Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems

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

Abstract

The paper argues that evaluation of anaphora resolution algorithms and anaphora resolution systems should be carried out separately and shows that recall and precision are imperfect as measures for anaphora resolution algorithms. The paper proposes a package of evaluation measures and tasks for anaphora resolution which provide a clearer, more comprehensive picture of the performance of both anaphora resolution algorithms and systems. Finally, the development of a consistent evaluation environment for anaphora resolution is outlined.

1. Introduction

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

Against the background of growing interest in the field, it seems that still insufficient attention has been paid to the evaluation of the systems developed. Even though the number of works reporting extensively on evaluation in anaphora resolution is increasing (Azzam et al. 1998; Baldwin 1997; Cardie & Wagstaff 1999; Gaizauskas & Humphreys 1996; Lappin & Leass 1994; Mitkov 1998, 2000; Mitkov & Stys 1997; Tetrault 1999; Walker 1989), the forms of evaluation that have been proposed are neither sufficient nor perspicuous.

The studies carried out so far have not distinguished between the evaluation of an anaphora resolution algorithm and the evaluation of an anaphora resolution system. As a result, the findings reported often vary significantly and fail to provide common ground for comparison.

The MUCs (Message Understanding Conferences) promoted the use of recall and precision for evaluating the performance of coreference resolution systems (Aone and Bennett 1995; Baldwin 1997, Gaizauskas and Humphreys 1996). While these measures have been successfully used for fully implemented coreference resolution systems, we argue that evaluating an anaphora resolution algorithm or system in terms of these measures does not always contribute to its consistent evaluation. As an alternative, we propose the simple measure success rate which should be computed separately for anaphora resolution algorithms and anaphora resolution systems. Our view is that this measure should be backed by a number of additional measures and tasks with a view to providing a comprehensive overall assessment of an approach (or a system). In order to see how much a certain algorithm (or system) is “worth”, it would be necessary to assess it against other “benchmarks”, e.g. against other existing or baseline models. It also makes sense to evaluate the performance on anaphors which do not point to sole candidates for antecedents and which cannot be disambiguated on the basis of gender and number agreement alone (see the notions of non-trivial success rate and critical success rate, section 5.1). Finally, a comparison with other similar or well-known approaches/systems would serve to indicate where the approach/system stands in the state of play of anaphora resolution.

Furthermore, the evaluation would be more revealing if, in addition to evaluating a specific approach as a whole, we break down the evaluation process by looking at the different components involved. In the case of factor-based anaphora resolution, we propose methods for evaluation of each individual factor employed in the algorithm. Such evaluation would provide important insights as to how the overall performance of factor-based systems

1 Anaphora and coreference are not identical linguistic phenomena: anaphora is the pointing back to a previously mentioned item in the text as opposed to coreference which is the act of referring to the same referent in the real world. Anaphora resolution and coreference resolution differ as tasks as well: the objective of coreference resolution is to identify all coreference classes, whereas that of anaphora resolution is to identify an antecedent of the anaphor. In the case of identity-of-reference nominal anaphora the latter boils down to tracking down a preceding NP from the coreferential chain of the anaphor (this class of anaphora involves coreference).
could be improved (e.g. through changing the weights/scores of the factors). In this work we propose the notion of *decision power* of anaphora resolution factors which can play an important role in preferential architectures.

The paper is structured as follows. Section 2 briefly outlines the approach which we use as a testbed for our evaluation. Section 3 proposes to evaluate anaphora resolution algorithms and systems separately in terms of success rate. Section 4 comments on the lack of clarity and insufficient coverage of the measures recall and precision when used for anaphora resolution algorithms. Section 5 elaborates on the evaluation measures and tasks that we have taken on board and reports on some results from the evaluation of our robust algorithm. Section 6 discusses the evaluation of anaphora resolution systems and in particular the automatic, as opposed to the human-mediated, resolution of anaphors. Section 7 discusses the reliability of the evaluation results, whereas section 8 outlines the on-going development of a new evaluation environment (evaluation workbench) for anaphora resolution.

### 2. Evaluation: using our robust, knowledge-poor pronoun resolution approach as a testbed

The approach which we used as a testbed for our evaluating methodology was Mitkov’s robust, knowledge-poor approach to pronoun resolution (Mitkov 1998) which will be referred to as the *knowledge-poor approach*. Since the evaluation methodology presented in this paper also includes evaluation of the components of algorithms, we deem it appropriate to outline this approach first.

#### 2.1. The knowledge-poor approach: a brief outline

With a view to avoiding complex syntactic, semantic and discourse analysis, we developed a robust, knowledge-poor preference-based approach to pronoun resolution which does not make use of syntactic and semantic knowledge or any form of non-linguistic information. The core of our approach lies in activating a set of antecedent indicators after filtering candidates from the current and three preceding sentences on the basis of gender and number agreement. The approach operates as follows: it works from the output of a text processed by a part-of-speech tagger and an NP extractor, locates noun phrases which precede the anaphor within a distance of 3 sentences, checks them for gender and number agreement with the anaphor and then applies the indicators to the remaining candidates by assigning a positive or negative score (-1, 0, 1 or 2). The noun phrase with the highest composite score is proposed as antecedent.

The indicators employed can be either *boosting* or *impeding*. The boosting ones apply a positive score to an NP, reflecting a positive likelihood that it is the antecedent of the current pronoun. In contrast, the impeding indicators 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) or to structural matches, whereas others are genre-specific. In the following we shall outline the indicators used and illustrate some of them by examples.

The *boosting indicators* are:

**First Noun Phrases:** A score of +1 is assigned to the first NP in a sentence.

**Indicating Verbs:** A score of +1 is assigned to those NPs immediately following a verb which is a member of a predefined set (including verbs such as *discuss*, *introduce*, *summarise*, *highlight* etc.).

**Lexical Reiteration:** A score of +2 is assigned to those NPs repeated twice or more in the paragraph in which the pronoun appears, a score of +1 is assigned to those NPs repeated once in that paragraph.

**Section Heading Preference:** A score of +1 is assigned to those NPs that also occur in the heading of the section in which the pronoun appears.

**Collocation Pattern Preference:** A score of +2 is assigned to those NPs that have an identical collocation pattern to the pronoun.

This preference is given to candidates which have an identical collocation pattern with a pronoun. The collocation preference here is restricted to the patterns `<noun phrase (pronoun), verb>` and `<verb, noun phrase (pronoun)>` or if the verb is *to be*, `<noun phrase (pronoun), verb, adjective/past participle>`.

Example:

*Press the key down and turn the volume up... Press it again.*

Owing to lack of syntactic information, this preference is somewhat weaker than the collocation preference described in (Dagan and Itai 1990).

The collocation pattern preference has been extended to patterns `<(un)V-NP, anaphor>` and `<NP/anaphor, (un)V>`, i.e. verbs with an "undoing action" meaning are considered to fall into collocation patterns along with their "doing action" counterparts. This extended new rule helps in cases such as "Loading a cassette or unloading it". Also, we would consider a certain pattern still a collocation, if the verb featured as a gerund (e.g. "When you plug in the power adapter, the print head moves to its protected position (you’ll hear it moving)").

**Immediate Reference:** A score of +2 is assigned to those NPs appearing in constructions of the form "...(You)
V
1
NP ... con (you) V
2
it (con (you) V
3
it)”, where con ∈ {and/or/before/after/then...}.

This preference can be viewed as a modification of the collocation preference. It also occurs quite frequently in imperative constructions:

To turn on the printer, press the Power button and hold it down for a moment.

**Sequential Instructions:** A score of +2 is applied to NPs in the NP
1
position of constructions of the form:
"To V
1
NP
1
, V
2
NP
2
. (Sentence). To V
3
it, V
4
NP
4
, the noun phrase NP
1
is the likely antecedent of the anaphor it (NP
1
is assigned a score of 2).

To turn on the video recorder, press the red button.
To programme it, press the ‘Programme’ key.

**Term Preference:** A score of +1 is applied to those NPs identified as representing terms in the genre of the text.

The last three indicators (immediate reference, sequential instructions and term preference) are genre-specific.

The **impeding indicators** are:

**Indefiniteness:** Indefinite NPs are assigned a score of -1.

**Prepositional Noun Phrases:** NPs appearing in prepositional phrases are assigned a score of -1.

Insert the cassette into the VCR making sure it is suitable for the length of recording.

Here the noun phrase the VCR is penalised for being part of the prepositional phrase into the VCR.

One indicator, **Referential Distance**, may impede or boost a candidate’s chances of being selected as the antecedent of a pronoun depending on that NP’s distance in terms of clause and sentence boundaries from the pronoun. NPs in the previous clause to the pronoun are assigned a score of +2, those in the previous sentence to the pronoun are assigned a score of +1, those in the sentence prior to that are assigned a score of 0 and more distant pronouns are assigned a score of -1.

The robust algorithm can be summarised as a three-step process. In step one, an agreement filter is applied so that no NP may be considered a suitable candidate for antecedent of a pronoun if it does not agree with the pronoun in terms of number and gender. In step two, a set of boosting and impeding indicators are applied to each candidate NP. In step three, the total score of each candidate is computed by adding the scores of each of its indicators and the candidate with the highest score is selected as the antecedent of the current pronoun. When a number of candidates jointly have the highest score, a number of heuristics are applied to distinguish one as the antecedent.

A more detailed description of each stage of the approach is described in (Mitkov 1998).

### 3. Evaluation in anaphora resolution: two different perspectives

One of our main arguments of this paper is that the evaluation in anaphora resolution should be addressed from two different perspectives depending on whether the evaluation focuses on the anaphora resolution algorithm only or if it covers the performance of the anaphora resolution system. We propose a distinction between evaluation of anaphora resolution algorithms and evaluation of anaphora resolution systems. By anaphora resolution system we refer here to a whole implemented system which processes the text at various levels such as morphologic, syntactic, semantic, discourse etc., in order to produce analysed text which is then fed to the anaphora resolution algorithm. In section 5 and 6 we define the measures success rate of an anaphora resolution algorithm and success rate of an anaphora resolution system.

A natural way to test an anaphora resolution algorithm is to let it run in an “ideal environment” without taking into consideration any possible errors or complications which occur at various pre-processing stages. In contrast, when evaluating an anaphora resolution system, one will certainly have to face performance drop due to the inability of analysing natural language with absolute accuracy. A number of anaphora resolution systems operate either on human-controlled inputs (e.g. pre-analysed corpora or human-corrected outputs from pre-processing modules) or are manually simulated, which suggests that the evaluation they report is concerned with the anaphora resolution algorithm only. On the other hand, there are systems which fully process the text before it is sent to the anaphora resolution algorithm and their evaluation is usually concerned with the evaluation of the anaphora resolution system. A further discussion on automatic anaphora resolution (as opposed to non-automatic) may be found in section 6.

### 4. Evaluation of anaphora resolution algorithms: consistent measures are needed

The Message Understanding Conferences, and in particular MUC-6 and MUC-7 (Hirschman and Chinchor 1997) introduced the measures recall and precision for coreference resolution which have been adopted by a number of researchers for the evaluation of anaphora resolution algorithms or systems. We argue that these measures, as defined, are not satisfactory in terms of clarity and coverage when applied to the evaluation of anaphora resolution algorithms. Consider the following definitions.

**Definition 1** (Aone and Bennett 1995)

\[
\text{Recall} = \frac{\text{Number of correctly resolved anaphors}}{\text{Number of all anaphors identified by the program}}
\]
We propose an evaluation package for evaluating anaphora resolution algorithms consisting of (i) performance measures (ii) comparative evaluation tasks and (iii) component measures. The first cover the overall performance of the algorithm, the second compare the algorithm with other approaches whereas the third look at the efficiency of the separate components of the algorithm. These measures are transferrable to the evaluation of anaphora resolution systems, but the figures obtained in this case will reflect the performance of the whole system and not the resolution module only.

The performance measures are success rate, non-trivial success rate and critical success rate. The comparative evaluation tasks include evaluation against baseline models, comparison with similar approaches and comparison with classical, “benchmark” algorithms. The measures applied to evaluate separate components of the algorithm are decision power and relative importance (see below).

5. Towards a more comprehensive framework of evaluation measures and tasks

We propose an evaluation package for evaluating anaphora resolution algorithms consisting of (i) performance measures (ii) comparative evaluation tasks and (iii) component measures. The first cover the overall performance of the algorithm, the second compare the algorithm with other approaches whereas the third look at the efficiency of the separate components of the algorithm. These measures are transferrable to the evaluation of anaphora resolution systems, but the figures obtained in this case will reflect the performance of the whole system and not the resolution module only.

The performance measures are success rate, non-trivial success rate and critical success rate. The comparative evaluation tasks include evaluation against baseline models, comparison with similar approaches and comparison with classical, “benchmark” algorithms. The measures applied to evaluate separate components of the algorithm are decision power and relative importance (see below).

5.1 Evaluation measures covering the resolution performance of the algorithm

The measures that we propose are illustrated and have been tested on pronominal anaphors, but they can be equally applied to noun phrase anaphora. Note that we restrict the validity of most measures to nominal anaphora which is most extensively studied and best understood in Computational Linguistics.

5.1.1 Success rate

The success rate for an anaphora resolution algorithm

\[
\text{Success rate}_{\text{Anaphora resolution algorithm}} = \frac{\text{Number of successfully resolved anaphors}}{\text{Number of all anaphors}}
\]

reflects the resolution success of an algorithm against all anaphors in the evaluation corpus. Since this measure focuses on the performance of the algorithm and not on any pre-processing modules, the exact success rate will be obtained if the input to the anaphora is either post-edited by humans or is extracted from an already tagged corpus.

Table 1 summarises the success rate of our knowledge-poor algorithm on samples from different manuals; the evaluation texts were automatically pre-processed (POS

---

6 Baldwin’s definition is in line with that used by Gaizauskas and Humphreys (1996) and by Harabagiu and Maiorano (2000).

7 Nominal anaphora is exhibited by NPs – pronouns, definite descriptions and proper names referring to antecedents which are NPs.

8 As marked by humans.

9 On the other hand, the success rate of an anaphora resolution system reflects the performance of the whole system; in this case the text to be processed is normally not expected to be analysed by humans.
tagging, NP identification) but were then manually post-edited to ensure that the input to the algorithm was correct.

<table>
<thead>
<tr>
<th>Manual</th>
<th>Number of anaphoric pronouns</th>
<th>Success rate in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minolta Photocopier</td>
<td>48</td>
<td>95.8</td>
</tr>
<tr>
<td>Portable Style Writer (PSW)</td>
<td>54</td>
<td>83.8</td>
</tr>
<tr>
<td>Alba Twin Speed Recorder</td>
<td>13</td>
<td>100.0</td>
</tr>
<tr>
<td>Seagate Medalist Hard Drive</td>
<td>18</td>
<td>77.8</td>
</tr>
<tr>
<td>Haynes Car Manual</td>
<td>50</td>
<td>80.0</td>
</tr>
<tr>
<td>Sony Video Recorder</td>
<td>40</td>
<td>90.6</td>
</tr>
<tr>
<td>All manuals</td>
<td>223</td>
<td>89.7</td>
</tr>
</tbody>
</table>

Table 1: Success rate(s) of the knowledge-poor approach on different manuals

5.1.2 Non-trivial success rate

The measure non-trivial success rate applies only to the anaphors which have more than one candidate for antecedent, removing those preceded by only one NP in the search scope of the algorithm (and therefore having only one candidate) since their resolution would be trivial.

5.1.3 Critical success rate

The measure critical success rate applies only to those “tough” anaphors which still have more than one candidate for antecedent after gender and number filters. This measure can be very indicative in that it can point to misleading results based on the evaluation of data containing only very easy-to-resolve anaphors (e.g. anaphors that can be resolved directly after gender agreement checks).

More formally, let $N$ be the set of all anaphors involved an evaluation, and $S$ the set of anaphors which have successfully resolved. Further, let $K$ be the set of anaphors which have only one candidate for antecedent (and which are therefore correctly resolved in a trivial way), $M$ the set of anaphors which are resolved on the basis of gender and number agreement and let $n = \text{card } (N)$, $s = \text{card } (S)$, $k = \text{card } (K)$ and $m = \text{card } (M)$. Clearly $s \leq t$, $k \leq t$, $k+m \leq s$, $k \geq 0$, $m \geq 0$, $s \geq 0$. The following relation holds:\(^{10}\)

\[
\text{success rate} \geq \text{non-trivial success rate} \geq \text{critical success rate}
\]

since

\[
\text{success rate} = \frac{s}{n}, \quad \text{non-trivial success rate} = \frac{s-k}{n-k},
\]

\[
\text{critical success rate} = \frac{s-k-m}{n-k-m}
\]

As an illustration, consider evaluation data containing 100 anaphors assuming that 20 of these anaphors have only one candidate for antecedent and that the antecedents of further 10 anaphors can be determined only on the basis of gender and number agreement. Furthermore, let us assume that the algorithm resolves 80 of the anaphors correctly. The success rate would then be $80/100 = 80\%$, the non-trivial success rate would be $60/80 = 75\%$ and the critical success rate $- 50/70 = 71.4\%$.

The non-trivial success rate is indicative of the performance of the algorithm in that it removes from the evaluation anaphors that have no competing candidates for antecedents. The critical success rate is an important criterion for evaluating the efficiency of the factors employed by the anaphora resolution algorithms in "critical cases" where agreement constraints alone cannot point to the antecedent.\(^{11}\) It is logical to assume that good anaphora resolution algorithms have high critical success rates which are close to the overall success rates. In fact, it is really the critical success rate that matters: high critical success rate naturally implies high overall success rate.

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

5.2 Comparative evaluation tasks

The majority of the comparative evaluation tasks described in this section are not a novel idea in that some of them have already been used by other researchers. What is significant in our case is that we have compared the performance of our approach to a fairly representative set of benchmark approaches and models which as a whole should be sufficiently indicative of where the approach stands in the state of art of anaphora resolution. We discuss three classes of benchmark evaluations: evaluation against baseline models, evaluation against approaches that share a similar “philosophy” and evaluation against “classical”, well established approaches in the field.

5.2.1 Evaluation against baseline models

The evaluation against baseline models is important to provide information as to how effective an approach is, by comparing it with typical baseline models. This type of evaluation also justifies the usefulness of the approach developed: however high the success rate may be, it may not be worthwhile developing a specific approach unless the approach demonstrates clear superiority over simple baseline models. We compared our method with (i) a baseline model which checks agreement in number and gender and, where more than one candidate remains, picks

\[s \geq \frac{s-k}{n-k} \geq \frac{s-k-m}{n-k-m}, \quad k \geq 0, m \geq 0, s \geq 0.\]

\(^{10}\) Note that these relations hold in an ‘ideal environment’, when the input to the anaphora resolution algorithm is correctly analysed. For different outcomes in the evaluation of anaphora resolution systems see section 6.

\(^{11}\) Factor-based algorithms typically employ a number of factors after gender and number checks. Factors can be preferences or constraints.
out as antecedent the most recent subject matching the gender and number of the anaphor, and (ii) a baseline model which selects as antecedent the most recent noun phrase that matches the gender and number of the anaphor (Table 2).

<table>
<thead>
<tr>
<th>Approach</th>
<th>Number of anaphoric pronouns</th>
<th>Success rate in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge-poor approach</td>
<td>223</td>
<td>89.7</td>
</tr>
<tr>
<td>Baseline Most Recent</td>
<td>223</td>
<td>65.9</td>
</tr>
<tr>
<td>Baseline Subject</td>
<td>223</td>
<td>48.6</td>
</tr>
</tbody>
</table>

Table 2: Comparison of the success rates of the knowledge-poor approach and two baseline models

The most recent version of the knowledge-poor approach referred to as MARS was also compared to a baseline model which randomly selects the antecedent from all candidates surviving the agreement restrictions (see Section 6 and Table 6). An even weaker baseline model would be to randomly select any candidate before any agreement checks.

5.2.2 Comparison to similar approaches

A comparison with similar methods (if available) or with other well-known (classical) approaches helps to discover what the new approach brings to the current state of the field. Our comparison with similar approaches included running Breck Baldwin’s CogNIAC approach (Baldwin 1997) on part of the evaluation texts (Table 3). The reason for choosing CogNIAC is because both our approach and that of Baldwin share common principles - both are regarded as knowledge-poor and use POS taggers rather than parsers. The MARS version of approach was compared both with Baldwin’s approach and with Kennedy and Boguraev’s (1996) parser-free method. Section 8 provides more details on that evaluation.

5.2.3 Comparison to "classical" approaches: Hobbs’ Naïve Algorithm

We carried out comparative evaluation of Jerry Hobbs’ naïve algorithm (Hobbs 1976) on the basis of the same texts used for the comparative evaluation of Baldwin’s approach (Stylewriter 1994). The results obtained suggest a success rate in the range of 71%.

Hobbs’ naïve algorithm has already been used by other researchers for benchmark evaluation (Baldwin 1997; Tetreault 1999; Walker 1989). The BFP algorithm (Brennan et al. 1987) has also been used for comparison (Tetreault 1999).

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

We believe that it is important to evaluate the performance of the separate components of anaphora resolution algorithms because this type of assessment provides useful insights as to how the approach can be further improved. In particular the evaluation of each resolution factor gives us an idea of the significance or contribution of each factor and provides a basis upon which the factor scores can be adjusted with a view to attaining an overall improvement of the approach. We carried out an evaluation of each antecedent indicator of the knowledge-poor algorithm and concluded that there are two measures of significance: the decision power, which reflects the influence of each indicator on the final decision of the antecedent, and the relative importance which is regarded as the relative contribution of a specific factor in that it is computed as the drop of performance if this indicator were removed. In what follows these measures will be illustrated on the set of antecedent indicators, but they can be computed for any set of anaphora resolution factors.

We define decision power as the measure of the influence of each factor (indicator in the case of our approach) on the final decision, its ability to ‘impose’ its preference in line with, or contrary to, the preference of the remaining factors (indicators). We define the decision power (DP<sub>K</sub>) of a boosting (rewards) indicator K in the following way:

\[ DP_K = \frac{SI_K}{A_K} \]

where SI<sub>K</sub> is the number of successful antecedent identifications (resolutions) when this indicator is applied, and A<sub>K</sub> is the number of applications of this indicator. For the penalising indicators prepositional noun phrase and indefiniteness this figure is calculated as

\[ DP_K = \frac{UI_K}{A_K} \]

where UI<sub>K</sub> is the number of unsuccessful antecedent identifications and A<sub>K</sub> the number of applications of this indicator. The immediate reference emerges as the most “influential” indicator, followed by prepositional noun phrases and collocation pattern preference (Table 4). The relatively low figures for the majority of (seemingly very useful) indicators should not be regarded as a surprise: firstly, one should bear in mind that in most cases a canti-

![Table](image-url)
date is picked (or rejected) as an antecedent on the basis of applying a number of different indicators and secondly, most anaphors have a relatively high number of candidates for antecedent.

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

Table 4: Decision power values for the antecedent indicators

Another way of measuring the importance of a specific factor (indicator) would be to evaluate the approach with this factor "switched off". This measure is called relative importance since it shows how important the presence of a specific factor is. Relative importance ($RI_K$) for a given indicator $K$ is defined as

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

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

6. Evaluation of anaphora resolution systems

In section 3 we proposed a distinction between evaluation of anaphora resolution approaches and evaluation of anaphora resolution systems. We believe that such a distinction is necessary because it would not be fair to compare the success rate of an approach operating on texts which are perfectly analysed by humans, with the success rate of an anaphora resolution system which has to process the text at different levels before activating its anaphora resolution algorithm. In fact the evaluation of many anaphora resolution approaches focus on the accuracy of resolution algorithms and do not take into consideration the possible errors which inevitably occur in the pre-processing stage. The vast majority of approaches rely on some kind of pre-editing of the text which is fed to the anaphora resolution algorithm; some of the methods have only been manually simulated. As an illustration, Hobbs' naive approach (1976, 1978) was not implemented in its original version. In (Dagan 1990, 1991), (Aone and Bennett 1995) and (Kennedy and Boguraev 1996) pleonastic pronouns are removed manually, whereas in (Mitkov 1998) and (Ferrandez et al. 1997) the outputs of the PoS tagger and the NP extractor/partial parser are post-edited in a similar way to (Lappin and Leass 1994) where the output of the Slot Unification Grammar parser is corrected manually. Finally, Ge at al.'s (1998) and Tetrau'll's approaches (1999) make use of annotated corpus and thus do not perform any pre-processing.

We implemented a fully automatic anaphora resolution system based on our knowledge-poor approach (Orasan, Evans and Mitkov 2000); we also implemented fully automatic versions of Baldwin's as well as Kennedy and Boguraev's approaches (Barbu and Mitkov 2000). Our results provide compelling evidence that fully automatic anaphora resolution is more difficult than previous work has suggested. 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 NP. 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, unknown word recognition, NP extraction, identification of pleo-

---

13 Similar techniques have been used in (Lappin & Leass, 1994).
nastic 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 much below 100%. 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.

We propose the measure **success rate** of anaphora resolution systems which is defined in a similar way as for anaphora resolution algorithms. However, the success rate for anaphora resolution systems reflects in addition to the resolution rate of the algorithm implemented, the overall performance of the system as a whole, and also its ability to carry out successful pre-processing which, includes among other things, the correct identification of noun phrases (which are regarded as candidates of antecedents in the case of nominal anaphora) and the ability to recognise all anaphoric occurrences in the text. The success rate of a specific anaphora resolution system is expressed as the ratio:

\[
\text{success rate} = \frac{\text{Number of successfully resolved anaphors}}{\text{Number of all anaphors}}
\]

where **Number of all anaphors** is all anaphoric occurrences in the evaluation text as identified by humans. This definition assumes that the identification of anaphors (and therefore the identification of non-anaphoric NPs including non-anaphoric pronouns) is the responsibility of the system. Since the pre-processing is expected to be automatic, it is likely that the system may miss some anaphors or candidates for antecedents which will result in a drop in the success rate.

We propose that in addition to measuring the success rate of the anaphora resolution system, it would be useful to calculate the success rate of the anaphora resolution algorithm by running it on perfectly analysed inputs (see Fukumoto, Yamada and Mitkov 2000; see also Table 6, the MAX columns). Such a measure will shed light on the limitation of a specific algorithm, provided the pre-processing was 100% correct.

The measures **non-trivial success rate** and **critical success rate** can be applied to anaphora resolution systems as well. It should be noted, however, that the inequality relations formulated earlier may not hold in a fully automatic processing environment and therefore in this case these measures may not be as indicative as they are for the evaluation of algorithms. As an illustration, consider the scenario when an anaphora resolution system extracts no candidates for an anaphor due to pre-processing errors. The “standard” success rate includes this set and none of them are correctly resolved, resulting in a fall in the success rate. In critical success rate, however, these “always wrong” anaphors are excluded because they do not have more than one candidate after agreement filters have been applied, and so at times there can be a higher score for critical success rate than for “standard” success rate.

Comparison with baseline models is particularly important when evaluating anaphora resolution systems. Table 6 shows the results from comparing MARS with a baseline model which selects as antecedent the most recent NP matching the anaphor in gender and number, and with a baseline model which picks as antecedent a randomly generated NP from the list of candidates.

The question that still remains is how to evaluate systems which are almost ‘automatic’ in the sense that they may involve some (but not full) human intervention – for instance the elimination of anaphors whose antecedents are VPs and other non-NP constituents in the case of anaphora resolution systems that handle nominal anaphora only. One way of ensuring a fair comparison would be to run such systems in a fully automatic mode and provide these results as well.

The fully automatic version of the knowledge-poor approach (MARS) was evaluated on six different files, featuring 52 187 words and 581 anaphoric pronouns (see Table 5). MARS incorporated a module for recognition of pleonastic pronouns as well as recognition of instances of non-nominal anaphoric it.

The overall success rate of (the fully automatic) MARS was 54.65% (323/591). After optimisation (Ora san, Evans and Mitkov 2000), the success rate rose to 62.44% (369/591). Table 6 gives details on the comparative evaluation of MARS as opposed to the original version of the approach, the optimised version, the MAX version (assuming that the input to the anaphora resolution algorithm was 100% correct) and the baseline models picking the most recent and a random candidate, respectively. On this table, the **Original** column presents the success rate of both the optimised and non-optimised versions of the knowledge-poor approach in its original version. The **MARS** column presents the success rate of both the optimised and non-optimised versions when it is run in its full version **Default**, a version in which non-nominal it has been identified (w/o it) and a version in which the agreement filter was switched off (w/o agr). The **MAX** column shows the upper-bound for the success rate due to pre-processing errors. Two baseline models, presented in the **Baseline** column, were evaluated, one in which the most recent candidate was selected as the antecedent and one in which a candidate was selected at random - both after agreement restrictions had been applied.

---

17 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 Paice and Husk (1987) reported 92% for identification of strictly pleonastic it.

18 In Section 5.1.3 we showed that success rate ≥ non-trivial success rate ≥ critical success rate.
### Table 5: The characteristics of the texts used for evaluation of MARS

<table>
<thead>
<tr>
<th>Text</th>
<th># Words</th>
<th># Anaphoric pronouns</th>
<th># Non-nominal anaphoric it</th>
<th># Pleonastic it</th>
<th>Classification accuracy for it</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>9 753</td>
<td>159</td>
<td>5</td>
<td>17</td>
<td>83.97%</td>
</tr>
<tr>
<td>CDR</td>
<td>10 453</td>
<td>82</td>
<td>0</td>
<td>8</td>
<td>89.29%</td>
</tr>
<tr>
<td>BEO</td>
<td>7 493</td>
<td>68</td>
<td>1</td>
<td>23</td>
<td>79.24%</td>
</tr>
<tr>
<td>MAC</td>
<td>15 131</td>
<td>148</td>
<td>0</td>
<td>17</td>
<td>89.65%</td>
</tr>
<tr>
<td>PSW</td>
<td>6 475</td>
<td>76</td>
<td>0</td>
<td>3</td>
<td>93.22%</td>
</tr>
<tr>
<td>WIN</td>
<td>2 882</td>
<td>48</td>
<td>0</td>
<td>3</td>
<td>97.06%</td>
</tr>
<tr>
<td>Total</td>
<td>52 187</td>
<td>581</td>
<td>6</td>
<td>71</td>
<td>87.75%</td>
</tr>
</tbody>
</table>

Table 6: Evaluation of the knowledge-poor approach and its fully automatic, enhanced version MARS

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>MARS</th>
<th>MAX</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-optimised</td>
<td>Optimised</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Files</td>
<td>Base</td>
<td>Opt</td>
<td>Default</td>
<td>w/o it</td>
</tr>
<tr>
<td>PSW</td>
<td>64.55</td>
<td>74.68</td>
<td>72.15</td>
<td>72.00</td>
</tr>
<tr>
<td>MAC</td>
<td>53.93</td>
<td>63.03</td>
<td>61.21</td>
<td>65.33</td>
</tr>
<tr>
<td>WIN</td>
<td>33.33</td>
<td>47.05</td>
<td>45.09</td>
<td>43.10</td>
</tr>
<tr>
<td>ACC</td>
<td>33.33</td>
<td>37.22</td>
<td>36.11</td>
<td>38.80</td>
</tr>
<tr>
<td>CDR</td>
<td>53.84</td>
<td>56.04</td>
<td>59.34</td>
<td>62.02</td>
</tr>
<tr>
<td>BEO</td>
<td>35.48</td>
<td>43.01</td>
<td>36.55</td>
<td>44.20</td>
</tr>
<tr>
<td>PSW + MAC</td>
<td>57.37</td>
<td>65.98</td>
<td>64.75</td>
<td>67.60</td>
</tr>
</tbody>
</table>

7. Reliability of the evaluation results

A major issue in the evaluation of an anaphora resolution algorithm or anaphora resolution system is the reliability of results obtained. One mandatory question that has to be asked is how definitive the evaluation results can be considered. To start with, it has to be pointed out that the majority of anaphora resolution systems report results from tests on one genre only. Next, whether the evaluation is restricted to one genre only or not, the validity of evaluation greatly depends on the size, representativeness and statistical significance of the evaluation corpus. What has emerged is that the evaluation has to cover not hundreds of anaphors but many thousands: it has already been seen that even in the same genre, results may differ if the samples are not large enough (Table 1). Theoretically speaking, the success rate or other evaluation measures could be regarded as definitive only if the approach were tested on all naturally occurring texts, which of course is an unrealistic task. Nevertheless, this consideration highlights the advantages of carrying out the evaluation task automatically. Automatic evaluation requires a large corpus with annotated coreferential links, against which the output of the anaphora resolution systems is to be matched. We have been working actively on the development of large-size coreferentially annotated corpora, with a view to using them in the evaluation process (Mitkov et al. 2000b).

An alternative method would be to employ comprehensive sampling procedures. We are currently experimenting not only with the selection of random samples, but also with selecting them in such a way that no two anaphors are located within a window of 100 sentences. We believe that such a sampling process will produce statistically more significant results.

Finally, the issue as to how reliable or realistic the obtained performance figures are largely depends on the nature of the data used for evaluation. Some evaluation data may contain anaphors which are more difficult to resolve, such as anaphors that are (slightly) ambiguous and require real-world knowledge for their resolution, or anaphors that have a high number of competing candidates, or that have their antecedents far away both in terms of sentences/clauses and in terms of number of ‘intervening’ NPs etc. Therefore it is suggested that in addition to the evaluation results, information should be provided as to how difficult to resolve are the anaphors in the evalua-
tion data. To this end more research is needed to come up with suitable measures for quantifying the average ‘resolution complexity’ of the anaphors in a certain text. In the meantime, simple statistics such as the number of anaphors with more than one candidate, and more generally, the average number of candidates per anaphor, or statistics showing the average distance between the anaphors and their antecedents, would be more indicative of how “easy” or “difficult” the evaluation data is, and should be provided in addition to the information on the numbers or types of anaphors (e.g. intrasentential vs. intersentential) occurring in the evaluation data.

The next section addresses the problem of comparative evaluation in anaphora resolution by postulating that comparison on same data only is insufficient; what also matters is comparison on the basis of the same pre-processing tools.

8. A way forward: evaluation workbench for anaphora resolution

In order to secure a fair, consistent and accurate evaluation environment, and to address the problems identified above, we developed an evaluation workbench for anaphora resolution which allows the comparison of anaphora resolution approaches sharing common principles (e.g. POS tagger, NP extractor, parser). The workbench enables the ‘plugging in’ and testing of anaphora resolution algorithms on the basis of the same pre-processing tools and data. This development is a time-consuming project, given that we have to re-implement most of the algorithms but it is expected to produce a better picture as to the advantages and disadvantages of the different approaches. Developing our own evaluation environment (and even re-implementing some of the key algorithms) also alleviates the formidable difficulties associated with obtaining the codes of the original programs. Another advantage of the evaluation workbench can be seen in the fact that all approaches incorporated operate in fully automatic mode.

The current version of the evaluation workbench\(^\text{19}\) employs one of the best available ‘super-taggers’ in English - Conexor’s FDG Parser (Tapanainen and Jarvinen 1997). This super-tagger provides information on the dependency relations between words which allows the extraction of complex NPs. It also gives morphological information and the syntactic roles of words. Although FDG does not provide the identification of the noun phrases in the text, the dependencies established between words have served in the building of a noun phrase extractor and selecting only the anaphoric expressions (instances of pleonastic \textit{it} are removed). Each entry in this list consists of a record containing the following: the word form, the lemma of the word or of the head of the noun phrase, the starting position in the text, the ending position in the text, the part of speech, the grammatical function, the index of the sentence that contains the candidate and the index of the verb whose argument is the candidate. The list of candidates is implemented as a binary tree for optimum access.

The workbench incorporates an automatic scoring system that operates on an SGML input file where the correct antecedents for every anaphor have been marked. The annotation scheme recognised by the system at this moment is MUC, but support for the MATE annotation scheme is being developed. The results are visually displayed on the screen and they can also be saved on file. For easier visual comparison, each anaphor is displayed in parallel with the antecedents proposed by each of the algorithms.

Three approaches that have been extensively cited in the literature were first selected for comparative evaluation by the workbench: Kennedy and Boguraev’s parser-free version of Lappin and Leass’ RAP (Kennedy and Boguraev 1996), Baldwin’s pronoun resolution method Cogniac which uses limited knowledge (Baldwin 1997) and Mitkov’s knowledge-poor pronoun resolution approach (Mitkov 1998). All three of these algorithms share a similar pre-processing methodology: they do not rely on a parser to process the input and use instead POS taggers and NP extractors; none of the methods make use of semantic or real-world knowledge. Kennedy and Boguraev’s and Baldwin’s algorithms were re-implemented, and the standard, non-optimised version of MARS was used to represent Mitkov’s algorithm. Since the original version of Cogniac is non-robust and resolves only anaphors that obey certain rules, for fairer and comparable results the ‘resolve-all’ version as described in (Baldwin 1997) was implemented. Both Kennedy and Boguraev’s and Baldwin’s approaches benefit from Evans’ (2000) program for identifying and filtering instances of non-nominal anaphora (which includes occurrences of pleonastic pronouns).

The comparative evaluation was based on a corpus of technical texts that was manually annotated for coreference. The corpus contains more than 50 000 words, with 19 305 noun phrases and 484 anaphoric pronouns. The files that were used are: “Beowulf HOW TO” (referred in Table 7 as BEO), “Linux CD-Rom HOW TO” (CDR), “Macintosh Help file” (MAC), “Portable StyleWriter Help File” (PSW), “Windows Help file” (WIN).

Table 7 shows the success rate of the three anaphora resolution algorithms on a set of the above files. The overall success rate calculated for the 426 anaphoric pronouns found in the texts was 62.5% for MARS, 59.02% for Cogniac and 63.64% for Kennedy and Boguraev’s method.

\(^{19}\) To a certain extent, the critical success rate addresses this issue in the evaluation of anaphora resolution algorithms by providing the success rate for the anaphors that are more difficult to resolve.

\(^{20}\) Implemented by Catalina Barbu.
<table>
<thead>
<tr>
<th>File</th>
<th>Number of pronouns</th>
<th>MARS</th>
<th>Cogniac</th>
<th>Kennedy&amp;Boguraev</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSW</td>
<td>77</td>
<td>79.74</td>
<td>72.1</td>
<td>79.8</td>
</tr>
<tr>
<td>MAC</td>
<td>148</td>
<td>66.06</td>
<td>60.8</td>
<td>67.1</td>
</tr>
<tr>
<td>WIN</td>
<td>51</td>
<td>56.86</td>
<td>55.9</td>
<td>58.7</td>
</tr>
<tr>
<td>BEO</td>
<td>67</td>
<td>45.16</td>
<td>45.0</td>
<td>46.3</td>
</tr>
<tr>
<td>CDR</td>
<td>83</td>
<td>64.83</td>
<td>61.3</td>
<td>66.3</td>
</tr>
<tr>
<td>Total</td>
<td>426</td>
<td>62.53</td>
<td>59.02</td>
<td>63.64</td>
</tr>
</tbody>
</table>

Table 7: Comparative evaluation carried out with the help of the evaluation workbench

Besides the evaluation system, the workbench also incorporates a basic statistical calculator of the anaphoric occurrences in the input file. The parameters calculated are: the total number of anaphors, the number of anaphors in each morphological category (personal pronoun, noun, reflexive, possessive), the number of inter- and intrasentential anaphors and average number of candidates per anaphor. More details on the current implementation of the evaluation workbench are reported in (Barbu and Mitkov 2000).

While the workbench is currently based on the FDG shallow parser\textsuperscript{21}, we plan to update the environment in such a way that two different modes will be available: one making use of a shallow parser (for approaches operating on partial analysis) and one employing a full parser (for algorithms making use of full analysis). Future versions of the workbench will include access to semantic information (WordNet) in order to accommodate approaches incorporating knowledge of this type.

9. Conclusion

The paper has argued that the evaluation of anaphora resolution algorithms and the evaluation of anaphora resolution systems should be done separately. To this end the simple measure of success rate has been proposed to counter the inadequacies of recall and precision in the case of robust algorithms. A package of resolution rate measures and benchmark models was proposed to provide a better picture of performance.

The paper has also argued that the only way in which anaphora resolution systems can be consistently evaluated is to develop an evaluation workbench which offers the use of the same pre-processing tools to approaches based on similar philosophies of text analysis.

10. Acknowledgments

The author expresses his gratitude to Amit Bagga, Richard Evans, Constantin Orasan and Catalina Barbu for their comments on an earlier version of this paper.

11. References


\textsuperscript{21} It is worth mentioning that experiments with different pre-processing tools are underway with a view to achieving optimal performance. At the time of completing this article the Xerox language processing suite was being incorporated as an alternative pre-processing option.


Tapanainen, P. and T. Jarvinen. 1997. “A non-projective Dependency Parser”. In the Proceedings of the 5th Conference of Applied Natural Language Processing, 64-71, ACL, USA.