Coreference Resolution in a Multilingual Information Extraction System

Saliha Azzam, Kevin Humphreys and Robert Gaizauskas

{s.azzam,k.humphreys,r.gaizauskas}@dcs.shef.ac.uk

Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello Street
Sheffield S1 4DP UK

Abstract

We present in this paper the coreference mechanism implemented in the M-LaSIE system, a prototype multilingual Information Extraction (IE) system. We describe an experiment in which texts from a parallel French/English corpus were marked up manually and processed by the system following the MUC coreference annotation scheme. This experiment allows us to assess the applicability of the MUC annotation scheme to a non-English language, to make several observations about differences in coreference behaviour in English and French, and to assess in a tentative way the cross-language portability of the M-LaSIE approach to coreference resolution.

1. Introduction

The M-LaSIE system (Gaizauskas et al., 1997) is a prototype multilingual Information Extraction (IE) system with a coreference mechanism which can conform to the MUC coreference task specification (Def, 1995; Def, 1998). Unlike many IE systems that skim texts and use large collections of shallow, domain-specific patterns and heuristics to fill in templates, M-LaSIE attempts a fuller text analysis, first translating individual sentences to a quasi-logical form (QLF), and then constructing a weak discourse model of the entire text. Underpinning the system is a language independent 'domain model', represented as a semantic net, which is extended during the processing of a text by adding the classes and instances described in that text. The coreference mechanism is of central importance in M-LaSIE, both for integrating the QLF representations of successive sentences into the discourse model and for allowing domain and world knowledge to be brought to bear in 'gluing' together the multiple fragments of QLF produced for single sentences by the system's robust but partial parser.

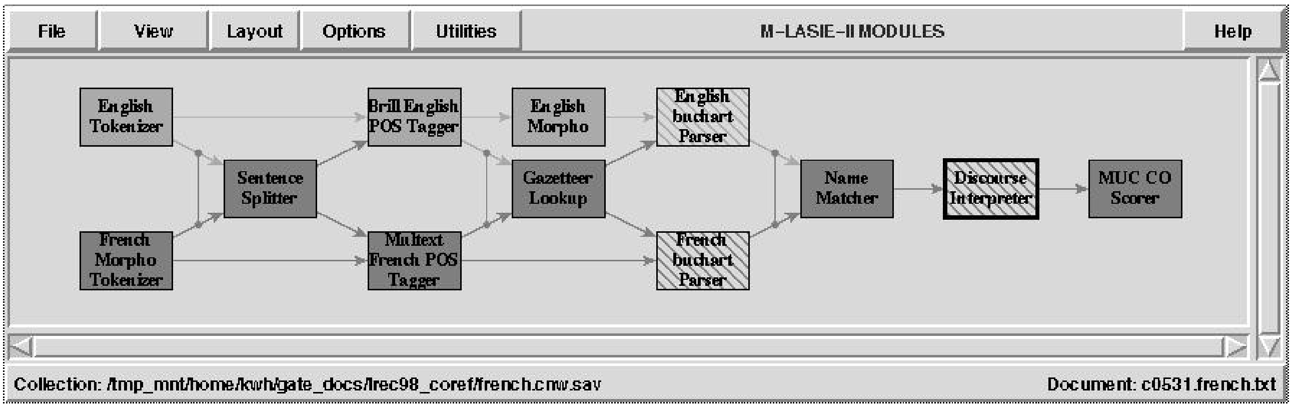
So far M-LaSIE can process texts in English and French only, though work is underway to develop Spanish and German versions of the system, as part of the EU AVENTINUS project (Thurmair, 1996). A small corpus of parallel French/-English newswire articles in the (MUC-6) domain of management succession events has been used in the development of M-LaSIE. In this paper we discuss recent investigations concerning the language (in)dependency of the coreference mechanism as revealed by experimentation with this corpus. Following an initial overview of M-LaSIE and its approach to coreference resolution, we describe those aspects of the approach that have needed modification in moving from English to French. We then describe an experiment in which texts from the parallel corpus were marked up following the MUC coreference annotation scheme. This experiment allows us to assess the applicability of the MUC annotation scheme to a non-English language, to make some observations about differences in coreference behaviour in English and French, and to assess in a tentative way the cross-language portability of our approach to coreference resolution.

2. M-LaSIE Overview

The prototype multilingual IE system M-LaSIE has been derived from the English-only LaSIE system (Gaizauskas et al., 1995; Gaizauskas and Humphreys, 1997). LaSIE was designed as a general purpose IE research system, geared towards, but not solely restricted to, carrying out the English language tasks specified in MUC-6 and MUC-7: named entity recognition, coreference resolution, template element filling, template relation filling and scenario template filling (see Def (1995) and Def (1998) for details of the tasks). In addition, the system can generate a brief natural language (NL) summary of any scenario templates it has filled from the text. Both LaSIE and M-LaSIE have been implemented within GATE, the General Architecture for Text Engineering (Cunningham et al., 1997) which facilitates modular development, integration and reuse of language processing components.

The LaSIE system is a pipelined architecture which processes a text sentence by sentence. It consists of three principal processing stages: lexical preprocessing, parsing plus semantic interpretation, and discourse interpretation. The overall contributions of these stages may be briefly described as follows:

- Lexical preprocessing reads and tokenises the raw input text, tags the tokens with parts-of-speech, performs morphological analysis, and performs multiword matching against lists of known proper names;
- Parsing does two pass chart parsing, pass one with a special named entity grammar, and pass two with a general phrasal grammar. A 'best parse' is then selected, which may be only a partial parse, and a predicate-argument representation, or quasi-logical form (QLF), of each sentence is constructed compositionally.



Collection: /tmp_mnt/home/kwh/gate_docs/lrec98_coref/french.cnw.sav

Document: c0531.french.txt

Figure 1: M-LaSIE architecture

- Discourse interpretation adds the QLF representation to a semantic net, which encodes the system's domain model as a hierarchy of concepts. Additional information presupposed by the input is also added to the model, then coreference resolution is performed between new and old instances. Information consequent upon the input is then added, resulting in an updated discourse model.

When an entire text has been processed, the result is a single, integrated discourse model. Templates for specific IE tasks and the NL summary are generated directly from this model.

M-LaSIE (pictured in Figure 1) is a relatively straightforward elaboration of LaSIE. The lexical preprocessing and parsing stages are necessarily language-specific, since separate languages are morphologically and syntactically distinct, though the same algorithms may be used for separate languages (e.g., trainable part-of-speech taggers, chart parsers). The target representation in the discourse model is, however, intended to be language independent. While this was meant in principle to be true of LaSIE, the development of M-LaSIE has lead us to see more clearly where language dependencies were in fact built into the representation and has helped us to correct these. Further it has enabled us to see to what extent algorithms that work on this language independent representation of the discourse, such as the coreference mechanism, carry across languages.

The QLF output by each language-specific parser marks the point where the language independent representation begins to emerge, so we proceed by describing it in more detail.

2.1. QLF

Semantic interpretations are assigned to each sentence in a text during parsing using what is essentially a classical compositional approach – each phrase structure rule has a corresponding semantic rule which specifies how a semantic representation is to be built up. The result is a *quasi-logical form* or QLF, much cruder than that used by Alshawi (1992), but sharing the characteristics of retaining various proximities to the surface form and of postponing some disambiguation, e.g., prepositional phrase and relative clause attachments, full analysis of quantifier scope and word sense disambiguation.

Syntactically, QLF expressions are simply conjunctions of first order logical terms. The predicates in the QLF representation are either derived from the appropriate lexical morphological roots of head words, or come from a closed class of relational predicates that express modification or semantic role relations. To be more specific:

1. NPs lead to the introduction of a unary predicate whose functor is the morphological root of the head of the NP and whose argument is a unique index which serves as an identifier for the entity referred to – e.g. *company* will map to something like company(e22).

   (a) Determiners such as *the*, *some* and *many* lead to the introduction of a det relation whose first argument is the index introduced by the head noun and whose second argument is the actual determiner. E.g. *the company* becomes
   company(e22), det(e22,the).

   (b) Cardinal quantifiers such as *three*, *10 million* lead to the introduction of a count relation. E.g. *three companies* becomes
   company(e22), count(e22,3).

   (c) Adjectives such as *big* and *old* are treated in the same way as determiners, by introducing an adj relation with the adjective itself as an argument. E.g. *big company* becomes
   company(e22), adj(e22,big).

   (d) Noun modifiers introduce new indices which are treated as the second argument to a qual relation, so that, e.g. *computer company* becomes
   computer(e21), company(e22), qual(e22,e21).

2. VPs lead to the introduction of a unary predicate whose functor is the morphological root of the head of the VP and whose argument is a unique index which serves as an identifier for the event referred to – e.g. *hired* will map to something like hire(e34), time(e34,past) [1].

3. Where complement structure has been recognised in the parser this is recorded in the QLF representation using binary relations of the form lsubj(e34,e22) (for logical subject), lobj(e34,e25) (for logical object)

[1] This treatment of VPs is in the tradition of (Davidson, 1967).

and, in the case of prepositional phrase complements, *prep*(e34,e29) (where *prep* is the actual preposition, e.g. beside(e34,e29)).

The non-lexically derived predicates, i.e. the grammatical relation predicates such as lsubj, lobj, adj, qual, det, etc., are language independent, though whether all of them are likely to be utlised in each language and what the full set is is not known. The lexically derived predicates are clearly language dependent, and are mapped onto language-independent 'concept' nodes in the discourse model as the first stage in discourse interpretation.

2.2. The Discourse Model

The discourse model is constructed by integrating the QLF of successive sentences into a pre-existing domain-specific semantic net. This net, which we refer to as an *ontology*, is represented as a hierarchically organised directed graph of 'concept' nodes connected by isa links or isinstance links, depending on whether the subordinate node is a subclass or an instance of a dominating node. Each node may have associated with it an attribute-value structure which can be inherited down the graph.

At the highest level, the hierarchy divides into object, event, and property nodes. Instances of unary predicates in the incoming QLF are generally added beneath the object node if they are nominal and beneath the event node if verbal, as indicated by the presence of time or aspect information. Event nominalizations, if recognised as such, are added beneath the event node. Relational (always binary) predicates in the QLF are added to the attribute-value structures of any instances referenced in them.

The semantic net that exists prior to the processing of any text reflects prior encoding of conceptual and world knowledge and may be as rich or impoverished as the IE application designer chooses. If an incoming predicate (e.g. company(e9)) is already represented by a node in the semantic net, then the new instance is recorded beneath that node; if not, a new class node is added directly beneath the object or event node and the instance placed there.

One special object is the text object which records information about the text. In particular it records the division of the text into sections and sentences and records in which sentence and in what order within the sentence each instance index was introduced. This information is later used by the coreference mechanism to implement recency constraints on coreference between surface referring expressions and to direct hypothesis-driven resolution of prepositional phrase, relative clause, and complement attachment ambiguities.

This representation for discourse modelling is intended to be language independent. The nodes in the semantic net are not language specific (though they may correspond to concepts which are lexicalised only in a particular language) and neither is the text representation. However, the processes which map into and out of the discourse representation and which manipulate it need to make use of language specific information. A lexeme-to-concept mapping at the initial stage of adding the QLF to the semantic net means source language lexical dependencies are left behind for those words/concepts which are recorded in the mapping table, but means that such a mapping table needs to be constructed for each language. This mechanism is crude and in particular does not address the well-known problems of word sense ambiguity (though such ambiguities are rare in limited domain IE applications) and lexical gaps. However, these problems do not invalidate the basic model and more sophisticated solutions can be incorporated into our framework as they emerge. The converse problems arise on mapping out of the discourse representation into NL; again, these are well-known problems in multilingual NL generation and general solutions can be applied in this context.

Of more interest is the question of whether any processing which is carried out solely on the discourse representation itself is language dependent, and if so whether the processing apparatus can be parameterised so as to make portability across languages possible. The key process of concern here is coreference resolution.

3. Coreference Resolution

3.1. The Base Algorithm

After the QLF representation of a sentence has been added to the discourse model, all new indices (those introduced by that sentence) are compared with other indices in the model to determine whether any pair can be merged, representing a coreference in the text. The comparison of indices is carried out in several stages:

1. new indices with proper_name attributes are compared with all existing indices with proper_name attributes;
2. all new indices are compared with each other (intrasentential coreference resolution);
3. new indices introduced by pronouns are compared with existing indices from the current paragraph, and then each previous paragraph in turn until an antecedent is found;
4. other new indices are compared with all existing indices in the model.

Each comparison involves first determining if the indices' classes lie on the same branch in the ontology (type-compatibility). If not, then the indices are not considered further for coreference. If they are on the same branch then the attributes of the indices are compared to ensure there are no conflicts (attribute-compatibility). Certain attributes, such as animate, are defined in the ontology as taking single, fixed values for a particular index and so indices with conflicting values for these attributes cannot be the same. If such conflicts are discovered then the comparison is abandoned.

If no attribute conflicts are found between two indices, a similarity score is calculated based on the number of common attributes and on a semantic distance measure, determined simply in terms of the number of nodes on the path in the semantic graph between them. After a newly input index has been compared with all others in a particular comparison set, it, together with its attributes, is merged in the discourse model with the index with the highest similarity score (if any score).

There is nothing language specific in this base algorithm since it operates solely on the language independent discourse representation. For example, consider the example of definite noun phrases, i.e, NPs with a definite determiner. In the following French text:

> *Lafarge Corporation est l'un des principaux fournisseurs de ciment, de béton, de granulats et d'autres matériaux de construction pour le secteur résidentiel, commercial, et des travaux publics en Amérique du Nord. La Société exploite actuellement 14 cimenteries et environ 400 opérations de matériaux de construction au Canada et aux Etats-Unis.*

and its English translation:

> *Lafarge Corporation is one of North America's largest producer of cement, concrete, aggregates and other construction products for residential, commercial, institutional and public works construction. The company operates 14 cement plants and approximately 400 construction materials operations in the U.S. and Canada.*

the initial definite noun phrases of the second sentence – *La Société* and *The company* – both give rise to an instance of the same concept in the domain model, company, with the property 'definite', resulting from the determiners *la* and *the* respectively. Therefore, the same coreference rules apply to both instances, and give *Lafarge Corporation* of the previous paragraph as the antecedent.

3.2. A Focus-based Extension

In addition to the base coreference algorithm, we have also experimented with an approach based on Azzam's proposed extensions (Azzam, 1996) to Sidner's focusing approach (Sidner, 1981). This approach is based on the claim that anaphora generally refer to the current discourse focus, or 'center', and so modelling changes in focus through a discourse will allow the identification of antecedents for anaphors. So far, we have only applied the approach to pronominal coreference.

The algorithm makes use of several *focus registers* to represent the current state of a discourse, mainly *CF*, the current focus register and *AF*, the actor focus register. At first, the *CF* is initialised to the theme of the first sentence, the 'theme' being either the object of a transitive verb, or the subject of an intransitive or the copula (following Gruber (1976)), and the *AF* is initialised to the agent of the first sentence. A set of *interpretation rules* (*IR*s) applies whenever an anaphor is encountered, proposing potential antecedents from the registers from which one is chosen using other criteria: syntactic, semantic, inferential, etc. The focusing algorithm updates these registers after each simple clause is processed (the clause, as opposed to the sentence, being the processing unit in our system – see (Azzam, 1996)), confirming or rejecting the current focus. The main stages in pronominal coreference are the following.

- The class of the pronoun is determined by:
  - its animacy (animate, inanimate, unknown), in turn determined by its surface form and possibly by semantic role information if available (e.g. if the pronoun is the logical subject of a verb which requires an animate subject, such as *say*);
  - its syntactic type (personal, reflexive, possessive, demonstrative) provided by its lexical entry in the monolingual lexicon.

  Note that a pronoun and its literal translation in a different language may not be of the same class. For example, the translation of *he* in French, *il*, belongs to the class animate: unknown, syntype: personal (since *il* can be either inanimate or animate), whilst *he* in English belongs to the more specific class, animate: yes, syntype: personal. Thus, language specific knowledge is required to assign a pronoun to the correct class.

- The interpretation rules (*IR*s) propose an antecedent taking into account the class of the pronoun and the state of the focus registers. For example, one *IR* states: *if the pronoun is animate and personal it corefers with the current AF*. The *IR*s are language independent as they do not make use of language specific knowledge. They are expressed in terms of *pronoun classes* and *focus register states* only.

- Antecedents proposed by the *IR*s are accepted or rejected based on their semantic type and feature compatibility, using M-LaSIE's base coreference mechanism which relies on semantic and attribute value similarity scores.

- Finally, the focusing rules take into account the results of the resolution to decide whether the focus remains the same or changes. These rules apply only to the registers and are therefore completely language independent. An example is: *if a pronoun in a theme position corefers with the current focus, keep the current focus.*

4. Corpus Annotation for Coreference

We use the MUC annotation scheme for coreference relations, as defined in the MUC-6 Coreference Task Definition v2.3 (Def, 1995) and slightly revised for MUC-7. It is important to note that this definition in no way purports to exhaustively describe the coreference phenomena in natural language, that it is concerned primarily with a certain sort of text – newswire articles, and that some arbitrary decisions were taken to allow for automatic scoring in MUC. The following is a synopsis of the core parts of the MUC definition, borrowing heavily from (Def, 1995).

4.1. The Annotation Scheme

Coreferential expressions are annotated by adding SGML tags into the text. Given an antecedent A and an anaphor B, where both A and B are strings in the text, the basic coreference annotatation has the form

```
<COREF ID="100"> A </COREF> ...
<COREF ID="101" TYPE="IDENT" REF="100"> B </COREF>
```

The ID attribute is a unique identifier for each string in a coreference relation, and the REF attribute indicates which string is coreferential with the one which it tags. The TYPE attribute serves to indicate the relationship between anaphor and antecedent, with the value IDENT indicating identity, which, for the MUC task, is the only relationship to be marked.

An optional MIN attribute is also used to identify the minimum string that would be accepted by the scoring algorithm – either the head of the phrase or a named entity. Full credit is given for any string including at least the MIN string and at most the full string. This attempts to decouple the coreference task from the task of accurately parsing noun phrases.

4.2. Definition of the Task

Coreference relations are only marked between strings of certain classes of nouns, noun phrases, and pronouns, known as *markables*, and only if the string with which they corefer is also markable (so, e.g., a pronoun referring to a clause would not be markable).

Markables include:

- names and named entities (as defined in the MUC named entity task) – e.g. “Galactic Enterprises Inc.”;
- definite noun phrases – e.g. “the company”.
- conjoined noun phrases – e.g. in

  ```
  *The boys and girls* enjoyed *their* breakfast.
  ```

  the starred strings should be marked as coreferential.
- present participles modified by nouns or adjectives – e.g. “deficit financing”;
- pronouns (personal, demonstrative, possessive and reflexive forms) – e.g.

  ```
  *He* shot *himself* with *his* revolver.
  ```
- ‘bare’ nouns occurring as prenomial modifiers – e.g.

  ```
  Sheffield’s production of *steel* has dropped due
  to foreign competition in the *steel* industry.
  ```

Examples of non-markables are:

- names embedded in other names – e.g. “Kent” in

  ```
  The Duchess of Kent
  ```
- gerunds – e.g.

  ```
  Leaping over tall buildings
  ```
- implicit pronouns – e.g. in

  ```
  John posted the letter and walked home.
  ```

  the implicit subject of “walked” should not be linked to “John” by marking an empty string.

Given the definition of markable, the task definition identifies a set of coreference relationships to annotate. These are:

1. **basic coreference** Two markables that refer to the same object, set or activity.
2. **bound anaphors** Noun phrases and anaphors bound by them even if they are not coreferential in the usual sense, e.g.

   ```
   *Every student* discovered *their* grades.
   ```
3. **apposition** Appositional phrases in which both noun phrases are definite and which are explictly marked via overt punctuation, e.g.

   ```
   *Tony Blair*, *the Prime Minister*, ...
   ```

   but not

   ```
   *Bloggs*, *an old friend of mine*, ...
   *Treasury spokesman* *Jones* ...
   ```
4. **predicate nominals and time-dependent identity** Predicative nominals, regardless of time, provided they are definite, e.g.

   ```
   *Blair* is *Prime Minister of Great Britain*.
   *Major* was *Prime Minister of Great Britain*.
   ```

   are both marked, but not

   ```
   Hague might be Prime Minister of Great Britain.
   Politics is a profession for rogues.
   ```
5. **types and tokens** Markables referring to identical sets or types, though the distinction between sets and types is not always easy to define and in cases where there is residual doubt the links are marked as optional. For instance, in

   ```
   *Producers* don’t like to see a hit wine increase
   in price... *Producers* have seen this market
   opening up and *they*’re now creating wines that
   appeal to these people.
   ```

   the three starred markables, if taken as referring to the same sets, would not be marked as coreferential since the set of producers who have seen the market opening up is presumably not the same as the set of those who have created new wines in response to this. However, these markables are taken as referring to the same *type* and hence are marked as coreferential.
6. **functions and values** An expression may refer to the value of a function at certain arguments by mentioning the function and arguments explicitly, by assuming the arguments implicitly from context, or by simply stating the value. In

   ```
   GM announced *its third quarter profit*.
   *It* was *$0.02*.
   ```

   all three starred expressions are marked as coreferential. In

   ```
   *The temperature* is *90* ...
   The temperature is rising.
   ```

   the first occurrence of “The temperature” refers to the value of the function at arguments whose value is supplied by context and that value is 90. Hence the first two starred expressions are marked as coreferential. The second occurrence of “The temperature” refers to the function (indirectly by reference to its first derivative) and not to its value and hence is not marked as coreferential with either of the earlier two expressions.
7. **metonymy** Metonymy is viewed as type coercion. For example, in

```
*The White House* held a press conference today.
*The beleaguered administration* announced ...
```

the White House is coerced to the administration operating out of the White House. Metonymical markables such as this are marked as coreferential if the entities referred to *after* coercion are identical. Thus, in the preceding example the two starred references are marked as coreferential. However, in

```
I bought the New York Times this morning. I read
that the editor of the New York Times resigned.
```

the first reference to the New York Times is coerced into a copy of the paper published by the New York Times, while the second is coerced into the organisation; in this case no coreference is marked.

5. Initial Experiments in Multilingual Coreference

As a preliminary experiment in investigating multilingual coreference, we annotated and developed the M-LaSIE system using a small parallel corpus of 20 French and English texts available on the web from Canada NewsWire Ltd. (www.newswire.ca).

The MUC coreference annotation scheme has proved suitable for the French examples processed so far, as illustrated in the following text for the different types of coreference, e.g. apposition, pronouns, proper names, etc.

```
CHARLOTTE, Caroline du Nord, 13 septembre /CNW/ - <COREF
ID="1">United Dominion Industries Ltd.</COREF> (<COREF
ID="0" TYPE="IDENT" REF="1">UDI</COREF> aux bourses de
Toronto et de New York), <COREF ID="2" TYPE="IDENT"
REF="1">fabricant de produits usines
diversifis</COREF>, a annonce aujourd'hui la nomination
de <COREF ID="5">John G. Mackay</COREF>, qui est age de
56 ans, au poste nouvellement cree de vice-president
directeur pour l'<COREF ID="8">Europe</COREF>.

En outre, <COREF ID="4" TYPE="IDENT" REF="5">M.
MacKay</COREF> participera a la coordination des entites
d'exploitation de <COREF ID="6" TYPE="IDENT"
REF="1">United Dominion</COREF> en <COREF ID="7"
TYPE="IDENT" REF="8">Europe</COREF>, contribuera a
<COREF ID="9" TYPE="IDENT" REF="6">ses</COREF> activites
financieres et de planification fiscales internationales
et nouera des liens avec des banques, des fabricants
europeens et d'autres.  <COREF ID="10" TYPE="IDENT"
REF="4">Il</COREF> demeurera premier vice-president et
relevera de <COREF ID="12">William R. Holland</COREF>,
<COREF ID="11" TYPE="IDENT" REF="12">president du
conseil et chef de la direction</COREF>.
```

The MUC scoring software will therefore be applicable to French as well as English.

5.1. Evaluation of Pronominal Coreference

We now report some initial figures on the performance of the M-LaSIE coreference algorithm on French and English pronouns. For a total of 30 pronouns in 10 French texts the results are as follows, using the standard Information Retrieval metrics of ‘recall’ and ‘precision’.[2] The scoring of pronoun coreference was done manually, since the MUC scoring software cannot currently be restricted to a particular class of anaphor. In the following, a pronoun coreference counts as correct if and only if the entities in the coreference chain to which it belongs in the system's response are a subset of the set of entities in a coreference chain in the key.

For the base M-LaSIE system, extended with the focus-based algorithm for pronouns, pronoun resolution gave:

```
   Recall = 14/30  (47%)
Precision = 14/18  (78%)
```

The same system on the parallel 10 English texts, with a total of 19 pronouns, gave the following results:

```
   Recall = 12/19  (63%)
Precision = 12/14  (86%)
```

One of the most apparent distinctions between French and English pronouns is that different information is conveyed by third person singular pronouns in French, i.e. the animacy of the antecedent is not determinable from the pronoun (though the gender is, even for inanimate objects). This can generate ambiguous cases not found in English, as shown in the French and English text below: *Elle* can corefer to any singular feminine entity in focus, in this case *promotion*, while for *She*, *promotion* will be rejected, as *She* can only refer to a person, *Jane*.

> *“La promotion de Mme Baird est un autre exemple de la faĉon dont Cognos rcompense le travail acharné et le dévouement. Elle a clairement démontré sa capacité de relever les défis, ainsi que de diriger et de motiver une équipe des plus talentueuses”, a ajouté M. Zambonini*
>
> *“Jane's promotion is yet another example of how Cognos rewards hard work and dedication,” Zambonini continued. “She has clearly demonstrated her ability to rise to challenges, as well as lead and inspire a very talented team.”*

One solution to this problem is a more general model of verb subcategorisation than that currently used in the LaSIE approach, so that the semantic types of verb roles can be specified. Pronouns in role positions can then be classified more accurately. At present subcategorisation patterns are only included as part of the domain model for specific IE tasks, with no general purpose classification of event types and their roles.

After adding subcategorisation patterns to the M-LaSIE domain model for the particular verbs occuring in the corpus, thus allowing some disambiguation of third person pronouns, three previously incorrect pronouns could now be correctly resolved:

```
   Recall = 17/30  (57%)
Precision = 17/18  (94%)
```

The addition of subcategorisation patterns has no disambiguating effect on pronoun resolution in the English corpus, though potentially it will allow the use of animate

[2] Recall is a measure of how many correct (i.e. manually annotated) coreferences a system found, and precision is a measure of how many coreferences that the system proposed were actually correct. For example, suppose there are 100 manually annotated coreference relations in a corpus and a system proposes 75, of which 50 are correct. The system's recall is then $50/100$ or 50% and its precision is $50/75$ or 66.7%.

pronouns referring to organisations to be distinguished, i.e. to allow type coercion.

However, type coercion, or metonymy, is a general problem for both French and English, and accounts for some of the overall missing recall. For example in:

> *"We believe the company will benefit from their extraordinary talents in each of their respective new assignments," said Bertrand P. Collomb, Chairman of Lafarge Corp.*

the first person animate pronoun *we* should corefer with *Lafarge Corp.* but is ruled out in the current coreference mechanism due to conflicts in both type and number.

This example also illustrates the problem of cataphora, i.e. pronouns occuring in the text before their antecedents, which is only partially handled in the current mechanism. The focus-based algorithm looks for antecedents in the current simple clause, but this is disrupted with complex clauses. For example in:

> *At its recent quarterly meeting, held in Virginia, the Board of Directors of Lafarge Corporation, appointed a new president and CEO, John M. Piccuch.*

the algorithm will miss the resolution of *its*, whose antecedent, *Board of Directors of Lafarge Corporation*, occurs only in the subsequent clause. The cataphora *its* would be resolved correctly without the *held* clause.

This problem occurred more frequently in French texts for the pronoun *se*, for example in:

> *Avant de se joindre à Northern Telecom, M. Safarikas a travaillé quelques années chez Ogivar Inc.*

where the antecedent *M. Safarikas* occurs in the subsequent *travaillé* clause.

Recall in the French texts could also be noticeably improved with a more accurate French grammar, since the current one has only been developed for simple sentences, and handles relative and conjoined clauses poorly.

5.2. Non-pronominal coreference

In the current corpus, there are no noticable differences in M-LaSIE's resolution of non-pronominal anaphora between French and English. However, one potentiallly significant distinction occurs for definite determiners. While for both languages definite determiners do not always introduce anaphoric NPs (see, e.g. Vieira (1997)), the class of non-anaphoric NPs introduced by definite determiners differs for the two languages. This is illustrated in the example below with the determiner *la* in *de la recherche*, while in the English version this phenomenon does not occur, as mass nouns such as *research* do not need determiners.

> *Elle a également été directrice de la recherche en commercialisation pour North American Automotive Operations de Ford jusqu'en 1992.*
>
> *She also served as director of marketing research for Ford's North American Automotive Operations until 1992.*

This suggests a more restrictive definition for anaphoric definite noun phrases is required in French to avoid spurious coreference.

6. Conclusion

The experiment carried out for French and English coreference in M-LaSIE illustrated several interesting linguistic phenomenon, mainly those related to the different feature sets of pronouns in each language. The lower precision obtained on the French texts was due to the more ambiguous classifications of French pronouns, forcing less precise resolution rules. We then showed that additional disambiguating information, particularly verb subcategorisation frames, could be incorporated easily to improve the results, without changing the basic coreference approach.

Much of the lost recall in the results is due to the main drawback of the focus-based approach, that is the reliance on a robust input from the parser. Since often only partial parses are available, much information about verb roles, on which the focus-based approach relies, is lost. However, some of this can be recovered through the system's partial parse extension mechanism, which relies on (the currently limited) domain specific subcategorisation patterns.

Another significant problem is that of unknown words, where, for example, a resolution rule which requires an animate antecedent will fail if entities in the text are not recognised as such. A larger ontology, or methods for extending the ontology automatically, will reduce this problem.

Future work on both incorporating further subcategorisation information and extending the ontology is planned.

7. References

Alshawi, H., editor. 1992. *The Core Language Engine*. MIT Press, Cambridge MA.

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

Cunningham, H., K. Humphreys, R. Gaizauskas, and Y. Wilks. 1997. Software Infrastructure for Natural Language Processing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*, pages 237–244, March.

Davidson, D. 1967. The logical form of action sentences. In N. Rescher, editor, *The Logic of Decision and Action*. University of Pittsburgh Press, Pittsburgh.

Defense Advanced Research Projects Agency. 1995. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann.

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

Gaizauskas, R. and K. Humphreys. 1997. Using a semantic network for information extraction. *Journal of Natural Language Engineering*, 3(2/3):147–169.

Gaizauskas, R., K. Humphreys, S. Azzam, and Y. Wilks. 1997. Concepticons *vs.* lexicons: An architecture for multilingual information extraction. In M.T. Pazienza, editor, *Proceedings of the Summer School on Information Extraction (SCIE-97)*, LNCS/LNAI, pages 28–43. Springer-Verlag.

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

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

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

Thurmair, G. 1996. AVENTINUS System Architecture. LE project LE1-2238 AVENTINUS internal technical report, GMS, Germany.

Vieira, R. 1997. *Definite Description Processing in Unrestricted Text*. Ph.D. thesis, University of Edinburgh.