In Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC2), Lancaster, 1998, pp. 15-27.

# Extending a Simple Coreference Algorithm with a Focusing Mechanism

*Saliha Azzam, Kevin Humphreys, Robert Gaizauskas*
Department of Computer Science, University of Sheffield

## 1 Introduction

We compare two approaches to anaphora resolution: one based on a focusing mechanism [12] and a simpler approach that relies solely on semantic similarity of potential coreferents and distance restrictions between anaphor and antecedent. Both approaches have been implemented and quantitatively evaluated against a substantial corpus of texts which have been annotated with coreference relations as part of the Message Understanding Conference coreference task [3] and implemented within the general framework provided by the LaSIE (Large Scale Information Extraction) system [7, 9], Sheffield University's entry in the MUC-6 and 7 evaluations. The simple approach was reported on at DAARC-1 [6], at which time we claimed that the flexibility of the framework in which we carried out research on coreference facilitated experimentation with alternative techniques. Given this claim and the interest exhibited in focus-based approaches to anaphora resolution at DAARC-1 we decided to extend the simple approach with a representation of focus and compare it with the previous approach.

In this paper we present results about an evaluation of both approaches on real-world texts to determine the main drawbacks and advantages of each. We see this as a first step towards answering the question: does adding a notion of focus to a simple coreference mechanism buy you anything?

## 2 Coreference in LaSIE

The LaSIE system has been designed as a general purpose IE system which can conform to the MUC task specifications for named entity identification, coreference resolution, IE template element and relation identification, and the construction of scenario-specific IE templates (see [4] for a detailed description of the MUC-7 task definitions). The system is a pipeline architecture consisting of separate modules for tokenisation, sentence splitting, part-of-speech tagging, morphological stemming, domain-specific lexical lookup, parsing with semantic interpretation, intratextual proper name matching, and discourse interpretation. The latter stage constructs a discourse model, based on a predefined domain model, using the, often partial, semantic analyses supplied by the parser.

The domain model represents a hierarchy of domain-relevant concept nodes, together with associated properties. It is expressed in the XI formalism [5] which provides a basic inheritance mechanism for property values and the ability to represent multiple classificatory dimensions in the hierarchy. Instances of concepts mentioned in a text are added to the domain model, populating it to become a text-, or discourse-, specific model.

The basic coreference mechanism in LaSIE takes a set of instances newly added to the discourse model, and compares each one with a set of instances already in the discourse model. For object (or nominal) coreference, proper names, pronouns, and common nouns are handled separately, first attempting intra-sentential coreference for each set, and then inter-sentential coreference. The fundamental algorithm compares the semantic classes and attributes of each pair of new-old instances, and, if compatible, calculates a similarity score for the pair, based on the distance between the instances' parent classes in the concept hierarchy, and the number of shared properties. The highest

scoring pair, for each new instance, if any are compatible, is merged in the discourse model, deleting the instance with the least specific class in the ontology, and combining the properties of both instances. If more than one pair have an equal highest score, the pair with the closest realisation in the text is selected. This mechanism is basically unchanged from the LaSIE-I MUC-6 system, as reported in DAARC-1 [6].

## 2.1 LaSIE-II Coreference Restrictions

The fundamental resolution algorithm can be extended arbitrarily through the use of a special `distinct` attribute, associated with particular concept nodes in the ontology, and evaluated during the comparison of each instance pair. If a particular instance pair can inherit this attribute, further comparison is abandoned and the next pair tried. The `distinct` attributes implement various positional, syntactic, and semantic restrictions on the underlying 'eager' resolution algorithm ('eager' in the sense that two entities that have semantically compatible types and attributes *will* be coreferred unless something – such as a constraint implemented via the `distinct` attribute – prevents this). Some of the restrictions reflect general purpose linguistically motivated principles, and others reflect specific MUC guidelines for certain classes of potential anaphora

LaSIE-I used the `distinct` attribute to implement general restrictions for avoiding the attempted resolution of indefinite noun phrases and pleonastic pronouns (like *it* in *it is unbelievable*), preventing unresolved pronouns from becoming antecedents, etc., and MUC-specific restrictions for noun phrases acting as qualifiers, embedded proper nouns, etc. The following sections describe the main restrictions added to LaSIE-II for the MUC-7 evaluation, in addition to those present in LaSIE-I (not repeated here – see [6]).

### 2.1.1 General restrictions

1. *Long Distance Coreference* By 'long distance coreference' we mean cases where an anaphor can find its antecedent in a much earlier paragraph. In LaSIE-I antecedents for pronouns and bare nouns were sought only in the current and previous paragraphs, and no attempt was made to find an antecedent in earlier paragraphs even if the anaphor almost certainly required one, as, for example, in the case of pronouns. LaSIE-II was extended to search for potential antecedents in successively earlier paragraphs until a compatible one is found.

2. *Cataphora* LaSIE-II handles some cases of cataphora, involving pronouns occurring in the text before their antecedents. Specifically, two cases:

   (a) where pronouns occur in quotations, as in:

   > *"I caught Reggie when he was much younger counting his dad's trophies," McNair said.*

   where *I* corefers forwards with *McNair*. The coreference succeeds only if the quoted sentence is processed completely by the parser.

   (b) copular constructions, such as *he is a president*, where *he* corefers forwards with *president*, or *This is a mystery* where *This* corefers with *mystery*.

### 2.1.2 MUC-specific restrictions

1. *Co-ordinated NPs* One of the changes in the coreference task specification from MUC-6 to MUC-7 was to allow certain cases of conjoined NPs to become markables. Hence, coordinated or conjoined noun phrases are now taken into account in the LaSIE-II coreference algorithm. A new instance that represents the set of the coordinated instances is created in the discourse model and becomes a potential antecedent. The new set instance will have the attribute `plural` and have the semantic class of the coordinated instances if they are from the same class, or the lowest common parent class otherwise. For example, in:

   > *Bruce and his boys were not in the habit of sharing their feelings about each other.*

three instances *Bruce* (`e1`), *his boys* (`e2`), and the set *Bruce and his boys* (`e3`) are represented in the discourse model. The latter is given the semantic class `person` and the attribute `plural`. The pronoun *their* can then corefer with the set (`e3`), and *his* with *Bruce* (`e1`).

The system may, however, generate spurious coreference or fail to corefer coordinated noun phrases, if the parser fails to correctly recognise the coordinated phrases on which the identification of a set instance relies.

2. *Copular Constructions* Copular constructions of the type *NP1 'is' NP2*, where *NP1* should corefer with *NP2*[1] were not dealt with in LaSIE-I simply due to a lack of development time. These are now taken into account by the coreference algorithm, as in the following example:

   *The F14 "Tomcat" is the Navy's first-line fighter aircraft.*

   where *the Navy's first-line fighter aircraft* will corefer with *The F14 "Tomcat"*.

   Taking copular constructions into account necessitated reviewing all the coreference rules, to discover where exceptions needed to be made. For example, in a non-copular construction an indefinite noun phrase like *a president* cannot corefer backward, while in a copular construction it can. Thus, the coreference rule for indefinites needed to be relaxed to exclude copulars. (The same rule was also relaxed for appositions, to handle cases such as:

   *They bought the plane, a 1975 single-engine Cessna 177B Cardinal, about nine years ago.*

   where the indefinite *a 1975 single-engine Cessna* corefers with *the plane*.)

   Another important aspect of copulars is that they provide information that allows 'unknown' words, i.e. words whose semantic type is not known, to be classified in our ontology during processing. This is possible when the unknown word occurs as one argument of the verb *to be* and the semantic class of the other argument is present in the ontology. For example, in *Bill is a president*, if *president* is not known in the ontology and *Bill* is a known person name, then `president` is added automatically as a new class below the `person` node, with *Bill* as an instance. This can make subsequent coreference more accurate by preferring or preventing coreference with instances of the newly added class (for example, once *president* has been added as a subclass of *person*, subsequent occurrences of the impersonal pronoun *it* in the text would be prevented from coreferring with occurrences of *the president*).

3. *Bare Nouns* One of the more difficult types of MUC coreference is that of bare nouns (nouns without determiners). We distinguish between bare nouns acting as (prenominal) modifiers, for which the coreference rules were similar to LaSIE-I, and modified and unmodified head nouns. The latter 2 classes require the addition of distinct sets of restrictions.

# 3 Extending pronoun resolution with a focusing mechanism

The fundamental approach to coreference resolution, as used in LaSIE-I and LaSIE-II and described in the previous section, performs well, but as one would expect with a simple mechanism based solely on semantic compatibility and recency there are cases where inappropriate antecedents are selected. Consider the following sentence:

*Because of the low liability limit, many airlines are voluntarily paying more, either to spare themselves a lengthy trial at which families will try to show negligence; which allows higher damages; or for general good will, or because they do not think the limits will hold up in court.*

[1] These constructions are discussed in the section 'Predicate Nominals and Time-dependent Identity' in the MUC-7 coreference task definition [4].

Using only semantic compatibility plus recency, *they* will be coreferred with *damages*, if no information is known about the animacy of the arguments of the verb *think*, otherwise with *families*. The correct antecedent is, of course, *many airlines*.

This example suggests that some mechanism is needed to detect which entity the sentence is about, i.e. the *focus* or *center* (*many airlines* in the example) and to bias pronouns to corefer with it. We have experimented with plugging in a focus-based approach based on [2], that provides such a mechanism for pronoun resolution. This approach is described in the rest of this section and the following section describes corpus-based experiments we have carried out to assess to what extent the focussing mechanism can offer an improvement over the fundamental approach.

## 3.1 Focus in Anaphora Resolution

The term *focus*, along with its many relations such as *theme*, *topic*, *center*, etc., reflects an intuitive notion that utterances in discourse are usually 'about' some thing. For anaphora resolution, stemming from Sidner's work [12], focus has been given an algorithmic definition and a set of rules for its application. Sidner's approach is based on the claim that anaphora generally refer to the current discourse focus, and so modelling changes in focus through a discourse will allow the identification of antecedents. The algorithm makes use of several *focus registers* to represent the current state of a discourse, in particular the *current focus* (*CF*) register. A set of *Interpretation Rules* (*IR*s) applies whenever an anaphor is encountered, proposing potential antecedents from the registers, from which one is chosen.

An important limitation of Sidner's algorithm, noted by [3], [2] and [10], among others, is that the focus registers are only updated after each sentence. Thus antecedents proposed for an anaphor in the current sentence will always be from the previous sentence or before. Intrasentential references are therefore impossible. A related difficulty is that no antecedent will be proposed for an anaphor in the first sentence of a discourse, since the focus registers will always be empty at this point. Solutions for these problems have been proposed in [2], and we base our implementation on this account.[2]

## 3.2 Implementing Focus-Based Pronoun Resolution in LaSIE

Integration of the focus-based algorithm proposed in [2] into LaSIE proved straightforward, taking advantage of various mechanisms already available as part of the fundamental algorithm. In this approach *elementary events* (*EE*s, effectively simple clauses) are used as the basic processing units, rather than sentences, and updating the focus registers and applying *IR*s for pronoun resolution then takes place after each *EE*, permitting intrasentential references. In addition, an initial 'expected focus' is determined based on the first *EE* in a text, providing a potential antecedent for any pronoun within the first *EE*. The resolution algorithm is as follows:

```
for each sentence:
1. split the semantic representation into EEs
2. for each EE:
  a. if 1st EE of 1st sentence,
       initialise focus registers
       (apply 'expected focus' algorithm)
  b. for each pronoun
       apply IRs
  c. update focus registers
     (apply 'focusing' algorithm)
```

The expected focus algorithm selects an initial focus, the 'expected focus', generally the *theme* of the first *EE*, where this is either the object of a transitive verb, or the subject of an intransitive or the copula (following [8]).

[2]This account has also been previously implemented in the COBALT system [1], which was based on very different discourse and world knowledge representations than those of LaSIE.

The focusing algorithm updates the focus registers that represent the current state of a discourse: *CF*, the current focus; *AFL*, the alternate focus list, containing other candidate foci; and *FS*, the focus stack, containing previous *CF*'s. A parallel structure to the *CF*, *AF*, the actor focus, is also set to deal with agentive pronouns, together with *AFS*, the actor focus stack, used to record previous *AF*'s, and so allow a separate set of *IR*s for *agent* pronouns (animate verb subjects like *he* or *she*). Also *Intra-AFL*, the intrasentential alternate focus list, is used to record candidate foci from the current *EE* only. The algorithm updates these registers after each *EE*, confirming or rejecting the current focus.

Pronoun resolution uses the state of the focus registers and a set of *IR*s associated with each pronoun type to determine which element of the focus registers is the antecedent. Each IR suggests one or several antecedents depending on the focus and on the pronoun type. Pronouns are divided into three main classes, each with a distinct set of *IR*s proposing antecedents:

1. Personal pronouns acting as agents (animate subjects): the *IR*s propose antecedents from the 'agentive' registers, *AF* initially then animate members of *AFL* and *AFS* . The pronoun *he* in *Shotz said he knew the pilots* belongs to this class.
2. Non-agent pronouns: the *IR*s firstly propose the *CF* as antecedent, followed by the members of the *AFL* and *FS*. The pronoun *it* in the second sentence of the extended example below belongs to this class.
3. Possessive, reciprocal and reflexive pronouns (*PRR*s): for this class the *IR*s propose antecedents from the *Intra-AFL*, thus allowing intra-EE references. The pronoun *their* in *the brothers flew to Block Island and were on their way home* belongs to this class.

Antecedents proposed by the *IR*s are accepted or rejected based on their semantic type and feature compatibility, using the semantic and attribute value similarity scores of the fundamental LaSIE coreference mechanism.

We now illustrate the focus-based pronoun resolution algorithm, as it is implemented, by stepping through the processing of each EE from the following example:

> *The Russian airline Aeroflot has been hit with a writ for loss and damages.*
>
> *All 75 people on board the Aeroflot Airbus died when it ploughed into a Siberian mountain in March 1994.*
>
> *Aeroflot general manager for Hong Kong said on Tuesday he was unaware the writ had been filed.*
>
> *The writ is for "damages, interest and costs" of seven passengers who died. It claims the deaths were "caused by negligence."*

**EE-1** *The Russian airline Aeroflot has been hit with a writ for loss and damages*

The expected focus (theme) is *Russian airline Aeroflot*. The focusing algorithm initialises the registers. *Intra-AFL* is first initialised with all (non-pronominal) candidate foci in the *EE*:

*Intra-AFL* = `a writ, loss and damages`
All other registers, the *AFL* (alternate focus list), *AF* (actor focus), *FS* (focus stack) and *AFS* (actor focus stack), are unaffected by the expected focus and remain empty. There are no pronouns in EE-1 so no IRs apply. *Intra-AFL* is then added to the current *AFL*, as it is after each *EE* of the same sentence has been processed. The state of the registers is then:

*CF* (current focus) = `Russian airline Aeroflot`
*AFL* = `a writ, loss and damages`[3]

[3]Here conjunctions are considered as a single unit in the registers, i.e, 'A and B'. Alternatively, or additionally, the units 'A', 'B' can be represented separately, but for our implementation we currently only handle the compound case. Similar problems apply in the case of compound nouns, like 'Russian airline Aeroflot', where 'airline' and 'Aeroflot' can also be represented separately.

**EE-2** *All 75 people on board the Aeroflot Airbus died*

*Intra-AFL* is reinitialised with candidate foci from this *EE*:
*Intra-AFL* = `board, Aeroflot Airbus`

No pronouns occur in EE-2 so no *IR*s apply. The focusing algorithm initialises the *AF* to *75 people* (as the agent of EE-2), the *CF* remains unchanged, the *AFL* is reset at this point with the elements of the *Intra-AFL*, to take into account the elements of the new sentence only. The state of the registers is then:

*CF* = `Russian airline Aeroflot`
*AF* = `75 people`
*AFL* = `board, Aeroflot Airbus`

**EE-3** *when it ploughed into a Siberian mountain in March 1994.*

*Intra-AFL* = `Siberian mountain, March 1994`

*IR*s propose *Aeroflot Airbus*, member of the AFL, as the antecedent of *it* (we assume here that semantic restrictions, e.g. subcategorisation patterns of the verb *plough*, rule out the *CF*, *Russian airline Aeroflot*, proposed first as the antecedent, and *board*, the first item in the *AFL*). After the focusing algorithm applies, *Aeroflot Airbus* then becomes the new *CF* and the old one is added to the *FS*:

*CF* = `Aeroflot Airbus`
*FS* = `Russian airline Aeroflot`
*AF* = `75 people`
*AFL* = `board, Aeroflot Airbus,Siberian mountain, March 1994`

**EE-4** *Aeroflot general manager for Hong Kong said on Tuesday*

*Intra-AFL* = `Aeroflot general manager, Hong Kong, Tuesday`

EE-4 does not contain pronouns. The focusing algorithm updates the focus registers and changes the *AF* to the current agent of EE-4, *Aeroflot general manager*, given that no reference is made to the current *AF*. The old *AF* is then added to the *AFS*. *CF* remains unchanged. *AFL* is again reset for the new sentence:

*CF* = `Aeroflot Airbus`
*FS* = `Russian airline Aeroflot`
*AF* = `Aeroflot general manager`
*AFS* = `75 people`
*AFL* = `Aeroflot general manager,Hong Kong,Tuesday`

**EE-5** *he was unaware*

*Intra-AFL* is empty.

*IR*s corefer the pronoun *he* with the current *AF*. The focusing algorithm keeps then the same *AF*. *CF* remains unchanged too as the current *EE* lacks a new theme:

*CF* = `Aeroflot Airbus`
*FS* = `Russian airline Aeroflot`
*AF* = `Aeroflot general manager`
*AFS* = `75 people`
*AFL* = `Aeroflot general manager, Hong Kong, Tuesday`

**EE-6** *the writ had been filed*

*Intra-AFL* = `the writ`

EE-6 does not contain pronouns so no *IR*s apply. There is a new theme, *the writ*, that replaces then the current *CF*. The old *CF* is added to the *FS*, *AF* remains unchanged, and *Intra-AFL* is added to *AFL*:

*CF* = `the writ`
*FS* = `Aeroflot Airbus, Russian airline Aeroflot`
*AF* = `Aeroflot general manager`
*AFS* = `75 people`
*AFL* = `the writ, Aeroflot general manager,Hong Kong,Tuesday`

**EE-7** *The writ is for "damages, interest and costs" of seven passengers*

*Intra-AFL* = `the writ, damages, interests and costs, seven passengers`

There are no pronouns. The focusing algorithm keeps the same *CF*, *the writ*, that occurs as a theme in EE-7. The *AFL* is reset at this point, with the elements of the *Intra-AFL*, because EE-7 starts a new sentence:
*CF* = `the writ`
*FS* = `Aeroflot Airbus, Russian airline Aeroflot`
*AF* = `Aeroflot general manager`
*AFS* = `75 people`
*AFL* = `the writ, damages, interests and costs, seven passengers`

**EE-8** *who died*

*Intra-AFL* is empty. No pronouns occur in EE-8. The focusing algorithm changes *AF* to *seven passengers*, the current agent. The state of the registers is then:

*CF* = `the writ`
*FS* = `Aeroflot Airbus, Russian airline Aeroflot`
*AF* = `seven passengers`
*AFS* = `Aeroflot general manager, 75 people`
*AFL* = `the writ, damages, interests and costs`

**EE-9** *It claims*

*Intra-AFL* is empty.

*IR*s suggest the *CF*, *the writ*, as the antecedent for the pronoun *it*. All the registers remains the same, except that *AFL* is reinitialised:

*AFL* = `the writ`

**EE-10** *the deaths were "caused by negligence."*

No pronouns occur is EE-10. The registers change to the following:

*CF* = `the deaths`
*FS* = `the writ, Aeroflot Airbus, Russian airline Aeroflot`
*AF* = `seven passengers`
*AFS* = `Aeroflot general manager, 75 people`
*AFL* = `the writ, negligence`

# 4 Evaluation and Comparative Study

Several evaluations of the fundamental and focus-based pronoun resolution algorithms have been carried out with the system in various configurations. This section describes the system configurations evaluated, the corpora and metrics that were used for the evaluation and the results of the evaluation.

## 4.1 System Configurations

As a baseline measure the fundamental algorithm was tested alone (level 0), with no `distinct` properties used as heuristics to rule out any proposed semantically compatible antecedents. The contribution of the general purpose (non-domain/corpus specific) set of coreference restrictions was then tested (level 1) by adding `distinct` properties to test for pleonastic pronouns, indefinite noun phrases, etc. Finally, rules developed specifically for the MUC-7 coreference task (e.g. rules pertaining to bare nouns) and rules derived from observing the MUC-7 training data (e.g. rules preventing pronouns from coreferring with date expressions) were added (level 2); this is the system we actually used for MUC-7.

The focus-based algorithm differs from the fundamental algorithm only with respect to the treatment of pronouns. However, while for non-pronominal coreference its behaviour is identical to the fundamental algorithm, results may differ even for non-pronominal coreferents, since pronouns may form bridges or add properties (such as animacy or gender) in coreference chains that promote or prevent links between non-pronominal noun phrases. Further, the focus-based algorithm makes use of the compatibility tests carried out within the fundamental algorithm, and so its behaviour with respect to pronominal coreference is affected by the set of restrictions currently in use. Because of these interaction effects, we evaluated the focus-based algorithm at the same three levels: level 0, with no `distinct` properties in force; level 1 with general restrictions; and level 2 with both general and MUC-7-specific restrictions.

A final system configuration that we tested was the fundamental algorithm with full restrictions (level 2), but with pronoun resolution disabled altogether. This gives some indication of the significance of pronouns in the overall coreference task. However, because of various interaction effects this measure cannot be used as a straightforward baseline against which the fundamental and focus-based algorithms may be compared to see how well these two algorithms handle pronouns respectively.

## 4.2 Corpora

Two corpora were used for evaluation, both supplied as part of the MUC-7 coreference evaluation task. Both consisted of newswire articles from the *New York Times News Service*. Each article was supplied in raw form and in an annotated form, in which coreference relations had been manually added to the text as SGML tags according to the MUC-7 coreference annotation guidelines [4] (coreference relations are marked within single articles only).

The first corpus (Tr) was the training corpus used by MUC-7 participants for system development. It consists of 60 articles ranging in length from about 300 to 3000 words, averaging around 800 words. Our analysis reveals that the manually annotated texts, or 'keys' in MUC terminology, contain 5785 annotated referring expressions in 1427 coreference chains, with 1060 of the strings being pronouns (about 18%).

The second corpus (Eval) was the evaluation corpus which formed the blind data used in the final evaluation. We have continued to keep this data blind. It consists of 20 articles containing 1699 annotated referring expressions in 409 coreference chains, with 248 of the strings being pronouns (about 14.5%). The texts are from the same source and are of comparable length.

## 4.3 Metrics

The MUC-7 scoring software was used to calculate the metrics of 'recall' and 'precision', together with an overall $f$-measure. In general terms, recall is a measure of how much of what was to be identified the system has identified, precision is a measure of how much of what the system proposes is correct, and the $f$-measure is a single measure combining these two[4]. While these metrics may be straightforwardly applied in tasks like named entity recognition and template slot filling, coreference poses interesting differences, since what are to be found are equivalence classes

[4] $f$-measure is calculated according to the formula $F = \frac{(\beta^2+1)\times P\times R}{(\beta^2\times P)+R}$, where $\beta$ allows for a relative weighting of precision (P) and recall (R), and is set to 1 for equal weighting.

of referring expressions. These are represented by chains of linked text strings, but the same set of referring expressions may be linked in a variety of ways, each expressing equivalent information, but not directly comparable.

In the context of coreference scoring, precision and recall metrics may be defined informally as follows (see [13] for details). The key (manually annotated document) and response (system generated markup) each consist of a number of coreference chains. Each chain may be thought of as inducing an equivalence class, which is just the set of all members of the chain. Each key equivalence class is partitioned by the response equivalence classes into subsets that intersect with response equivalence classes and additionally those elements that do not occur in any response equivalence class. Recall error for each key equivalence class is just the number of "missing" links that need to be added to join up this partition and recall is the number of links defining the key equivalence class minus the recall error, normalised by the number of links defining the key equivalence class. For the whole test set, the recall is the sum for all key equivalence classes of the the number of links defining the class minus the recall error for that class all divided by the sum of the number of links defining all the key equivalence classes. Precision is defined symmetrically, by partitioning each response equivalence class by the key equivalence classes and seeing how many missing links need to be added to make the key up to the response (these may be viewed as the links the response has added incorrectly).

While these metrics provide measures of overall coreference performance, they do not allow more focussed assessment of performance on subclasses of coreference phenomena, such as pronominal anaphora. Since the MUC annotation scheme does not distinguish between different classes of anaphora (pronouns, definite noun phrases, bare nouns, and proper nouns) there is no way to directly score pronoun coreference. But even if these classes were distinguished, it is not clear how one should adapt the MUC coreference scoring procedure to evaluate pronominal coreference alone. One proposal would be to filter the key to include only coreference chains containing one or more pronouns; one could then filter the response to contain only chains which intersect with some chain in this filtered key and score this modified key and response. However, what this in fact measures is how well a system scores against chains containing pronouns, and, if one is interested in how well pronoun coreference is carried out, has the counter-intuitive consequence that a system that resolved no pronouns could still get non-zero precision and recall scores. Another possibility would be to score as correct only those pronoun-containing response chains whose elements form a subset of some key chain (so, in the response the pronoun only participates in correct links). This is a stricter metric than the general MUC coreference metric, which gives partial credit to overlapping key and response chains. Yet another option would be to score as correct any pronoun-containing response chain which in addition to the pronoun contained at least one element which occurs together with the pronoun in the same key chain (so, in the response the pronoun participates in at least one correct link). This is a more indulgent measure and is arguably too weak. Each of these three proposed metrics for evaluating pronoun coreference has problems, but each also gives at least partial information about how well pronoun coreference is being carried out and hence can be fruitfully used to compare different systems.

For various technical reasons we have adopted the third measure discussed above: it can be computed relatively straightforwardly from certain files which the MUC scorer produces while scoring all coreferences. As a byproduct of the scoring process the MUC scorer produces 'partition files' which contain the results of partitioning the key and response chains against each other. Since pronouns form a closed class, it is possible to examine these files directly to ascertain various facts about the treatment of pronouns. Counting the number of response-induced partitions of a key chain which contain a pronoun and at least one other element and dividing by the total number of pronouns in all key chains gives what we can call a *weak measure of pronoun recall*; dividing instead by the total number of pronouns in all response chains gives what we can call a *weak measure of pronoun precision*. Re-expressing this informally, weak pronoun recall is the proportion of coreferential pronouns the system manages to correctly link to at least one expression and weak pronoun precision is the proportion of pronouns proposed by the system as coreferential which are correctly linked to at least one expression.

## 4.4 Results

The results obtained for the fundamental and focus-based algorithms in each configuration against both the Tr and Eval corpora are given in Table 1.

| Configuration | Recall | | Precision | | f-measure | |
|---|---|---|---|---|---|---|
| | Tr | Eval | Tr | Eval | Tr | Eval |
| Fundamental (level 0) | 61.6 | 58.8 | 57.9 | 56.1 | 59.7 | 57.4 |
| Focus (level 0) | 59.9 | 55.5 | 57.3 | 51.1 | 58.6 | 53.2 |
| Fundamental (level 1) | 60.3 | 57.7 | 63.3 | 61.0 | 61.8 | 59.3 |
| Focus (level 1) | 58.0 | 54.8 | 62.6 | 60.5 | 60.2 | 57.5 |
| Fundamental (level 2) | 58.2 | 56.0 | 71.3 | 70.2 | 62.3 | 64.1 |
| Focus (level 2) | 55.4 | 53.3 | 69.8 | 69.7 | 61.8 | 60.4 |

Table 1: Results of Fundamental and Focus-based Algorithms

The results in Table 1 show better performance for the non-focus based approach against both corpora at all levels and on all metrics. Looking across the three levels at which the algorithms were evaluated we see that adding an extra 'layer' of linguistic rules (level 1), shows a significant increase in precision for both approaches with only a slight drop in recall. And, as expected, the addition of MUC specific rules (level 2) benefits both approaches. Again there is significant improvement in precision (an increase of approximately 9% for both approaches for unseen texts, and approximately 8% for the training texts) with only marginal decrease in recall.

Concentrating, now on pronominal coreference alone, we observe that running the fundamental algorithm with full restrictions (level 2) but with pronoun resolution disabled altogether gives the figures in Table 2. These figures show that while precision goes up slightly by not attempting pronoun coreference (2-4%), recall drops substantially (11-16%) leading to an f-measure drop of about 10%.

| Corpus | Recall | Precision | f-measure |
|---|---|---|---|
| Tr: | 42.4 | 73.6 | 52.6 |
| Eval: | 44.7 | 73.9 | 55.7 |

Table 2: Results of Disabling Pronoun Coreference in the Fundamental Algorithm

As described above in section 4.3, we can define 'weak' measures of pronoun coreference evaluation by how many system response chains containing a pronoun correctly link that pronoun to at least one other referring expression in the text. Results for these measures are presented in Table 3. Again on these measures we see that the focus-based algorithm fares worse than the fundamental algorithm.

To attempt to get some idea of the extent to which the focus-based algorithm is addressing a subset of the cases the fundamental algorithm is addressing, and to what extent it is addressing different cases we once again examined the partition files produced as a byproduct of the MUC-7 coreference scoring process, and this time attempted to see how many of the pronouns occurring in response chains for one approach also occurred in the other. An initial analysis of the differences between the scorer output for pronouns reveals that, for the level 2 systems:

- the focus-based algorithm proposes antecedents for 2 pronouns in Tr and 1 pronoun in Eval,

| Configuration | Weak Pronoun Recall | | Weak Pronoun Precision | |
|---|---|---|---|---|
| | Tr | Eval | Tr | Eval |
| Fundamental (level 0) | 72.9 | 66.5 | 66.9 | 62.0 |
| Focus (level 0) | 64.9 | 53.2 | 63.7 | 53.0 |
| Fundamental (level 1) | 76.0 | 71.7 | 71.3 | 66.9 |
| Focus (level 1) | 65.0 | 54.8 | 68.2 | 61.5 |
| Fundamental (level 2) | 77.5 | 73.0 | 73.0 | 68.6 |
| Focus (level 2) | 64.7 | 53.2 | 68.6 | 61.1 |

Table 3: 'Weak' Pronoun Coreference Evaluation

for which the fundamental algorithm fails to propose any antecedent at all.

- the fundamental algorithm proposes antecedents for 117 pronouns in Tr and 44 pronouns in Eval, for which the focus-based algorithm fails to propose any antecedent at all.
- the focus-based algorithm proposes antecedents for 3 pronouns in Tr and 1 pronoun in Eval, for which no antecedent exists (according to the key) and for which the fundamental algorithm did not propose an antecedent.
- the fundamental algorithm proposes antecedents for 15 pronouns in Tr and 6 pronouns in Eval, for which no antecedent exists (according to the key) and for which the focus-based algorithm did not propose an antecedent.

Thus, the response for the focus-based approach differed primarily by omitting cases that the fundamental algorithm covered. While it reduced spurious pronominal coreference to a certain degree, this effect is overshadowed by the much more substantial effect of failing to resolve so many cases. Note that the Tr corpus contains 1060 pronouns participating in coreference relations, and the Eval corpus 248. This means the focus-based approach is not resolving at all some 10-20% of the coreferential pronouns in the text.

## 4.5 Discussion

The lower performance of the focus-based approach against both corpora, at all levels, and on all metrics raises the question of whether the more complex focus-based algorithm has any real advantage over the simpler fundamental approach. In order to arrive at any firm conclusions about this we must identify possible sources of error in the focus-based approach and see to which of them errors in the current implementation are attributable.

There are three possible sources of error.

1. The input to the algorithm may be so noisy it cannot perform well. The algorithm relies on the parser correctly identifying the principal grammatical role players in a sentence, e.g., the logical object and subject. If these are not found then focus registers may be incorrect or empty leading to incorrect or missing coreferences. Partial parses will also affect the identification of *EE* boundaries, on which the focus update rules depend. For example, if the parser fails to attach a prepositional phrase containing an antecedent, it will then be missed from the focus registers and so not be available to the *IR*s.

2. Either or both of the register update rules and the interpretation rules may be incomplete or incorrect – i.e. the underlying focus model is sound but the rules (which may be viewed as parameters of the model) need adjusting.

3. There are pronominal coreference phenomena the focus model simply cannot account for and which no amount of fiddling with the rule set can fix (or may, but only at the cost of introducing other problems). That is, some of the fundamental assumptions of the focus-based approach, such as that the focus is favoured as an antecedent, may not always apply.

Without doing exhaustive manual failure analysis on individual cases across the test sets, there is no way to ascertain to what extent these each of factors is at play. This detailed work has not been done as yet. However, it is certainly clear that the parser we have used is far from perfect (it is designed as a partial parser and is conservative in attaching prepositional phrases) and this has likely lead the focusing mechanism both to fail to propose any antecedent and to make many errors. Certainly the high number of pronouns for which no antecedent is proposed by the focusing algorithm does suggest that the adequate information has not been available to set the focus registers.

Examples of pronouns not referring to entities in focus, can also be found in the corpus:

> *In June, a few weeks before the crash of TWA Flight 800, leaders of several Middle Eastern terrorist organizations met in Teheran to plan terrorist acts. Among them was the PFL of Palestine, an organization that has been linked to airplane bombings in the past.*

Here *leaders* is in focus at the beginning of the second sentence and would be chosen as the antecedent of *them*; however, the antecedent is, in fact, *organizations*. Further discussion of this sort of example may be found in [11].

In contrast, the fundamental algorithm has certain strengths.

1. It is much more robust in the face of partial or even incorrect syntactic analysis, since coreference is determined by semantic compatibility and recency only. While syntactic analysis does play a part here (e.g. in establishing which attributes an entity possesses), overall the algorithm is much less sensitive to parsing problems.

2. It proposes antecedents of pronouns from each preceding paragraph until one is accepted, while the focus-based approach suggests a single fixed set. This boosts recall for the fundamental approach at some cost in precision.

However, examples such as the one introduced at the beginning of section 3 remain as challenges to any simple approach. To what extent such examples account for the fundamental algorithm's loss in precision also remains to be determined.

# 5 Conclusion

A focus-based approach to pronoun resolution has been implemented within the LaSIE IE system and comparatively evaluated on real-world texts against a simpler approach based primarily on semantic compatibility and recency only. The results show the simpler approach performing somewhat better on all measures. One main limitation of the focus-based approach is its reliance on a robust syntactic/semantic analysis to find the focus on which the interpretation rules which assign antecedents to pronouns depend. Examining performance on the real-world data also raises questions about the theoretical assumptions of focus-based approaches, in particular whether the 'focus' is always a favoured antecedent.

Certain examples clearly show the inadequacy of the simpler approach which, when its simple syntactic and semantic rules propose a set of equivalent antecedents, can only select, say, the closest arbitrarily. A combined approach may therefore be suggested, where the focus-based approach is applied first and then, if a pronoun remains unresolved, the fundamental algorithm is used as a default to resolve it. Whether this would be more effective than further refining the update and resolution rules of the focus-based approach, or improving parse results and adding more detailed semantic constraints, remains an open question.

# References


[1] S. Azzam. *Computation of Ambiguities (Anaphors and PPs) in NL texts. CLAM: the prototype*. PhD thesis, Paris Sorbonne University, 1995.

[2] S. Azzam. Resolving anaphors in embedded sentences. In *Proceedings of the 34th meetings of the Asssociation for Computational Linguistics (ACL'96)*, Santa Cruz, CA, 1996.

[3] D. Carter. *Interpreting Anaphors in natural language texts.* Ellis Horwood, Chichester, 1987.

[4] Defense Advanced Research Projects Agency. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Morgan Kaufmann, 1998. Forthcoming.

[5] R. Gaizauskas. XI: A Knowledge Representation Language Based on Cross-Classification and Inheritance. Technical Report CS-95-24, Department of Computer Science, University of Sheffield, 1995.

[6] R. Gaizauskas and K. Humphreys. Quantitative Evaluation of Coreference Algorithms in an Information Extraction System. In S. Botley and T. McEnery, editors, *Discourse Anaphora and Anaphor Resolution*. Forthcoming. Also available as Department of Computer Science, University of Sheffield, Research Memorandum CS – 97 – 19, `http://www.dcs.shef.ac.uk/research/resmems`.

[7] R. Gaizauskas, T. Wakao, K Humphreys, H. Cunningham, and Y. Wilks. Description of the LaSIE system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 207–220. Morgan Kaufmann, 1995.

[8] J.S. Gruber. *Lexical structures in syntax and semantics*. North-Holland, 1976.

[9] K. Humphreys, R. Gaizauskas, S. Azzam, C Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. Description of the LaSIE-II system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)* [4]. Forthcoming.

[10] M. Kameyama. Intrasentential centering: A case study. In M. Walker, A. Joshi, and E. Prince, editors, *Centering Theory in Discourse*. Oxford University Press, Oxford, 1997.

[11] A Kehler. Current theories for centering for pronoun interpretation: A critical evaluation. *Computational Linguistics*, 23(3):467–475, 1997.

[12] C. Sidner. Focusing for interpretation of pronouns. *American Journal of Computational Linguistics*, 7:217–231, 1981.

[13] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 45–52. Morgan Kaufmann, 1995.