previously resolved to that NP. The conditions for boosting them remain the same.

Collocation Match (CM) was originally implemented to boost candidates found in the same paragraph as the pronoun, preceding or following a verb identical or morphologically related to a verb that the pronoun precedes or follows. CM was modified so that in the first step, for every appearance of a verb

[10] Examples of pleonastic *it* include non-referential instances as in 'It is important...', 'It is requested that...', 'It is high time that...' Examples of the pronoun *it* that exhibit non-nominal anaphora are the cases where the antecedent is not an NP but a clause or whole sentence.

in the document, the immediately preceding and immediately following heads (PHEAD and FHEAD respectively) of NP arguments are written to a file. In the case of prepositions, the immediately following NP argument is written. An extract from the resulting file is shown below:

```
VERB replace PHEAD you FHEAD it
VERB replace PHEAD battery FHEAD cover
VERB replace PHEAD printer FHEAD cartridge
VERB replace FHEAD cartridge
VERB replace PHEAD You FHEAD cartridge
VERB replace FHEAD battery
VERB replace PHEAD battery FHEAD it
VERB replace PHEAD You FHEAD battery
VERB replace PHEAD problem FHEAD battery
VERB replace PHEAD you FHEAD battery
VERB replace PHEAD this FHEAD cartridge
VERB replace PHEAD Ink FHEAD Cartridge
VERB replace FHEAD Cartridge
VERB replace FHEAD Ink
```

MARS then consults this data file when executing CM. When resolving the pronoun *it* in sentence 4 of the illustrative paragraph,

> (1) Do not touch the battery terminals with metal objects such as paper clips or keychains. (2) Doing so can cause burns or start a fire. (3) Carry batteries only within the printer or within their original packaging. (4) Leave *the battery* inside the printer until you need to charge or replace *it*.

the NP *the battery* is awarded a boosting score of +2 because the pronoun is in the FHEAD position with respect to the lemma of the verb *replace* and the lemma of the head of *the battery* also appears in the FHEAD position with respect to that verb in the database. Thus, the indicator applies on the basis of information taken from the whole document, rather than information only found in the paragraph.

We are currently investigating the generalisation of CM using semantic information from the WordNet ontology. The method under investigation involves post-processing the data file produced by CM so that each entry is replaced by the most general senses (unique beginners) in WordNet of its elements. It was assumed that patterns appearing with significant frequency in the post-processed file could be used in a more generalised version of CM in which predicates with pronoun arguments and competing candidates are associated with their unique beginners (which we will denote by *Pred-UB* and *Cand-UB* respectively). The data file is then consulted to see if the patterns *Cand-UB* - *Pred-UB* or *Pred-UB* - *Cand-UB* have a significant presence. Candidates involved in those patterns in the data file that have a significant frequency are awarded a boosting score.

Our experiments in using WordNet to generalise the CM indicator have not yielded an improvement in the system, and have diminished MARS's performance overall. There are three reasons for this. Firstly, we have not yet incorporated a word sense disambiguator into our system, though work is underway in that regard with reference to the method proposed in [29]. Instead, we associate each word with the first sense returned in the list by WordNet. Secondly, many of the senses appearing in the somewhat specialised domain of technical manuals are not present in the WordNet ontology. It would require the use of a more

specialised ontology to obtain optimum performance from the system. Thirdly, we have taken the mean frequency of appearance of a pattern in the datafile as the threshold level of significance. It may be possible to improve performance by using more sophisticated methods such as TF.IDF for patterns with respect to all the texts at our disposal.

First NPs has been renamed *obliqueness* (OBL). Following centering theory [15], where grammatical function is used as an indicator of discourse salience, MARS now awards subject NPs a score of +2, objects a score of +1, indirect objects no bonus, and NPs for which the FDG parser is unable to identify a function a penalising score of -1[11].

A clause splitter is not yet incorporated into MARS, so a simplified version of the *referential distance* indicator is implemented, with distance being calculated only in terms of sentences rather than clauses and sentences.

Regarding the *term preference* indicator, in the first implementation of MARS, significant terms were obtained by identifying the words in the text with the ten highest TF.IDF scores. Candidates containing any of these words were awarded the boosting score. In the current implementation, it is the ten NPs that appear with greatest frequency in the document that are considered significant. All candidates matching one of these most frequent NPs are awarded the boosting score.

2.2.2 Classification of *It*

MARS includes a program that automatically classifies instances of the pronoun *it* as pleonastic, examples of non-nominal anaphora, or nominal anaphora [10].

The method was developed by associating each instance of it in a 368830 word corpus with a vector of feature values. 35 feature-value pairs are used, the values being computed automatically by our software. Each feature belongs to one of six different types. *Type 1 features* carry information about the position of the instance in the text. *Type 2 features* describe the number of elements in the surrounding text, such as complementisers and prepositions, which are indicative of the pronoun's class. *Type 3 features* display the lemmas of elements such as verbs and adjectives in the same sentence as the instance. *Type 4 features* show the parts of speech of the tokens surrounding the instance. *Type 5 features* indicate the presence or otherwise of particular sequences of elements, such as *adjective* + NP or *complementiser* + NP, following the instance. *Type 6 features* indicate the proximity of suggestive material such as *-ing* forms of verbs or complementisers, following the instance in the text. The 3171 resultant vectors were then manually classified as belonging to one of the following classes: *nominal anaphoric*; *clause anaphoric*; *proaction*, *cataphoric*; *discourse topic*; *pleonastic*; or *idiomatic/stereotypic*. This manually annotated set of instances constitutes the training file.

[11] Note that the FDG parser proposes grammatical functions for most words. The POS tagger used in the original version was not able to identify syntactic functions and first NPs were used as approximations of subjects.

The classification system works by rendering new feature-value vectors for previously unseen instances of *it* and using TiMBL [6] to classify them with respect to the instances in the training file. The overall classification rate was reported to be 78.74% using ten-fold cross-validation. Table 1 gives more details on the accuracy of this classification over the texts processed in the current study.

2.2.3 Syntactic Constraints

The following constraints proposed by Kennedy and Boguraev [19] that act as knowledge-poor approximations of Lappin and Leass's [20] syntax filters, were also implemented in MARS's latest version: *A pronoun cannot corefer with a co-argument*, *a pronoun cannot co-refer with a non-pronominal constituent which it both commands and precedes*, and *a pronoun cannot corefer with a constituent which contains it.* These constraints are applied before activating the antecedent indicators and after the gender and number agreement tests.

2.2.4 Identification of Animate Entities

Evans and Orasan [9] presented a robust method for *identifying animate entities* in unrestricted texts, using a combination of statistics from WordNet [11] and heuristic rules.

Here, seven unique beginners from WordNet were taken to contain senses that in the case of nouns, usually refer to animate entities, and in the case of verbs, usually take animate subjects. For the NPs in a text, their heads were scrutinised in order to count the number of animate/inanimate senses that they can be associated with. In the case of subject NPs, their predicates were scrutinised in a similar fashion. The information concerning the number of an entity's animate/inanimate senses was then used when classifying the entity as being either animate or inanimate. The heuristic rules examined the specific form of the NPs in the text, reporting whether or not they contained suggestive complementisers such as *who*, or whether they were in fact pronouns whose gender could be determined in a trivial way. Once each NP was associated with all of this information, a simple rule-based method was used to classify the NP as animate or inanimate.

Overall, the method was shown to be a useful step towards enforcing gender agreement between pronouns and potential antecedents. The method worked adequately over texts containing a relatively high number of animate entities (+5.13% success rate in anaphora resolution), but it was ineffective over texts with relatively few animate entities as a result of the incorrect elimination of valid antecedents (-9.21% success rate on the technical document referred to in Section 3 as PSW).

In subsequent work, Orasan and Evans [26] refined the method for gender identification. In the original method, the unique beginners in WordNet were manually classified as *animate* or *inanimate* in line with the crude expectation that all their hyponyms were likely to refer to animate or inanimate entities. This approach was flawed in that the classification of a unique beginner is not a

very reliable indicator of the classifiction of all of its hyponyms. Addressing this problem, the new effort used files from the sense-annotated SEMCOR corpus. Head nouns and verbs in those files were then manually annotated as either animate or inanimate depending upon their use in the texts. Chi-squared was then used to classify the hypernyms of the senses whose animacy was known. More specific senses were then taken to share the classification of the hypernyms. Machine learning was coupled with an approach similar to that described in [9] in order to make an automatic classification of NPs in unseen texts. The method described in [26] obtained an accuracy of around 97% in identifying animate entities.

Despite the greater accuracy of this method, we found that it still hinders MARS's performance in the domain of technical manuals, as was the case for the earlier work. Although, with respect to the PSW text, the error rate dropped from 9.21% to 1.33%, application of the method still induces deterioration in system performance in the domain of technical manuals. There are two reasons for this. Firstly, the technical domain refers to specialised senses that cannot be found in WordNet. Secondly, for those senses that are found, they are usually used with a highly specialised meaning. In many cases there is strong evidence from WordNet that nouns such as *computer* or *printer* are normally used to refer to animate entities when in fact they are only used with inanimate senses in computer technical manuals. It may be possible to improve the performance of the system by first performing word sense disambiguation (WSD) in order to limit the number of animate senses that particular nouns are permitted to have with respect to documents from particular domains. Work is currently underway to implement the method for WSD described in [29].

Due to these problems, our methods for identification of animate entities have not been incorporated when running MARS over the technical documents described in Section 3. Instead, gender agreement was only enforced using a gazetteer of first names.

2.3 The algorithm

MARS operates in five phases. In *phase 1*, the text to be processed is parsed syntactically, using Conexor's FDG Parser [30] which returns the parts of speech, morphological lemmas, syntactic functions, grammatical number, and most crucially, dependency relations between tokens in the text which facilitates 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 [10]. In its current implementation, MARS is only intended to resolve third person pronouns and possessives of singular and plural number that demonstrate identity-of-reference nominal anaphora.

In *phase 3*, for each pronoun identified as anaphoric, potential antecedents (candidates), are extracted from the NPs in the heading of the section in which the pronoun appears, and from NPs in the text preceding the pronoun up to the limit of either three sentence boundaries or one paragraph boundary, whichever

contains the smallest amount of text. 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. Firstly, competing candidates are required to agree with the pronoun with respect to number and gender, as was the case in the original version of MARS. Secondly, they must obey the syntactic constraints described in Section 2.2.3.

In *phase 4*, preferential and impeding factors (a total of 14) 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.

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.

2.4 Using genetic algorithms to search for optimal performance

The scores of the antecedent indicators as proposed in Mitkov's original method were derived on the basis of empirical observations, taking their decision power into consideration, and have never been regarded as definite or optimal. By changing the scores applied by the antecedent indicators, it is possible to obtain better success rates.

Given that the score of a competing candidate is computed by adding the scores applied by each of the indicators, the algorithm can be represented as a function with 14 parameters, each one representing an antecedent indicator

$$score_k = \sum_{i=1}^{i=14} x_{k_i} \tag{1}$$

where $score_k$ is the composite score assigned to the candidate k, and x_{k_i} is the score assigned to the candidate k by the indicator i. The goal of a search method would be to find the set of indicator scores for which the composite score is maximum for the antecedents and lower for the rest of candidates. This would lead to a high success rate.

Genetic algorithms (GA) seemed the most appropriate way of finding the optimal solution. First proposed by Holland [18], GA mimic reproduction and selection of natural populations to find the solution that maximises a function, called fitness. The GA maintains a population of candidate solutions to the fitness function represented in the form of chromosomes. For our problem, each chromosome, representing a set of indicator scores, is a string of 34 real numbers; each value representing the outcome of an indicator application. The alphabet used to represent chromosomes is the set of real numbers. As a fitness function we used the number of anaphors correctly resolved by the system when a candidate solution's indicator scores are applied by the algorithm. Therefore, maximisation of the fitness function leads to an increase in the success rate.

The main use of the GA is to find the upper limits of a method based on numerical preferences. In this case, the algorithm does not try to find a general set of scores that could be useful over general texts. Instead, it searches the solution space for a set which maximises the fitness function (success rate) for a certain text. This value represents the maximum success rate that the given preference-based algorithm can obtain for that text. A secondary usage of the GA is as an optimisation method. In this case, the set of indicators which maximises the success rate for a particular file is applied by the algorithm when processing different files. The results of such cross-evaluation are presented in Section 3 and discussed in Section 4.

3 Evaluation

MARS was evaluated on eight different files, from the domains of computer hardware and software technical manuals, featuring 247,401 words and 2,263 anaphoric pronouns (Table 1). Each text was annotated coreferentially in accordance with the methodology reported in [23]. Applied over this corpus, MARS obtained an average success rate of 59.35%. Success rate is defined as the ratio of the number of anaphoric pronouns that MARS resolves correctly to the number of anaphoric pronouns in the text. We do not take the number of pronouns that the system attempts to resolve as the denominator because this would mean that a system that only attempted to resolve pronouns with a single candidate could obtain unfairly high levels of performance.

Each technical manual is identified by an abbreviation in column 1 of Table 1. Column 2 shows the size of the text in words, column 3 displays the number of anaphoric pronouns[12], column 4 shows the number of pronouns in the text that are instances of non-nominal anaphora or pleonastic *it*. Column 5 shows the accuracy with which the system is able to classify instances of the pronoun *it*. The reader will note that these figures are markedly improved over those reported in [10]. This is explained by the fact that in that paper, the system was tested over texts from many different genres, which included free narrative and direct speech. In the domain of technical manuals, instances of *it* are found in far more constrained and predicable linguistic contexts, resulting in greater reliability on the part of the machine learning method. Of the anaphoric pronouns, 1709 were intrasentential anaphors and 554 - intersentential. In 238 cases the antecedents were not on the list of candidates due to pre-processing errors.

The overall success rate of MARS was 59.35% (1343/2263). After using GA [25], the success rate rose to 61.55% (1393/2263). Table 2 gives details on the evaluation of MARS - covering the standard version and the version in which the GA was used to obtain the set of scores leading to optimal performance. As a result of errors at the level of NP extraction, and therefore possible omission of antecedents, the success rate of MARS cannot reach 100%. In the MAX column, the theoretical maximum success rate that MARS can obtain as a result of pre-processing errors is indicated. The column *Sct* represents the maximum possible

[12] More accurately, pronouns that demonstrate nominal identity-of-reference anaphora.

Table 1. The characteristics of the texts used for evaluation

Text	#Words	#Anaphoric pronouns	#Non-nominal anaphoric/pleonastic *it*	Classification accuracy for *it*
ACC	9753	157	22	81.54%
BEO	7456	70	22	83.02%
CDR	10453	83	7	92.86%
GIMP	155923	1468	313	83.42%
MAC	15131	149	16	89.65%
PSW	6475	75	3	94.91%
SCAN	39328	213	22	95.32%
WIN	2882	48	3	97.06%
Total	247401	2263	408	85.54%

success rate when a pronoun is considered correctly resolved only if the whole NP representing its antecedent is selected as such, in its entirety. As can be seen, this figure does not exceed 92%. Given the preprocessing errors, inevitable in an automatic system, we considered a pronoun correctly resolved if only part of a pronoun's antecedent was identified and that part included the head of the NP (as proposed in MUC-7 [16]). When this partial matching is considered, the maximum success rate can reach the values presented in the column *Ptl*. Two baseline models, presented in the *Baseline* column, were evaluated, one in which the most recent candidate was selected as the antecedent (*Rcnt*) and one in which a candidate was selected at random (*Rand*) - both after agreement restrictions had been applied.

The column *Old* displays the performance of a fully automatic implementation of the algorithm proposed in [21]. We should emphasise that it follows the method briefly discussed in Section 1 without including any additional components such as new or modified indicators or recognition of pleonastic pronouns. The values in this column are noticeably lower than those obtained for any of the subsequent systems.

We evaluated MARS in four different configurations: Default (*Dflt*), in which the system described in Section 2.3 is run in its entirety; *no it filter*, where the system is run without attempting to identify pleonastic/non-nominal instances of *it*; *no num/gend agr*, where the system is run without applying number and gender agreement constraints between pronouns and competing candidates, and *no syn constr*, where no syntactic constraints are enforced between pronouns and intrasentential candidates. Of course, more combinations are possible, but due to space and time constraints, we did not evaluate them. By comparing these columns with the *dflt* column, for example, it is possible to see that, overall, MARS gains around 30% in performance as a result of enforcing number and gender agreement between pronouns and competing candidates. For each configuration and each text, we obtained MARS's success rate, displayed in the column *Standard*. Additionally, we used the GA described in Section 2.4 to find the upper limit of MARS's performance when the optimal set of indicator scores is applied, displayed in the column *Upper bound*. In this case, the GA was used

Table 2. Success rates for different versions of MARS

Files	Old (2000)	MARS								MAX		Baseline	
		Standard				Upper bound							
		Dflt	no *it* filter	no num /gend agr	no syn constr	Dflt	no *it* filter	no num /gend agr	no syn constr	Sct	Ptl	Rcnt	Rand
ACC	33.33	51.59	52.87	35.67	49.04	55.41	55.41	43.31	43.31	73.88	96.18	28.02	26.75
BEO	35.48	60.00	60.00	45.71	60.00	67.14	64.28	50.00	67.14	81.43	95.71	35.71	22.86
CDR	53.84	67.47	68.67	51.81	67.47	75.90	74.69	54.22	74.69	78.31	95.18	36.14	43.37
GIMP	-	57.15	60.42	17.57	57.63	57.83	60.83	18.94	57.22	79.70	91.69	37.80	30.72
MAC	53.93	71.81	69.79	60.40	71.14	75.84	77.85	67.11	76.51	83.89	96.64	51.68	44.97
PSW	64.55	82.67	84.00	80.00	82.67	86.67	90.67	80.00	89.33	92.00	97.33	49.33	45.33
SCAN	-	61.50	62.44	46.48	60.56	63.85	64.79	51.64	63.85	79.81	87.32	32.39	30.52
WIN	33.32	52.08	62.50	39.58	52.08	68.75	66.67	60.42	68.75	81.25	87.50	37.50	18.75
TOTAL	45.81	59.35	61.82	29.03	59.35	61.55	63.68	32.04	60.41	80.03	92.27	37.78	31.82

as a search algorithm and not as a general optimisation method. It allowed us to explore the limitations of this knowledge poor pronoun resolution system.

The optimal indicator scores obtained after applying the GA to a specific text were applied when running the algorithm on different texts, in order to make a blind test and to ascertain the general usefulness of genetic optimisation. The results of the cross-evaluation were quite disappointing.

Table 3. The results of cross-evaluation

Inds/ Texts	ACC	BEO	CDR	MAC	PSW	WIN	SCAN	GIMP
ACC	55.41	47.77	47.13	45.22	42.67	45.86	44.59	51.59
BEO	48.57	67.14	52.86	45.72	51.43	60.00	58.57	65.71
CDR	60.24	71.08	75.90	48.19	57.83	57.83	62.65	71.08
MAC	61.74	64.43	63.76	75.84	63.09	61.74	65.77	65.77
PSW	81.33	73.33	72.00	77.33	86.67	74.67	74.67	78.67
WIN	41.67	47.92	52.08	47.92	43.75	68.75	52.08	52.08
SCAN	50.23	55.87	54.46	54.93	47.42	54.93	63.85	53.05
GIMP	51.43	55.04	51.91	53.06	49.80	51.77	50.89	57.83

In most cases the success rates obtained were lower that the ones obtained by the *Standard* version of MARS. The application of the GA will be discussed further in Section 4.

3.1 The influence of indicators

Relative importance is a measure showing how much the system's performance is degraded when an indicator is removed from the algorithm [24][13]. We computed this measure for each indicator and each file, before and after the GA was applied.

[13] Similar to the measure used in [20].

Table 4. Standard relative importance

W/O	ACC	BEO	CDR	MAC	PSW	WIN	SCAN	GIMP	TOTAL
INDEF	-0.64%	-1.43%	0%	-2.01%	+3.95%	0%	-1.88%	+0.14%	-0.18%
OBL	+7.01%	+11.43%	+6.02%	-2.01%	-1.31%	-10.42%	+7.98%	+4.90%	+0.62%
IV	0%	0%	0%	0%	0%	0%	0%	+0.14%	+neg%
REI	-2.55%	-2.86%	+2.41%	+2.01%	-1.31%	-10.42%	-1.41%	+0.27%	-0.26%
SH	-0.64%	+2.86%	+2.41%	+0.67%	0%	-6.25%	+0.94%	+0.82%	+0.66%
PNP	0%	0%	0%	-3.35%	0%	0%	-0.47%	+0.48%	+neg%
CM	+1.27%	0%	0%	+0.67%	+2.63%	+2.08%	+3.29%	+0.82%	+1.10%
IR	0%	0%	-1.20%	+2.01%	0%	0%	+0.47%	+0.14%	+0.22%
SI	0%	0%	0%	0%	0%	0%	0%	0%	0%
RD	+3.18%	+5.71%	+1.20%	+1.34%	+2.63%	+12.50%	+3.29%	+5.31%	+4.64%
TP	0%	0%	-2.40%	-0.67%	0%	+2.08%	0%	-0.61%	-0.49%
BP	+3.82%	0%	+2.40%	-0.67%	0%	0%	+0.47%	+0.54%	+0.71%
SP	+1.27%	0%	+1.20%	-1.34%	-1.31%	+2.08%	+2.35%	+1.02%	+0.93%
FC	+0.64%	0%	0%	-0.67%	0%	+2.08%	0%	-0.34%	-0.18%

In some cases, there were negative values for relative importance reflecting the fact that in some isolated cases, depending on the particular characteristics of the text, removing one of the indicators actually improved MARS's performance. The relative importance of each indicator is displayed in Table 4 (before the GA is applied) and Table 5 (following application of the GA). Our findings are discussed in Sections 3.1.1 and 3.1.2.

Interestingly, after we had made the assessment of the importance of each indicator, and deactivated those with no importance or negative importance so that only the positively important were in effect, overall, MARS performed slightly worse than when all indicators were active (success rate of 59.21 vs. 59.35).

3.1.1 Original Indicators Our examination of the relative importance of each indicator with respect to each file showed that for the *Standard* version of MARS, the most important of the original indicators was SH in most of the cases. Due to the differences in the current implementation of RD, and its original statement, the importance of that indicator is discussed in 3.1.2. On the texts used for evaluation, the relative importance of SI and INDEF is negligible. The rest of the indicators have a moderate influence. A similar observation can be made for the version of the algorithm after the GA was applied, though the difference in importance between indicators is somewhat reduced. SH, PNP, and IR are the most important of the original indicators after application of the GA.

3.1.2 New/modified indicators With respect to the new and modified indicators presented for the first time in this paper, we noted the following. RD, even without access to information on a sentence's internal structure, is the most

Table 5. Relative importance after the GA was applied

W/O	ACC	BEO	CDR	MAC	PSW	WIN	SCAN	GIMP	TOTAL
INDEF	0%	0%	0%	0%	+1.33%	-2.08%	+1.88%	-0.27%	0%
OBL	+2.55%	-1.43%	+1.20%	-1.34%	+1.33%	+6.25%	+2.82%	+1.97%	+1.81%
IV	0%	-1.43%	0%	-2.01%	0%	0%	+0.47%	+0.82%	+0.40%
REI	0%	-1.43%	-1.20%	+1.34%	0%	0%	0%	-0.34%	-0.22%
SH	+1.27%	0%	+3.61%	-1.34%	0%	-2.08%	0%	+0.27%	+0.26%
PNP	-1.27%	0%	-2.40%	-0.67%	+1.33%	-2.08%	+1.41%	+0.14%	0%
CM	+1.27%	-1.43%	0%	-1.34%	+1.33%	0%	+1.88%	+2.11%	+1.55%
IR	+0.64%	-1.43%	-1.20%	0%	0%	0%	+0.94%	+1.29%	+0.88%
SI	-0.64%	-1.43%	0%	0%	0%	0%	+1.41%	+0.48%	+0.35%
RD	+1.27%	+10.00%	+2.40%	+4.03%	+5.33%	+8.33%	+5.63%	+6.33%	+5.74%
TP	+1.27%	0%	0%	-0.67%	0%	0%	+1.88%	+1.02%	+0.88%
BP	+1.27%	-2.86%	-1.20%	+0.67%	0%	-2.08%	+1.41%	-0.20%	-neg%
SP	-1.27%	0%	+1.20%	+0.67%	0%	0%	+3.29%	-0.14%	+0.22%
FC	-0.64%	-2.86%	0%	0%	+1.33%	-2.08%	+1.88%	-2.11%	-1.31%

important of the modified indicators, followed by CM. Although of variable importance over different texts, overall, OBL and SP make a positive contribution in both the *Standard* and *Upper bound* versions of MARS. On the other hand, REI has negative importance. We can account for this because the pronoun resolution process is itself imprecise and the fact that REI counts pronouns resolved by MARS to NPs as additional mentions of those NPs will make it somewhat inaccurate. Perhaps for similar reasons, the importance of BP was variable, having positive importance in the *Standard* version and negligibly negative importance in the *Upper bound* version. The importance of TP was negative in the *Standard* version of MARS but positive in the *Upper bound* version. It is very probable that the implementation of this indicator can be improved by using better algorithms to identify the significant terms in the texts. Of variably negative and positive importance when applied over different texts, the FC indicator was of negative importance overall, despite the observations and justification for this indicator presented in Section 2.2.1.

3.2 The influence of an automatic classification of *it*

The reader will note, by comparison of columns 3 and 4 in Table 2, that MARS's performance is slightly better, in terms of success rate, when no attempt is made at recognition of pleonastic/non-nominal *it*. Overall, as a result of classifying *it*, the success rate drops by more than 2%. This is due to inaccuracies in the classification module with some anaphoric instances of *it* being incorrectly filtered. In light of this, one may conclude that the pronoun classification module should be eliminated. However, we argue that the reader is drawn to this conclusion by inadequacies in the definition of success rate. In Section 4, we argue that success rate cannot capture the positive contribution made by the classification module and a new measure of performance is proposed.

3.3 The influence of syntactic constraints

In Table 2, the column *no syn constr* shows MARS's performance when the syntactic constraints described in Section 2.2.3 are not applied between pronouns and their competing candidates. Comparison of the *Dflt* columns with these shows the scale of the contribution to the system made by syntactic and agreement constraints. The contribution made by the syntactic constraints (around +2% success rate overall for the *Upper bound* version of MARS and no contribution in the *Standard* version) is not as great as may be expected. This is due to their reliance on an accurate global parse of sentences, which was not always obtained for the texts that we processed.

4 Discussion

The evaluation results give rise to a number of interesting conclusions that can be made with regard to the approach presented and with regard, more generally to anaphora resolution.

To start with, a close look at the MAX columns in Table 2 clearly shows the limits of fully automatic anaphora resolution, based on a given pre-processing tool, with candidates extracted from a range of two sentences from the pronoun. Systems depend on the efficiency of the pre-processing tools which analyse the input before feeding it to the resolution algorithm. Inaccurate pre-processing can lead to a considerable drop in the performance of the system, however accurate an anaphora resolution algorithm may be. The accuracy of today's pre-processing is still unsatisfactory from the point of view of anaphora resolution. Whereas POS taggers are fairly reliable, full or partial parsers are not. Named entity recognition is still a challenge, with the development of a product name recogniser being a vital task for a number of genres. While recent progress in areas such as identification of pleonastic pronouns [10], identification of non-anaphoric definite descriptions [3]; [32] and recognition of animacy [9] have been reported, these tasks and other vital pre-processing tasks such as gender recognition and term recognition, have a long way to go. For instance, the best accuracy reported in robust parsing of unrestricted texts is around the 86% mark [5]; the accuracy of identification of non-nominal pronouns normally does not exceed 80%[14] [10]; though the accuracy of identification of NP gender has reached 97% [26]. Other tasks may be more accurate but are still far from perfect. The state of the art of NP chunking which does not include NPs with post-modifiers, is 90-93% in terms of recall and precision. The best-performing named entity taggers achieve an accuracy of about 96% when trained and tested on news about a specific topic, and about 93% when trained on news about one topic and tested on news about another [14]. Finally, comparison of MARS which employs arguably one of the best shallow parsers for English with Mitkov's original approach which operated on correctly pre-processed texts, shows a drop of up to 25% of the success rate!

[14] However, Paice and Husk [27] reported 92% for identification of pleonastic *it*.

The results also show that the reported success rate is reduced if we consider resolution correct only if the full NP representing the antecedent is identified and if similarly to MUC-7 [16], the task is not simplified to tracking down only a part of the full NP as long as that part contains the head.

The use of the GA allowed us to gain an insight into the limits of this preference-based anaphora resolution method. It was shown that by choosing the right set of indicator scores, it is possible to improve the success rate of the system by up to 3% over all files tested. However, at this stage, we cannot find a method which can determine the optimal set of scores for unseen texts. Cross-evaluation showed that the optimal scores derived by the GA for a text are specific to it and attempts to use them when processing different texts led to low success rates. This result can be explained by over-fitting on the part of the GA with respect to the characteristics of a particular text. Further research on this topic is necessary in order to design a generally applicable optimisation method.

We should note that MARS employs a knowledge-poor algorithm: we do not have any access to real-world knowledge, or even to any semantic knowledge. MARS does not employ full parsing either and works from the output of a POS tagger enhanced with syntactic roles (in most cases) and functional dependency relations. Recent research [28] shows that approaches operating without any semantic knowledge (e.g. in the form of selectional restrictions) usually do not achieve a success rate higher than 75%. In light of this, we find MARS's success rate on a number of files to be encouraging.

The evaluation carried out raises another important issue. We have adopted the measure of success rate since it has been shown [24]; [4] that recall and precision are not always suitable for anaphora resolution. The current definition of success rate as the number of successfully resolved pronouns divided by the total number of pronouns (as marked by humans), however, does not capture cases where the program incorrectly tries to resolve instances of non-nominal anaphora. For programs handling nominal anaphora, we feel it is important to be able to judge the efficiency of the program in terms of removing instances of non-nominal anaphora and not incorrectly attempting to resolve these instances to NPs. Therefore, we believe that a measure which reflects this efficacy would be appropriate.

If an anaphora resolution system is presented with a set P of pronouns, where the subset A are instances of nominal anaphora and subset N are not nominally anaphoric, it may be useful to assess that system using a measure that captures the correctness of its response to all P pronouns. Ideally, such a system will attempt to resolve the set A and filter out the set N. If the system correctly resolves A' of the nominally anaphoric pronouns and correctly filters, N' of the non-nominally anaphoric ones, it can be evaluated using the single ratio, which we call *Resolution Etiquette*, $RE = 100*(N'+A')/P$. This measure captures the contribution made to the system by both recognition modules for non nominal and pleonastic pronouns and the anaphora resolution module itself, in a way that our previous measure, *success rate* (SR), did not. This measure is intended

Table 6. Evaluation of different configurations of MARS using SR and RE

File	Default		no *it* filter	
	SR	RE	SR	RE
ACC	51.59	49.17	52.87	45.86
BEO	60.00	60.21	60.00	45.16
CDR	67.47	67.03	68.67	62.64
GIMP	57.15	56.03	60.42	49.75
MAC	71.81	70.30	69.79	63.03
PSW	82.67	81.01	84.00	79.75
SCAN	61.50	62.13	62.44	56.60
WIN	52.08	54.90	62.50	58.82
TOTAL	59.35	58.21	61.82	52.24

to describe a system's ability to "behave appropriately" in response to a set of pronouns. Table 6 compares the *success rate* and *resolution etiquette* scores obtained by MARS when run with and without the recognition component for non-nominal and non-anaphoric pronouns.

It should be pointed out that a direct comparison between SR and RE is not appropriate. The purpose of Table 6 is not to compare them, but to show the ability of RE to capture the contribution made by the pronoun classification module.

When the pronoun classification module is deactivated, we notice an increase in SR. This is caused because the pronoun classification module incorrectly filters some nominal-anaphors. By definition, SR can only capture errors made by the classification module; its successful filtration of non-nominal anaphora/pleonastic pronouns is ignored by that measure. In contrast, the measure RE is much reduced when the pronoun classification module is deactivated. Even though the module incorrectly filters some nominally anaphoric pronouns, this side effect is outweighed by the correct filtration of non-nominal and pleonastic pronouns. Deactivating the module reduces MARS's ability to respond appropriately to the pronouns it is presented with, making it less useful in further NLP applications such as MT, information retrieval, information extraction, or document summarisation. The RE measure reflects this whereas SR does not. However, we appreciate that as this is a new measure, a comparison of MARS with other systems, using this measure, is not possible.

5 Conclusion

A new, advanced, and fully automatic version of Mitkov's knowledge-poor approach to pronominal anaphora resolution has been proposed in this paper. We have argued that there is a big difference between previously proposed anaphora resolution methods that were tested over small texts, in which most of the preprocessing steps were post-edited, and fully automatic systems which have to

deal with messy data, and errors. The new method has been thoroughly evaluated with respect to 8 technical manuals. By means of a GA, the practical limitations of the system have been revealed. As a result of the insights gained during the evaluation phase, a new measure that is argued to better reflect the performance of fully automatic anaphora resolution systems has been proposed.

References

1. Aone, C. and Bennett, S. W. (1995) Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the ACL (ACL'95)*, pp. 122-129. ACL.
2. Barbu, C. and Mitkov, R. (2000) Evaluation environment for anaphora resolution. In *Proceedings of the International Conference on Machine Translation and Multilingual Applications (MT2000)*, 18-1-18-8. Exeter, UK.
3. Bean, D. L. and Riloff, E. (1999) Corpus-based Identification of Non-anaphoric Noun Phrases. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*. pp. 373-380. ACL.
4. Byron, D. (2001) A proposal for consistent evaluation of pronoun resolution algorithms. *Computational Linguistics*. Forthcoming. MIT Press.
5. Collins, M. (1997) Three Generative, Lexicalised Models for Statistical Parsing. In *Proceedings of the 35th Annual Meeting of the ACL (jointly with the 8th Conference of the EACL)*, Madrid.
6. Daelemans, W. (1999) *TiMBL: Tilburg Memory Based Learner version 2 Reference Guide*, ILK Technical Report - ILK 99-01, Tilburg University, The Netherlands
7. Dagan, I. and Itai, A. (1990) Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, Vol. III, 1-3, Helsinki, Finland.
8. Dagan, I. and Itai, A. (1991) A statistical filter for resolving pronoun references. In Y.A. Feldman and A. Bruckstein (Eds) *Artificial Intelligence and Computer Vision*, pp. 125-135. Elsevier Science Publishers B.V. (North-Holland).
9. Evans R. and Orasan, C. (2000) Improving anaphora resolution by identifying animate entities in texts. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC 2000)*. pp. 154-162. Lancaster, UK.
10. Evans, R. (2001) Applying Machine Learning Toward an Automatic Classification of It. *Journal of Literary and Linguistic Computing*. 16(1) pp. 45-57. Oxford University Press.
11. Fellbaum, C. (ed) (1998) *WordNet An Eletronic Lexical Database*. MIT Press
12. Ferrández, A., Palomar, M., and Moreno, L. (1998) Anaphora resolution in unrestricted texts with partial parsing. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*, pp. 385-391. Montreal, Canada.
13. Ge, N., Hale, J. and Charniak, E. (1998) A statistical approach to anaphora resolution. In *Proceedings of the Workshop on Very Large Corpora*, pp. 161-170. Montreal, Canada.
14. Grishman, R. (Forthcoming) Information Extraction. In R. Mitkov (Ed.), *Oxford Handbook of Computational Linguistics*. Oxford University Press, forthcoming.
15. Grosz, B. J., Joshi, A., and Weinstein, S. (1995) Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2), pp. 44-50. MIT Press.

16. Hirschman, L. and Chinchor, N. (1997) *MUC-7 Coreference Task Definition* at *http : //www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task .html*
17. Hobbs, J. R. (1978) Resolving pronoun references. *Lingua*, 44, pp. 339-352.
18. Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems*, University of Mitchigan Press, US.
19. Kennedy, C. and Boguraev, B. (1996) Anaphora for everyone: pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pp. 113-118. Copenhagen, Denmark.
20. Lappin, S. and Leass, H.J. (1994) An Algorithm for Pronominal Anaphora Resolution, in *Computational Linguistics* Volume 20, Number 4
21. Mitkov, R. (1998) Robust pronoun resolution with limited knowledge. In *Proceedings of the 18.th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*, pp. 869-875. Montreal, Canada.
22. Mitkov, R. (2000) Towards more comprehensive evaluation in anaphora resolution. In Proceedings of *The Second International Conference on Language Resources and Evaluation, volume III*, pp. 1309-1314, Athens, Greece. ELRA.
23. Mitkov, R., Evans, R., Orasan, C., Barbu, C., Jones, L., and Sotirova, V. (2000) Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, pp. 49-58. Lancaster, UK.
24. Mitkov, R. (2000) Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, pp. 96-107. Lancaster, UK.
25. Orasan C., Evans, R. and Mitkov, R. (2000) Enhancing Preference-Based Anaphora Resolution with Genetic Algorithms. In *Proceedings of NLP'2000*, Patras, Greece. pp. 185-195.
26. Orasan, C. and Evans, R. (2001) Learning to identify animate references. In *Proceedings of the Workshop Computational Natural Language Learning 2001 (CoNLL-2001)*. ACL. Toulouse.
27. Paice, C.D. And Husk, G.D. (1987) Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun 'it,' in *Computer Speech and Language*, 2 pp. 109-132, Academic Press, US.
28. Palomar, M., Moreno, L., Peral, J., Munoz, R., Ferrandez, A., Martinez-Barco, P., and Saiz-Noeda, M. (2001) An algorithm for anaphora resolution in Spanish texts. Forthcoming.
29. Resnik, P. (1995) Disambiguating noun groupings with respect to WordNet senses. In *Proceedings of the Third Workshop on Very Large Corpora*. ACL. New Jersey. pp. 54-68.
30. Tapanainen, P. and Järvinen, T. (1997) A Non-Projective Dependency Parser, in The Proceedings of The *5th Conference of Applied Natural Language Processing*, pages 64-71, ACL, US.
31. Tetreault, J. R. 1999. Analysis of Syntax-Based Pronoun Resolution Methods. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pp. 602 - 605, ACL, University College Maryland. US.
32. Vieira, R. and Poesio, M. (2000) An Empirically-Based System for Processing Definite Descriptions. *Computational Linguistics*, v. 26, n.4.