RC22627 (W0211-041) November 5, 2002
Computer Science

IBM Research Report

Euphoria -- A Reference Resolution System for Machine Translation

Arendse Bernth
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

Euphoria – A Reference Resolution System for Machine Translation

Arendse Bernth

November 1, 2002

Abstract

A system for reference resolution in the context of the multilingual environment of Machine Translation is proposed, and a description of the first stage of implementation described. We argue that a semantics-based approach is more language-independent than a syntax-based, and show how WordNet functions provide a basis for resolution of pronouns and anaphoric nouns. The focus of the inital stage is design and treatment of third person pronouns and of speaker recognition. The system has been integrated with the English Slot Grammar Parser, so that parse structures show reference relations.

1 Introduction

Anaphora resolution is obviously useful for *any* natural language processing (NLP) system. In this paper we shall focus on anaphora resolution in the context of machine translation (MT). The most obvious use of pronominal anaphora resolution in MT is to render the translation of the pronoun correct with respect to morpho-syntactic features such as person, number, and gender, as applicable. The second most obvious use is to ensure proper subject-predicate agreement for languages that require this.

A third use in MT that may not be as obvious, but is nevertheless also very important, is to enable any semantic tests associated with the antecedent in the transfer lexicon to apply. An example is shown in (1). English *carry out* translates differently into German depending on the semantic type of the object. If the object is a physical thing like trash, then the proper translation is *wegtragen* as shown in (1a). However, if the object is non-physical like a plan, then the German should be *ausführen* as shown in (1b).

(1) a. He carried out the trash.

Er trug den Abfall weg
HE CARRIED THE TRASH AWAY

b. He carried out the plan.

Er führte den Plan aus
HE CARRIED THE PLAN OUT

If the object is pronominalized as in (2), then the proper translation requires identifying the antecedent in order to enable the semantic tests to apply. Without the tests, the translation would be as in (2a) and (2b) where we get *ausführen* for both cases. Additionally, the gender is the (wrong) default neuter gender *es* instead of the correct masculine *ihn*.

(2) a. He was working on his trash. He carried it out carefully.

Er arbeitete an seinem Abfall. Er führte es vorsichtig aus.
HE WORKED ON HIS TRASH. HE CARRIED IT CAREFULLY AWAY

b. He was working on his plan. He carried it out carefully.

Er arbeitete an seinem Plan. Er führte es vorsichtig aus.
HE WORKED ON HIS PLAN. HE CARRIED IT CAREFULLY OUT

With proper identification of the antecedent, we are able to get much better translations as shown in (3).

(3) a. He was working on his trash. He carried it out carefully.

Er arbeitete an seinem Abfall. Er trug ihn vorsichtig weg.
HE WORKED ON HIS TRASH. HE CARRIED IT CAREFULLY AWAY

b. He was working on his plan. He carried it out carefully.

Er arbeitete an seinem Plan. Er führte ihn vorsichtig aus.
HE WORKED ON HIS PLAN. HE CARRIED IT CAREFULLY OUT

[Mit02] describes the distinction between *identity-of-reference* and *identity-of-sense*. The former indicates *coreference* of antecedent and anaphor, as in (4). Here *it* and *the cottage* are coreferential.

(4) In Barcombe, East Sussex, a family had to flee their cottage when it was hit by lightning.

Identity-of-sense merely indicates *similarity* as in (5), where *it* is a paycheck, but a different one from the one mentioned in the same sentence.[1] [Gea68] named this *pronouns-of-laziness*. Pronouns-of-laziness may be substituted for expressions that are identical to, but not necessarily coreferential with, antecedent expressions.

(5) The man who gave his paycheck to his wife was wiser than the man who gave it to his mistress.

For our purposes, *viz.* MT, the distinction between identity-of-reference and identity-of-sense need not concern us, except as it might have an impact on the proper identification of the antecedent for a particular pronoun. What is important in example (5) is that *it* inherits the properties of *paycheck*, regardless of the lack of coreference.

Similarly, we need not be concerned with the complex interactions with quantifier scope that theories such as Discourse Representation Theory [KR93] address, illustrated by the so-called *donkey sentences*, an example of which is given in (6).

(6) Every man who owns a donkey beats it.

The important part is that *it* inherits the syntactic and semantic properties of *donkey*.

Another point worth making is that the *immediate* antecedent may not be the most useful antecedent, because it may be another pronoun. As [Stu01] rightly points out, applications need a non-pronominal antecedent. We might add that the most suitable non-pronominal antecedent for our purposes is the *most recent* non-pronominal antecedent, not just *any* non-pronominal antecedent.

In other words, our concern is to find the most recent non-pronominal antecedent. We shall somewhat sloppily conflate the notion of *antecedence relation* and *coreference* and use the term *coreference* indiscriminately for both cases.

The work reported on here is set in the context of Slot Grammar (SG) technology [McC80, McC90] used for the LMT [MB98] system. LMT has broad-coverage grammars for English, French, German, Italian, Spanish and Portuguese, which provide deep analysis of the source. Thus, we have multiple languages to provide pronoun resolution for, while at the same time good grammars are available. [KB96] reports on an anaphora resolution system where the situation is completely the opposite: One language (English) and little analysis available. Their approach is to push shallow analysis to its limits, using only a part-of-speech tagger enhanced with annotations of grammatical functions,

[1] The examples are quoted from [Mit02]; example (4) is attributed to *The Daily Mail* and (5) to Lauri Karttunen [Kar69].

and then apply various syntactic constraints. This system is basically a parser-free version of RAP [LL94] with some extensions. The RAP system uses *deep* analysis (English Slot Grammar, in fact) together with syntactic constraints. [GC01] reports on a pronoun resolution system that uses statistical techniques to learn syntactic constraints similar to those used by the two above-mentioned systems.[2]

What all these systems have in common is the recognition that sentence structure (syntax) tells us something about what words are in focus, or, in other words, where the center of attention is. And according to Rule 1 of Centering Theory ([GJW83a]), the center of attention is the entity that is most likely to be pronominalized. (Or conversely, pronominalization serves to give this status to a word or phrase). This is often referred to as *saliency* in the anaphora resolution literature. The more salient a word is, the more likely it is that it will be the antecedent.[3]

Different languages have different mechanisms for creating focus, and hence different rules for determining salience apply for different languages. This is due to the fact that constraints on e.g. word order vary from language to language. See e.g. [PFM01] for a description of the problems they encountered when trying to adapt to Spanish a number of syntax-based approaches designed for English. So we would have to write separate rules for each language in our MT system, were we to pursue a more syntactic approach. Given our multilingual environment, it makes more sense to minimize dependence on syntactic constraints as much as possible, while still taking advantage of the broad-coverage grammars to the extent it can be done in a language-independent manner.

As with the above-mentioned systems, our solution utilizes a "bag-and-filter" approach; i.e. we gather up potential antecedents as we process the text, and then apply filters and preferences to this bag of referents during anaphora resolution. Unlike these systems, however, our system additionally emphasizes semantics, in particular WordNet [Fel98] (which exists for several European languages),[4] as well as "predications" mined from corpora by SG technology.

While we cannot totally avoid syntactic rules, our goal is to minimize them, and to make them as language-independent as possible – at least within the languages we are currently working on, the Germanic and Romance languages. The deep analyses provided by the broad-coverage SG grammars provide the foundation for this.

This paper describes the first step in making a "language-independent" anaphora resolution system, Euphoria. This first step focuses on the general setup and its application to English only. In Section 2 we give an overview of

[2]These papers are representative of a number of different approaches to anaphora resolution; it is beyond the scope of this paper to review the totality of the extensive literature on this subject. For an excellent overview the reader is referred to [Mit02].

[3]Here we also have the explanation why the immediate antecedent may well be just another pronoun, as "useless" to an application as the pronoun we are trying to resolve.

[4]see Section 5.1 for a brief description of WordNet.

Euphoria, and in Sections 3, 4, and 5 we go into more detail about the syntactic, discourse, and semantic rules, respectively. Section 6 proposes a way of using predications mined from corpora in the anaphora resolution process and extending them by WordNet relations. Section 7 provides a brief description of the selection mechanism for choosing the best candidate. Finally in Section 8 we report on some results from the first stage of designing and implementing Euphoria.

2 Overview

2.1 A Hybrid Approach

Euphoria employs a hybrid approach, combining syntax, semantics, statistics, corpus data, document characteristics, and inference.

Syntax helps provide accurate identification of pronouns. [Stu01] points out the difficulties in relying on a part-of-speech tagger only, when deciding on whether a word like *her* is a possessive or personal pronoun and emphasizes that context is necessary to determine this. Of course fully parsing the sentence takes the totality of the syntactic context into consideration. Similarly, identification of pleonastic *it* is greatly facilitated by a full parse that can be used as the basis for syntactic rules. More accurate identification of multinouns is also possible when a full parse is available, as argued in [BMW02]. Some other things, described in Section 3, are also best handled by syntactic rules. The syntactic aspect is based on the full parse by ESG.

Semantics gives some robustness because even a fitted or partially wrong parse can be given a treatment when resolution is based on semantics rather than syntax. This lack of dependence on syntax alone also provides a measure of language independence, an important factor for our purposes. Semantics plays two roles: First, as constraints on possible antecedents. This is accomplished by semantic type checking. Second, as preferences for possible antecedents. This is accomplished by WordNet relations and inference. Inference may be particularly useful for resolving pronouns referring to propositions or events, such as the example given in (7).

(7) [The cat meowed all night]$_i$. It$_i$ prevented me from sleeping.

Inference (and translation to Logical Form) is an issue still to be addressed. Semantic type checking is based on semantic types in the SG lexicons as well as on WordNet relations such as synonomy, meronomy, hypernymy etc.

Statistics and corpus data: Simple frequency of antecedent candidates gives an important clue to the salience of an antecedent candidate as we shall argue in Section 4.2. Tuples of syntactic roles (or predications) with frequencies can

be mined from corpora and generalized by WordNet relations as described in Section 6. This provides a guide to selectional constraints.

Document characteristics such as length of document and paragraph structure as well as document classification may also play an important role. This still needs to be explored.

2.2 Order of Resolution

Careful consideration of the order of resolution of referential items is important. For example, resolving non-pronominal noun phrases before attempting resolution of pronouns provides the richest base of antecedents to draw from. It is also crucial for the resolution of cataphora, a topic we shall not address further in this report. Similarly, resolving the speaker, when applicable, before nouns in an utterance makes it possible to rule out coreference of speaker with nouns, or second or third person personal pronouns in the utterance[5] while at the same time providing a basis for the resolution of first person personal pronouns in the utterance.

Possessive pronouns and their head nouns should ideally be resolved simultaneously. For example, for the phrase *his dog* we need to know who *his* refers to in order to correctly resolve *dog*. And in order to resolve *his* it would help to know the reference of *dog*. For a sequential approach, it seems reasonable to try to resolve the pronoun first, given that antecedents of nouns may be found at a further distance than those of pronouns, and hence there will be a too large number of antecedents to choose among for the noun.

Summarizing these arguments, the proposed order of resolution is the following:

1. Speaker, if any (noun or pronoun).
2. (Multi-)nouns without possessive premodifiers.
3. Pronouns and remaining nouns: left to right.

Other systems, such as the one described by [PFM01], resolve in a strict left-to-right manner.

2.3 Resolution Procedure

The following gives an overview of the resolution procedure for each anaphoric candidate. The individual ingredients are described in the following sections. Some ingredients function as *constraints* that rule out coreference; others contribute to an overall *preference score*. The selection mechanism computes the

[5]The assumption is that a speaker will not refer to himself in the third or second person.

overall score for each antecedent candidate and returns the candidate with the highest score.

1. Identify type of anaphoric candidate (pronoun or noun).
   Case pronoun:
   (a) *If first person* (personal or possessive) resolve to current speaker.
   (b) *If pleonastic or weather*, resolve accordingly.
   (c) Go through possible antecedents:
      i. Apply sister constraint.
      ii. Apply number constraint.
      iii. Apply generic constraint.
      iv. Apply semantic type constraint.
      v. Apply selectional constraint.
      vi. *If possessive*, look for a holonym relation between antecedent candidate and syntactic mother of pronoun (WordNet).
      vii. Apply predication constraints.

   Case noun:
   (a) Go through possible antecedents:
      i. Apply semantic type tests.
      ii. Resolve appositives.
      iii. Identical word
      else
      iv. Synonym (WordNet)
      else
      v. Subsumption
      else
      vi. Hypernym (WordNet)
2. Select best candidate
3. *If no candidate:*
   Insert new referential item
   else
   Update referential item.

2.4 Communicating the Resolution Result to Other Applications

Euphoria is part of the Slot Grammar effort, and as such it is natural to mark the result of anaphora resolution in the parse output, as well as make it available to other applications through access functions.

A hook-up to ESG has been done, where Euphoria results are incorporated in the parse output. An example is shown in Fig. 1. For each referring expression (nouns, pronouns, and verbs[6]), the feature list (at the end of each line) has been updated to include a parenthesized field that gives the most recent non-pronominal antecedent and a unique referent ID. For first occurrences of nouns, these will be identical, except for the index added to make up the ID.

Two accessing functions, providing the antecedent and the ID, respectively, are available.

Additionally, the output can be annotated with the antecedent and ID; this is the format we shall use in most of the examples.

2.5 Example

The example in (8) illustrates some of the issues mentioned in Section 2. Fig. 1 shows the ESG parses of the example, including the reference resolution.

(8) A lion, searching for prey, caught a goat on a hill. It was devoured by the big cat.

As can be seen from the annotated parse tree, Euphoria introduces new referent IDs for *lion* (`lion#1`), *prey* (`prey#2`), *goat* (`goat#3`), and *hill* (`hill#4`) in the first sentence.

In the second sentence, *It* is resolved to `goat#3` and *big cat* to `lion#1`. According to the order of resolution explained in Section 2.2, *big cat* is resolved first. It is found to be corefential with the lion in the first sentence by use of the WordNet hyperonymy relation. Next, the pronoun is resolved. Since it is a sister of *cat*, coreference between these two items, as well as with the antecedent of *cat*, is ruled out. Resolving the nouns before the pronouns as described in Section 2.2 makes this possible.

3 Syntactic Rules

As mentioned in Section 1, our goal is to minimize morpho-syntactic rules because these are more likely to be language-dependent. However, there are a few rules that it seems prudent to include.

3.1 Gender and Number Agreement

Gender and number agreement can be powerful constraints on anaphoric references. In fact, [Mit02] p. 170-171 states that gender and number agreement

[6] The verbs are not yet used.

```
A lion, searching for prey, caught a goat on a hill.
------------------------------------------------------------------------
  .--- ndet       a1(1)            det sg indef
.----- subj(n)    lion1(2)         noun cn sg st_feline (lion/lion#1)
.----- vnfvp      search1(3,2,u,4) verb ving (search/search#6)
| `--- comp(p)    for1(4,5)        prep pprefv nonlocp pobjp
|   `- objprep(n) prey1(5,u)       noun cn sg (prey/prey#2)
o----- top        catch2(6,2,8,9)  verb vfin vpast sg vsubj (catch/catch#8)
| .--- ndet       a1(7)            det sg indef
`----- obj(n)     goat1(8)         noun cn sg st_mammal (goat/goat#3)
`----- comp(lo)   on1(9,11)        prep staticp
  | .- ndet       a1(10)           det sg indef
  `--- objprep(n) hill1(11)        noun cn sg st_place (hill/hill#4)
------------------------------------------------------------------------
It was devoured by the big cat.
------------------------------------------------------------------------
.------- subj(n)     it(1)          noun pron sg def perspron (goat/goat#3)
o------- top         be1(2,1,3)     verb vfin vpast sg vsubj stative
`------- pred(en)    devour1(3,4,1) verb ven vpass
  `----- subj(agent) by1(4,7)       prep pprefv
    | .- ndet        the1(5)        det sg def the ingdet
    | .- nadj        big1(6,u)      adj erest
    `--- objprep(n)  cat1(7)        noun cn sg st_feline (cat/lion#1)
------------------------------------------------------------------------
```

Figure 1: Annotated ESG Parses for Two-Sentence Discourse

alone accounted for a delta of about +30 percent in the success rate for the MARS system.

Number agreement is a difficult issue. In English, for example, some nouns like *council* and *family* indicate collective referents that can either be referred to by singular, following the morphological form of the antecedent, or by plural, following the semantic property of the antecedent. Currently, we are enforcing full syntactic number agreement, well aware that this may be too restrictive.

3.2 Generic Nouns

Identification of generic uses is also important since a specific *instance* of the generic class is not really coreferent with the generic class. Specific instances will have premodifiers that identify the subset or instance in question. There could be several instances in a discourse that need to be kept separate.

The following simple heuristics for classifying NPs as generic have been iden-

tified and implemented:

1. Plural common nouns not pre-modified by definite determiners or possessives (determiners or nouns): *Elephants are big.*
2. Singular common nouns filling the subject slot and not premodified by definite determiners or possessives (determiners or nouns): *Water is wet.*[7]
3. Nominal adjective phrases such as: *The best*, *His best*, and *The poor*.

3.3 Pleonastic *it* and Other Prop-*it*s

Recognizing pleonastic uses of *it* is important because treatment of *it* as a referring pronoun can have a bad effect for translation. Furthermore, wrongly ascribing a referent will affect the anaphoric chain, thus potentially wreaking havoc during further processing. A similar case holds for *it* used as the subject of the so-called "weather verbs" like *rain* and *snow* and for time expressions.[8]

3.3.1 Pleonastic *it*

[Mit02] and [GC01] indicate that identification of pleonastic *it* is a very difficult task and report low results on this. However, RAP shows a very good precision for this,[9] and a major feature that our system has in common with RAP is the use of a broad-coverage parser (ESG), whereas [Mit02] uses a POS tagger only, and [GC01] uses statistical means to identity pleonastic uses of *it*.

We have adapted the majority of the rules proposed by [LL94], viz. the following cases of extraposition, and their syntactic variants:

1. It be <Modaladj> that <S>
2. It be <Modaladj> (for <NP>) to <VP>
3. It be <Cogv>-ed that <S>
4. It seems/appears/ ... (that) <S>

Additionally, the following rule has been identified and implemented:

5. It be <Indef-NP> (for <NP>) to <VP>

[7] This will wrongly ascribe genericity in constructions such as *Water spilled on the table* where *water* really is only a mass noun.

[8] Some lump these uses of *it* together with the pleonastic use, but our view is that one could postulate *the weather* or *the time*, respectively as referent of *it*. This makes them exophoric rather than pleonastic.

[9] Lappin 2002, personal communication.

Rule (5) is an extension of rule (2) to cover such cases as *It was a mistake for him to do that.*

[LL94] indicates that the implementation depends on a list of actual modal adjectives, cognitive verbs, etc. Our implemention takes advantage of the slot-filling information provided by ESG to identify the constructions generally, without having to depend on checking for specific words.

Cleft constructions have an initial dummy *it*: *It was the dog who caught the mouse this time, not the cat.* This construction provides contrastive stress.

6. It be <NP> <Rel-Pron> <VP>

This rule is straightforward to implement.

These rules are clearly syntactic in nature; however, they appear to be shared among the languages of concern, though additional rules may need to be added for the other languages.

3.3.2 Idioms

Certain idioms also contain pleonastic pronouns: *Get it together.* These are language-specific by their very nature as idioms, and we will not address this issue further in this report.

3.3.3 Weather Verbs

A dictionary of "weather verbs" such as *rain*, *snow*, and *sleet* has been derived from WordNet to account for occurrences of *it* as the subject of these, and this use of *it* is recognized. Other uses of the so-called prop-*it* ([QGLS72]) include ambient conditions: *It was cold in the hall* and time expressions: *It is a quarter past 3.* These are not recognized yet.

3.4 Other Syntactic Rules

A few other syntactic rules have been identified and implemented:

1. A sister node of the pronoun cannot be an antecedent unless the pronoun is reflexive. This rule is probably language-independent.
2. A pronoun in a main clause cannot be coreferential with an NP in a subsequent dependent clause (see e.g. [Mit02] p. 30).
3. Multinouns can be identified syntactically by using techniques similar to those described in [BMW02]. This is crucial for providing the proper bag of antecedent candidates.

4 Discourse Rules

In this section we describe some general discourse rules that are useful in anaphora resolution.

4.1 Recency

[LL94] makes the observation that the sentence recency preference is the single most important factor for the RAP system, and [Mit02] p. 18 cites six studies giving empirical evidence that most antecedents are found very close to a pronominal anaphor, typically within three sentences. This is in contrast to non-pronominal anaphora, whose antecedents may occur at a much longer distance.

Consequently, we limit the search for antecedents of pronominal anaphora to the current sentence plus the last three preceding sentences, but let the search for antecedents of non-pronominal anaphora span the total document.

For selection of the best candidate, recency is given a fairly high prominence. The weight is basically the reciprocal of the sentence distance[10] between the sentence that the anaphoric expression occurs in and that of the proposed antecedent. We also experimented with a more fine-grained distance measure where the word distance *within* the sentences was taken into account, but that did not give as good results as the simpler measure.

4.2 Frequency of Mention

Frequency of mention is a good indication of the salience of a discourse entity (see e.g. [Mit02]). A frequently mentioned item is more likely to be foremost in the speaker's mind than an infrequent one. However, it is important to take into account that an entity can be referred to in many different ways, as we note in Section 5.3, and count the frequency not only on direct matches, but on the whole anaphoric chain, including pronominal references.

4.3 Speaker Identification

As explained in Section 2.2, speaker identification for direct speech is important. Euphoria uses the ESG parse to identify direct speech and the speaker by utilizing the semantic type marked on utterance verbs, and by exploiting the parse structure and the information in quotes.

We keep track of the current speaker separately from the other referential items, since it represents another dimension of the discourse.

[10] Reciprocal, because the scheme is that a higher number is a better number.

5 Semantic Rules

5.1 WordNet

WordNet is an English lexical ontology whose building block is the so-called *synset*, i.e. the synonym set that contains all the words that express a given concept. The synsets (and words) are related by other, explicit, semantic relations such as hyponymy, meronymy, entailment, and troponomy. EuroWordNet contains additional relations such as semantic constraints on verb arguments etc., and, since EuroWordNet includes a number of languages, it can also be viewed as a multilingual lexical database.

5.2 Semantic Type Checking

Semantic type checking provides a useful filter in that the antecedent should have a semantic type that is compatible with the anaphor. We have two sources of semantic types available; one is the type hierarchy provided with SG, the types of which are marked on the lexical entries. These types are made available for applications through the parse structure. The other source is WordNet.

By classifying the SG types into mutually exclusive sets, it is possible to utilize these types for type checking. This is done by checking that the semantic types of the antecedent candidate and the anaphor are not both in one of these sets. For example, under normal circumstances, `fem` and `masc` are mutually exclusive.

Unfortunately the organization of the type hierarchy appears not to have been designed with this purpose in mind,[11] and we have identified 27 disjoint sets, with more possible.

The SG type hierarchy comes with an *is-subtype* function. For each semantic type marked on a noun, the relevant disjoint sets are searched for conflict with the types marked on a proposed antecedent noun, searching the hierarchy as appropriate.

The WordNet hypernymy function provides additional possibilities for type checking.

For pronouns, the semantic type checking is used as follows. The numbers indicate contribution to the overall score; a higher number is better.

1. Pronoun is `fem` or `masc`:
   (a) Complete match between pronoun and antecedent: 1.0

[11] For example, the semantic type `size` has the following four subtypes `small`, `large`, `shallow`, and `deep`. Here `small` and `large` form a mutually exclusive set, and `shallow` and `deep` another set, but `small` and `shallow` are not incompatible.

else if

(b) WordNet: antecedent is a hyponym of `fem` or `masc` : 0.9
else if

(c) Antecedent is `person` and neither `fem` nor `masc`: 0.75
else if

(d) WordNet: Antecedent is a hyponym of `person`: 0.75

2. Pronoun is `neut`:

   (a) Wordnet: Antecedent not `person` (directly or hypernym) and antecedent is hyponym of `object` or of `animal` or of `plant`: 1.0
   else if

   (b) Antecedent neither `person` nor `fem` nor `masc`: 0.75
   else

   (c) 0.0

For non-pronominal anaphora, the 27 disjoint sets provide the semantic checking as described above.

5.3 Noun Phrase Coreference

Noun phrase[12] coreference is an interesting topic in itself. Whereas it is not an end in itself for our purposes, it is obviously necessary to consider coreference of NPs in order to be able to identify the most recent non-pronominal antecedent of a pronoun. This is useful e.g. when applying the sister rule.

We apply the following rules for resolving NP coreference, in this order. The contributions to the total score follow the scheme described in Section 6.3.

1. Identical NP (exact match)
   else if

2. Subsumption (generalization)
   else if

3. Synonym (WordNet; same level, all senses)
   else if

4. Antecedent is a hyponym of current NP (generalization) (WordNet; 3 levels, all senses)

[12] We shall here somewhat sloppily refer to non-pronominal NPs as noun phrases and use the term "pronoun" for pronominal NPs.

It has proved difficult, if not impossible, to map words into WordNet senses, so no sense distinction is attempted. Furthermore, WordNet synsets become very general at the upper levels, so a limit of three levels up was set. Obviously, this limit would be better if it was dynamically related to the level of the word we are exploring.

An example of hyponym antecedent was provided in Fig. 1. In (9) an example of synonym antecedent is given. WordNet gives *paper* and *composition* as synonyms.

(9) The student (student/student#1) turned in his (pronoun student/student#1) paper (paper/paper#6) at the end (end/end#2) of the semester (semester/semester#3). The composition(composition/paper#6) was 70 pages long.

5.4 Pronominal Anaphora

The following semantic tests are applied for pronouns:

1. Holonym for possessives (WordNet; 3 levels, all senses)
2. Predications match (WordNet; 3 levels, all senses)

An example of the use of the WordNet holonym function is provided in (10). The idea is that a *wing* is a part of a *bird*, and hence *bird* as antecedent of *its* is preferred.

(10) The bird (bird/bird#1) flew up into a tree (tree/tree#2) and flexed its (pronoun bird/bird#1) wings (wing/wing#7).

The predications match is described in Section 6.

6 Predications

[DI91] describes an approach to anaphora resolution that uses binary relations mined from corpora. These relations include subject-verb, verb-object, and adjective-noun relations. These relations provide preferences for selectional constraints. In other words, if there is a high frequency of collocational patterns of e.g. *eat vegetables*, and a low (or zero) frequency of collocations of *eat wood*, then, other things being equal, the system would prefer *vegetables* over *wood* as the antecedent of *it* in *eats it*.

The setup reported on in [DI91] has also been incorporated into RAP to boost RAP's performance, with a resulting improvement in performance of

about 2 percent. [GC01] reports on an experiment with similar binary relations. They conclude that no significant improvements were gained by this, partly because the many occurrences of *be* and *have* did not provide patterns that were selective enough.

Our approach significantly expands the notion of using statistics mined from corpora in three ways. First, we do not use binary-only relations, but utilize the full range of arguments that a word may take. This partially addresses the issue of *be* and *have* not being informative enough – just think of the difference between *Face is flushed* vs. *Face is* and *is flushed* separately. Our wide-coverage parser makes not only this possible, but also makes it possible to disregard the use of *be* and *have* as auxiliaries, which is bound to introduce a good deal of noise.

Second, we expand the statistics by WordNet relations. This allows us to gain much wider coverage.[13]

Third, we use as input to the acquisition phase not only large-scale web-crawling, but also WordNet itself.

In 6.1 we describe the statistics and their use, in 6.2 we describe expanding the coverage by WordNet relations, and in 6.4 we describe the use of WordNet as a knowledge source.

6.1 The Predications

The relations take the form of dictionary-style entries that have the head word of the phrase as the dictionary head word; the modifiers together with slot-filling information and frequency appear in the body of the entry. The body of the entry is actually a *list* of all relations relating to the head word.

A partial example is shown in 11. Here, *eat* is the head word, for which two different sets of modifiers were found. One set has the subject *man* and the object *potato* for which 1 occurrence was found. The other pattern has subject *woman* and object *rice*; this occurred 5 times. The frequency count collapses morphological variants.

(11) eat<(subj n man)(obj n potato)<1<(subj n woman)(obj n rice)<5

We shall refer to a head word and its modifiers as a "predication" since each relation can be viewed as a predicate, and the modifiers as its arguments. In this study, we have only used predications that show verbs and their complements.

The predications show slots and their fillers, but leave out other information such as adverbs. Thus negations and focalizers [McC87] like *always* and *never* are lost. For this reason, we cannot claim that the predications provide

[13] [Mit02] p. 175 mentions a similar idea in a footnote. We had this idea independently.

real-world knowledge, but they do give an indication of selectional constraints [QGLS72] p. 371 (or *pertinence* of a given complement to a given verb). Even if we accept as a real-world fact that "Tigers never eat hay", we *also* know that it makes sense to talk about tigers eating hay, whereas it does not make sense to talk of "books eating hay".

6.2 Extending the Predications Data Base

So far the predications have been mined from the Web. During this process we discovered that several different Web pages may contain exactly the same text. For example, a newspaper may have the same story in many different locations. Obviously, this creates a skewed picture of the frequency counts, so we devised a scheme to extract only *one* occurrence of each sentence and gather the predications on this reduced set.

The number of predications can be extended significantly by using WordNet relations to vary the different parts of the predication. Obviously, it is possible to vary both predicates and arguments, but in this study we have limited ourselves to varying the arguments. Useful relations include synonomy, antonymy, and hypernymy. For example, the data base may not include a perfect match, but does include a match with a synonym or hypernym.

6.3 Matching

In order to make the best use of the WordNet relations for matching against predications, it is important to decide on a good matching algorithm between what is actually in the predications database and the derived data. So if we have a sentence containing a pronoun like (12), the task is to find the antecedent providing the closest match of the clause with an entry in the predications database.

(12) The tiger ate it.

This is obviously related to the matching task of Example-Based Machine Translation (EBMT) [Som99].[14] EBMT is essentially translation by analogy. An EBMT system uses a database of example sentences in both source and target language to translate other, similar source-language sentences into the target language. There are many different ways of representing the examples, with as many different ways of matching. The method that comes closest to

[14] At first glance, the task also seems related to the BLEU translating evaluation metric [PRWZ02]. BLEU attempts to match MT output with one or more reference translations to determine the quality of the MT output, and takes the following factors into account: sentence length, word choice, and word order. However, the word choice measure, which would be the relevant part for us, only operates on *perfect* match. Variations like synonyms etc. are not considered.

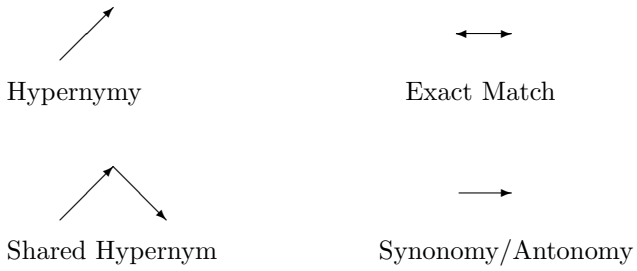

Figure 2: Allowable Paths

our WordNet relations is to measure the distance between the actual word and the example word in a thesaurus, counted as number of levels in the hierarchy. However, given that WordNet has several types of relations, not just the hierachical structure typical of thesauri, and that links in WordNet do not all reflect the same semantic distance [HSO98], this does not seem adequate.

[HSO98] reports on a more useful measure of distance, directly related to WordNet. They distinguish between different types of relations. Relations can be *upward* (hypernymy and meronymy), *downward* (cause, entailment, holonymy, and hyponymy), and *horizontal* (also-see, antonymy, attribute, pertinence, and similarity).

Extra-strong relations hold between a word and its literal repetion. *Strong relations* are synonomy, a horizontal link, and any kind of link between a word and a phrase that subsumes it. *Medium-strong relations* are calculated when there is an allowable path of between two and five links between synsets. The longer the path and the more changes in direction, the less confidence.

Given the unevenness of semantic distance, we allow only the paths diagrammed in Fig. 2, with a distance of 1 for each direction, except for hypernyms, where the maximum distance is 3. We categorize these relations in the same manner as [HSO98], with some additions and weights as shown below.

1. Exact match (extra strong): 1.0
2. Synonym (strong): 0.8
3. Subsumption (strong): 0.8
4. First person singular (strong): 0.8
5. Hypernym (medium): 0.8 - (match level/8)
6. Shares Hypernym (medium): 0.6
7. Holonym (medium): 0.8 - (match level/8)
8. Pleonastic: 0.0
9. Weather: 0.0

The same classification is used for the general rules reported on in the previous sections.

6.4 Using WordNet as a Source of Relations

For many synsets, WordNet gives an example sentence. These sentences can be extracted and used as a corpora in its own right. This is a possibility that still remains to be explored.

7 Selection of Best Candidate

Numerical scores are computed for each candidate based on recency, semantic type check, predication match, and frequency. Currently, the weights contributed by these factors are just added up. A statistical training method like linear regression analysis could be applied in the future.

8 Results and Plans

This section describes preliminary results and plans.

8.1 Implementation

The general framework has been implemented and hooked up to ESG. Semantic type checks have been devised and implemented. In type checking, it is useful to know when types are mutually exclusive. We have identified 27 mutually exclusive type sets from the SG/LMT ontology. Syntactic rules, including recognition of pleonastic *it* and weather verbs, have been implemented to the degree described in Section 3. Discourse rules have been implemented as described in Section 4. WordNet relations have been integrated as described in Section 5. The predications scheme described in Section 6 has been implemented but temporarily abandoned due to crashes in WordNet for larger files. The focus has been on third person personal and possessive singular pronouns, but also recognition of first person pronouns was made possible by speaker recognition.

8.2 Evaluation

Evaluation has two major purposes, *viz.* comparison of methods, and improvement of a given method. Comparison of different methods of anaphora resolution is very difficult because of the lack of a common standard. Several papers point out the need for common standards and make proposals for these, e.g.

[Mit02, Stu01, Byr01], indicating that the issue is far from settled. Given that this paper reports on work in progress, it seems more useful to use evaluation for diagnosis of problems with the current approach, with a view to improvements.

Euphoria was initially trained on a rather large, single, document, namely a short story from *The New Yorker*, that is, fiction. After training, the precision was approximately 75%. Error analysis indicated that a major problem was recognition of shift of focus. More specifically, there is a tension between the contribution (weight) of recency and that of frequency. Since the document is rather large, high frequency is a good bet, but frequency throws off the resolution when there is a change of focus from the overall focus of the document. Centering approaches such as described by [GJW83b, GJW83a] address the issue of focus change, but, as [PFM01] mentions, these tend to be very language-specific. A better way of dealing with change of focus may be to analyze the noun phrases of the first sentence of each paragraph. For obvious reasons, change of focus tends to coincide with paragraph change. Paying special attention to the first sentence of a paragraph is also a well-known technique in document summarization. We conjecture that a high degree of anaphoric phrases, particularly pronouns in initial position, is indicative of *continuation* of focus, whereas introduction of new reference items signal a change of focus.

The system was then tried out on a shorter, unseen text of the newspaper style from the web. The initial precision was very low, around 50%, and an error analysis revealed the need for *speaker identification*, as described in Section 4.3. Implementation of this improved the precision to approximately 85%. However, when the revised version of Euphoria was run on the *New Yorker* short story, it was very clear, without counting, that a degradation had taken place, and that it was due to inadequacies in the speaker recognition. The inadequacy consists in not keeping proper track of change of speaker in dialogues consisting of two persons taking turns, where there are no utterance verbs, but where the only clue to change of speaker consists in use of quotes to begin and end each speaker's turn.

The revised version was then tested on a few articles (unseen) from Wall Street Journal, which is a rather different genre than the training corpus. The precision of finding the most recent non-pronominal antecedent was 55%. Here again, the major problem appeared to be proper identification of speaker, as reflected in worse scores for the pronoun *I*. Also *it* presented some problems, whereas the best results were obtained with *he*, *his*, *she* and *her*.

8.3 Plans

As indicated in the preceding section, more work needs to be done on dialogue management when there is no utterance verb. This will also help resolution of second person pronouns (even though it will not cover the cases where *you* is used "generically").

We will also work on shift of focus, as described above.

Not all known pleonastic constructions are implemented yet, and this definitely seems like a useful thing to pursue, for the reasons given in Section 3.3.

Furthermore, we want to investigate the problem with the predications part to determine which WordNet functions cause the problem, and determine the degree of their usefulness. Other ways of automatically deriving selectional constraints should also be considered.

In order to accurately identify coreference for non-pronominal noun phrases a good treatment of appositives is needed. We have already done some work in this area, but more is needed.

Expansion of coverage of types of pronouns will also be addressed in the future.

Plurals pose their own problems. E.g. the antecedent may be split among two or more entities: *The man$_i$ called his dog$_j$, and they$_{i+j}$ went on a walk.* Or the antecedent may be plural and the anaphoric expressions split up: *The animals$_{i+j}$ played wildly. The dog$_i$ ended up biting the cat$_j$.*

Expanding the coverage to include other languages can be addressed in two ways. We can adapt Euphoria to EuroWordNet, but it is also worth while investigating if useful results can be obtained by using our bilingual dictionaries to provide transfer into English and then use English WordNet on that.

References

[BMW02] Arendse Bernth, Michael McCord, and Kara Warburton. Terminology extraction for global content management. Technical Report RC22615, IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, November 2002.

[Byr01] Donna Byron. The uncommon denominator: A proposal for consistent reporting of pronoun resolution results. *Computational Linguistics*, 27(4):569–577, 2001. Special Issue On Anaphora Resolution.

[DI91] Ido Dagan and Alon Itai. A statistical filter for resolving pronoun references. In Y. A. Feldman and A. Bruckstein, editors, *Artificial Intelligence and Computer Vision*, pages 125–135. Elsevier Science Publisher North-Holland, 1991.

[Fel98] Christiane Fellbaum, editor. *WordNet: An Electronic Database*. MIT Press, Cambridge, Masschusetts, 1998.

[GC01] Niyu Ge and Eugene Charniak. Statistical anaphoric pronouns resolution. Technical report, Brown University, 2001.

[Gea68] Peter T. Geach. *Reference and Generality*. Cornell University Press, Ithaca, 1968.

[GJW83a] Barbara Grosz, Aravind Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):201–225, 1983.

[GJW83b] Barbara Grosz, Aravind Joshi, and Scott Weinstein. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual meeting of the Association of Computational Linguistics (ACL'83)*, pages 44–50, Cambridge, MA., 1983.

[HSO98] Graeme Hirst and David St-Onge. Lexical chains as representations of context for the detection and correction of malappropisms. In Christiane Fellbaum, editor, *WordNet: An Electronic Database*, pages 305–332. MIT Press, Cambridge, Masschusetts, 1998.

[Kar69] Lauri Karttunen. Pronouns and variables. In *Papers from the Fifth Regional Meeting of the Chocago Linguistic Society*, pages 1108–116, Chicago, 1969. Chicago Linguistic Society.

[KB96] Christopher Kennedy and Bran Boguraev. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on on Computational Linguistics (Coling96)*, pages 113–118, Copenhagen, Denmark, 1996.

[KR93] Hans Kamp and Uwe Reyle. *From Discourse to Logic*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.

[LL94] Shalom Lappin and Herbert J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.

[MB98] Michael C. McCord and Arendse Bernth. The LMT transformational system. In *Proceedings of AMTA-98*, pages 344–355. Association for Machine Translation in the Americas, 1998.

[McC80] Michael C. McCord. Slot Grammars. *Computational Linguistics*, 6:31–43, 1980.

[McC87] Michael C. McCord. Natural language processing in prolog. In Adrian Walker, editor, *Knowledge Systems and Prolog*, pages 291–402. Addison-Wesley, Reading, MA, 1987.

[McC90] Michael C. McCord. Slot Grammar: A system for simpler construction of practical natural language grammars. In R. Studer, editor, *Natural Language and Logic: International Scientific Symposium, Lecture Notes in Computer Science*, pages 118–145. Springer Verlag, Berlin, 1990.

[Mit02] Ruslan Mitkov, editor. *Anaphora Resolution*. Longman, London, 2002.

[PFM01] Manuel Palomar, Antonio Ferrández, Lidia Moreno, Patricio Martínez-Barco, Jesús Peral, Maximiliano Saiz-Noeda, and Rafael Muñoz. An algorithm for anaphora resolution in spanish texts. *Computational Linguistics*, 27(4):555–567, 2001. Special Issue On Anaphora Resolution.

[PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, July 2002. University of Pennsylvania.

[QGLS72] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A Grammar of Contemporary English*. Longman, London, 1972.

[Som99] Harold Somers. Reviev article: Example-based Machine Translation. *Machine Translation*, 14:113–157, 1999.

[Stu01] Roland Stuckardt. Design and enhanced evaluation of a robust anaphor resolution algorithm. *Computational Linguistics*, 27(4):479–506, 2001. Special Issue On Anaphora Resolution.