Rule-based reference resolution for unrestricted text using part-of-speech tagging and noun phrase parsing.

Sandra Williams, Mark Harvey and Keith Preston, BT Labs.

Abstract

This paper describes an experimental syntactic rule-based method for reference resolution in unrestricted texts. References can be resolved automatically and this overcomes a major hurdle in text analysis and provides a key advantage in text `understanding' and information extraction.

A shortcoming of systems that locate and extract sentences from unrestricted text to help people assimilate information, and cope with `information overload', arises from references to sentences not present in the extract. Reference resolution is a technology that can be used in such systems and one of the motivations for the current work was to investigate ways of identifying and resolving such references, thus allowing us to enhance the performance and usability of these systems.

The reference resolution system accomplishes two tasks: firstly it identifies noun phrases, either as references or as non-references, and secondly it resolves references. The identification of reference/non-reference noun phrases is achieved in two stages: 1) part-of-speech tagging; 2) noun-phrase parsing. The reference resolution is achieved with rules and knowledge-bases of names, titles, and General Knowledge.

Tests results show that up to 97% of reference and non-reference noun phrases are correctly identified; and up to 76% of all references that the system attempts are resolved correctly within the first three estimates. Of these, up to 61% are resolved correctly on the first estimate.

1. Introduction

As computer and telecommunications technologies continue to advance, people are able to create, store and access more information than ever before. Businesses now depend on accurate, detailed and timely information in order to survive, while at home we find that information is playing an increasingly important role in education and leisure activities. However, each technological advance brings its own problems, and in this case, we have reached a state where people simply cannot cope with the rapidly accumulating mountains of information. We already suffer from 'information overload' and the problem is getting worse.

If we are to cope with the growing information overload, we need more than search engines and document retrieval systems; we need systems to help us to assimilate information, and ultimately systems which understand information for us [1].

Much of the information available today is in the form of unstructured texts, or texts loosely structured in SGML (Standard Generalised Mark-up Language) or HTML (Hypertext Mark-up Language). Information is available on a vast range of subjects from a wide range of sources. Therefore it is important that systems are able to handle unrestricted texts, and can analyse and `understand' those texts in as much depth as is necessary to aid people to locate and assimilate the information they need.

NetSumm [2] is a Text Abridgement program developed at BT Laboratories, which automatically picks out the "most important" sentences from documents. The summarisation

process is interactive, allowing the user to choose longer or shorter extracts at will, ranging from a single sentence to the full text. The techniques used by NetSumm appear to be very robust, giving it the ability to work on almost any text, independent of subject. Equally good results have been achieved with articles on such disparate subjects as semiconductor lasers and red squirrels!

One of the shortcomings of NetSumm, and indeed any system that extracts sentences from a longer text, arises from references to sentences not present in the extract. One of the motivations for the current work was to investigate ways of identifying and resolving such references, thus allowing us to enhance the performance and usability of NetSumm in the future. It was also realised that if we can resolve references, we will overcome a major hurdle in text analysis and provide a key advantage in text `understanding' and information extraction.

Since one of NetSumm's key features is its ability to handle almost unrestricted texts, we wanted to maintain this capability in the reference identification and resolution system. This led us to concentrate on "shallow" syntactically-based techniques, although we recognise that references are essentially a semantic phenomenon [3].

Research projects which do take a deeper, semantic approach tend to be limited to a small domain, e.g. Dahl [4], or operate on `toy' texts, e.g. Carbonell and Brown [5]. Many research projects concentrate solely on pronoun resolution, e.g. McEnery and Bench-Capon [6], Dagan and Itai [7] and Hobbs[8]. Our broader coverage is similar to that of Leass and Schwall [9] who have implemented a system for machine translation of unrestricted German text. Unfortunately we cannot compare our results to theirs since they did not report on their success rate.

In this paper we discuss the theoretical background to our reference identification and resolution system (see part 2). In part 3 we describe the system itself and explain how it identifies and resolves textual references both within a sentence and between sentences. Our results are presented and discussed in part 4 and further discussion and future improvements in part 5.

2. TYPES OF REFERENCE

2.1 Identifying References

References occur in articles and documents when the author makes mention of something which can only be uniquely identified by looking elsewhere in the text, or, in some cases, even outside the text.

Syntactically, references will usually correspond to noun phrases, though not every noun phrase will be a reference. For example, in the sentences:

Mary saw a cat. The animal was black.

the noun phrase "a cat" is not a reference, though the noun phrase "the animal" is (it refers to "a cat" in the previous sentence). In the rest of this paper, we use the term Non-Reference Noun Phrase (NRNP) for noun phrases like "a cat", and Reference Noun Phrase (RNP) for noun phrases like "the animal" in the above example. As we show later, it is possible to identify and distinguish NRNPs and RNPs syntactically, using robust parsing techniques.

2.2 Types of Reference

Halliday and Hasan [3] say of references: `*These items are directives indicating that information is to be retrieved from elsewhere.*' If the information is located further back in the text, then the reference is termed `anaphoric'. If it is further forward in the text, it is termed `cataphoric' and if

it occurs outside the text it is termed `exophoric' [3].

1. Jill enjoys eating cakes. They are very fattening.

2. The child, named Anna, smiled at Bob who was looking for another child.

3. Of all trees in the world, John prefers willows.

In example 1 above, `they' is an anaphoric reference, referring back to `cakes'. In example 2, `the child' is a cataphoric reference, referring forwards to `Anna'. Whilst in example 3, `the world' is an exophoric reference where the meaning is not found from information within the text but it is generally understood by mutually shared knowledge.

Halliday and Hasan [3] identify three major groups of references: personals, demonstratives and comparatives. These groups, which are largely orthogonal to the classification of references as anaphors, cataphors and exophors, are summarised in Table 1 below:

Personal Reference	I, me, mine, my, you, yours, your, we, us, ours, our, he him, his, she, he, hers, they, them, theirs, their, it, its, one, one's
Demonstrative Reference	this, these, here, now, that, those, there, then, the
Comparative Reference	same, identical, equal, similar, additional, other, different, else, better, more, identically, similarly, likewise, so, such, differently, otherwise, so, less, equally

Table 1: Types of reference as defined by Halliday and Hasan [3]

Personal references indicated in Table 1 include pronouns and possessive pronouns. We also include reflexive pronouns (myself, yourself, himself, itself, etc.) in this group. In example 1 above, we can now identify `they' more fully as a personal anaphoric reference. Demonstrative references are indicated by determiners and adverbs as shown in Table 1. Comparative references are indicated by certain adjectives and certain adverbs as indicated. So we can now identify in example 2, above, a demonstrative cataphoric reference `the child' and a comparative reference `another child' (resolved by a child who is not the child Anna). In example 3, `the world' is an exophoric demonstrative reference.

It is not possible to distinguish anaphoric, cataphoric and exophoric references by syntactic means alone. Consider the sentence below:

Mr X said: if [the merger] goes through [it] will be [the chance] of a lifetime, but [it] would be asking for [the moon].

The five references in this sentence `the merger', `it', `the chance', `it' and `the moon' are denoted by square brackets. Three are syntactically identical noun phrases, but of these three `the merger' is anaphoric (resolved by `a merger' in an earlier sentence), whilst `the chance' is cataphoric (it is uniquely identified by the following prepositional phrase `of a lifetime') and `the moon' is exophoric (it is a unique entity widely known through shared knowledge and experience). For this reason we have included cataphoric and exophoric references, as well as anaphoric references, in our system.

3. Implementation

The analysis carried out by our system can be divided into two parts, the identification and

distinguishing of NRNPs and RNPs, followed by reference resolution.

The system, shown diagrammatically in Figure 1, consists of three major processing modules:

* a part-of-speech tagger;

* a chart parser and grammar which identifies and distinguishes between RNPs and NRNPs;

* and a reference resolver.

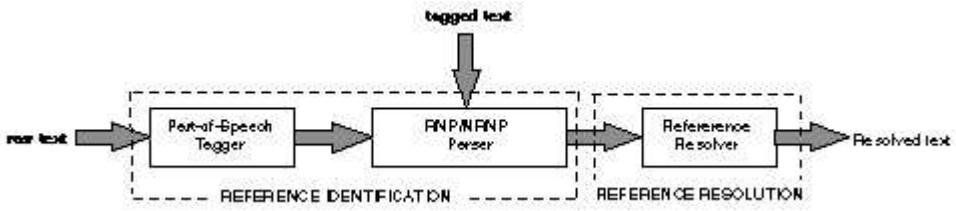

Fig. 1 System Diagram showing flow of data

Reference identification is performed by the first two modules which together implement a robust syntactic analyser, and reference resolution by the third. Note that tagged text may be introduced after the part-of-speech tagger, allowing us to evaluate the performance of the system in the absence of tagging errors.

The output of the Reference Identification stage (with part-of-speech tags omitted) is illustrated in the fragment of LOB a1 file [10] below:

<a move> to stop <Mr Gaitskell> from nominating <any more labour life peers> is to be made at <a meeting> of <labour MPs> <tomorrow>.

<Mr Michael Foot> has put down <a resolution> on [the subject] and [he] is to be backed by <Mr Will Griffiths> ,<MP> for <Manchester Exchange>.

though [they] may gather <some left-wing support> ,<a large majority> of <labour MPs> are likely to turn down <the Foot-Griffiths resolution>.

RNPs are marked with square brackets `[]' and NRNPs are marked with angled brackets `<>`.

The output from the Reference Resolution stage consists of an ordered list of hypothesised resolvers for each RNP attempted.

3.1 Part-of-speech tagging

A statistical part-of-speech tagger based on that proposed by Church [11] was used to tag the raw text. The tagger uses both lexical frequency and bigram word probabilities, as well as rules for handling unknown words and idioms.

We used an augmented LOB Corpus [10] tag set with extra tags added where we felt that the existing tags did not allow us to write grammar rules for our parser to distinguish sufficiently between those noun phrases which are references and those that are not. New tags were added for the following categories:

* **Singular and plural articles.** Some of these are not normally present in NP references (e.g. `no' in `no man') whilst others may well be (e.g. `the' in `the man').

* **Nominal pronouns.** Nominal pronoun like `so' is normally a reference as in `it ain't

necessarily so' whilst other nominal pronouns, such as `anybody', are not.

* **Singular determiners.** Singular determiners, such as `this', tend to be present in NP references whereas `each' does not.

* **Pre-qualifiers.** Pre-qualifiers like `such', in a noun phrase, are normally present in NP references, whereas other pre-qualifiers like `quite' are not.

Since our additional tags were chosen in a way which allowed them to be derived from combinations of words and their existing tags, we were able to retag the entire LOB corpus with our extended tag set automatically. We then retrained our part-of-speech tagger on this data.

3.2 Reference/Non-Reference Noun Phrase Parsing

Sentences from the tagger are analysed using a chart parser with a grammar which identifies and differentiates between RNPs and NRNPs. While we do not claim that the grammar rules we have devised are general rules for all English, they appear to work very well for this task in the majority of cases.

The grammar identifies as RNPs all the elements shown in Table 1 above**.** All other noun phrases are marked as NRNPs. These include noun phrases containing:

* capitalised nouns (e.g. `Internet' in `the Internet'),

* units of measurement (`km', etc.),

* titles (Mr, Mrs, etc.),

* proper names (e.g. `Michael Foot'),

* locative nouns (e.g. `bridge' in `the Bridge of Sighs'),

* compound nouns (e.g. `labour life peers'),

* adjectives (these more precisely specify the nouns they pre-modify, Quirk et al. [12]).

3.3 Reference Resolution

After eliminating within-sentence cataphoric references as described in 3.3.1 below, the system performs general anaphoric reference resolution by searching back through the text until it finds matching items. The matching items for different kinds of reference are described below in sections 3.3.2 and 3.3.3. Exophoric references are resolved as described in 3.3.4. The system does not at present resolve comparative references although it does identify them (see 3.1 and 3.2).

3.3.1 Within-sentence cataphoric reference resolution

Intra-sentential cataphoric demonstrative references are resolved by their immediate context. In order to do this we use three rules derived from Quirk et al. [12] to identify phrases which uniquely define the noun:

* Demonstrative reference identified by immediately following prepositional phrase: e.g. `those men **in the garden**'

* Demonstrative reference identified by immediately following ING verb phrase: e.g. `that girl

sitting down'

* Demonstrative reference identified by immediately following relative: e.g. `the chair **that Jack made**'.

Of course the application of these rules does not work in all cases. In the case of prepositional phrases (PPs), in `I hit the man **with the red shirt**', the PP `the red shirt' specifies which man, but in `I hit the nail **with the hammer**', the PP doesn't specify which nail. However, the PP seems to function as an NP modifier in the vast majority of cases in the texts we have studied rather than functioning as, for instance, a verb argument.

3.3.2 Demonstrative reference resolution

The system performs resolution by searching for co-occurrences of the head noun in a noun phrase reference with the head nouns of previous noun phrases. The two nominal heads must agree in number. Consider the sentences:

John bought a new car. He crashed the car the following week.

The reference `the car' in the second sentence is resolved by matching the singular nominal head, `car', with the singular nominal head `car' of `a new car' in the first sentence.

3.3.3 Personal reference resolution

Currently our system handles third person personal references only.

To resolve third person singular `his', `him', `her', `he', `she', `himself' and `herself' references, a database of common male and female names and a database of titles (Mrs, Dr, etc.) are held. These are searched to find any matches with contents of previous noun phrases in the text. Depending on the title used and the name, potential resolver noun phrases can be identified as male, female, or unisex. The system matches `her', or `hers' with a previous female name, and similarly for male references. If it cannot find a match, it will match with a unisex name, but not with a name of the opposite gender.

Third person singular references `it', `its' and `itself' are resolved by searching back in the text for singular noun phrases. A within-sentence example occurs in:

When a liquid is heated its temperature rises.

where the system resolves `its' with the singular noun phrase `a liquid'.

Similarly, third person plural references `them', `they', `theirs', etc. are resolved by searching back through the text for plural noun phrase resolvers. There is a problem when `they' refers to two or more persons who have been named previously in the text. The resolution is simple if the names are conjoined by `and' as in:

*`Jack and Jill went up the hill where **they** found a well',*

where `they' is resolved by `Jack and Jill'. But choosing which named persons are to be included, and which should not, in a more complex case like:

*A move to stop Mr Gaitskell from nominating any more labour life peers is to be made at a meeting of labour MPs tomorrow. Mr Michael Foot has put down a resolution on the subject and he is to be backed by Mr Will Griffiths MP for Manchester Exchange. Though **they** may gather some left-wing support...*

is much more difficult. In this case, `they' is resolved by `Mr Michael Foot' and `Mr Will Griffiths' but our system has no way of knowing that `Mr Gaitskell' ought not to be included as well. This would require semantics to understand the relationships described in the text. Our system does not yet attempt to resolve `multiple named persons' references.

With possessive personal references, the system attempts to resolve both the personal and the nominal head e.g. `his negotiations' in the sentence:

African delegates to the talks yesterday called on Mr Macmillan to cease his negotiations with Sir Roy's representative, Mr Julius Greenfield.

The system resolves the person referred to by `his' as Mr Macmillan, and it tries to link `negotiations' with a mention of `secret negotiations' 5 sentences back (incorrect in this case). The method for resolving the nominal head is as described in 3.3.2, demonstrative reference resolution.

3.3.4 Exophoric reference resolution

Exophoric references are what Quirk et al. [12] refer to as `larger situation' references, where `the identity of the referent may be evident from knowledge of the larger situation which speaker and hearer share' (page 266). The system resolves exophoric references by searching a database of definite article noun phrases in common usage, e.g. `the world'.

4. Results

Before discussing the results of our system it will be helpful to reiterate the ways in which we envisage it being used. One of the main initial applications is seen as enhancing the performance of the NetSumm text summariser, by presenting users with a short list of possible resolvers for each RNP in the extracted sentences. In this context, users will be able to use their own intelligence and judgement to determine which is correct, and we therefore do not require 100% accuracy in reference resolution. Other applications may impose more stringent requirements on accuracy.

We show results taken from two different texts. One consists of 50 sentences from the LOB file a1, a collection of newspaper articles. The other, from the New Scientist journal, is a text about a method for cooling silicon chips. The latter text was unseen in that it was not used to train the tagger, nor was it used to develop the RNP/NRNP grammar rules.

4.1 Part-of-Speech Tagging Results

The file a1 was already tagged since it was taken from the tagged LOB corpus, so the part-of-speech tagging was bypassed and the tags were (in theory) 100% correct.

Tags assigned by the part-of-speech tagger to the lexical items in unseen texts were not always correct. Particular problems occur when verbs are tagged incorrectly as common nouns. An example occurs in the sentence fragment below where tags are shown as codes following an underscore appended to each word:

Packing_VBG large_JJ numbers_NNS of_IN transistors_NNS into_IN less_AP space_NN results_NNS in_IN ...

where `results' has been tagged incorrectly as a plural common noun (NNS).

4.2 Identification of RNPs and NRNPs

4.2.1 LOB a1

The RNP/NRNP parser is very successful in correctly identifying reference noun phrases and non-reference noun phrases. In fact it gets them correct most of the time (97% were correctly identified for the tagged LOB a1 file) as shown in Table 2:

	Correctly Identified	Incorrectly Identified
Reference NPs	20%	1%
Non-reference NPs	77%	2%
TOTALS	97%	3%

Table 2: Results for Identification of Reference NPs and Non-Reference NPs for LOB a1

The few incorrectly identified non-reference NPs were due to incorrect boundary insertions in the centre of noun phrases so that two NPs were marked. An example occurs in the non-reference noun phrase:

`Mr Kenneth Kaunda's united national independence party'

which has been split into two: `Mr Kenneth' and `Kaunda's ... party'. The grammar tends to overgenerate on some very long noun phrases of this kind. It could be altered so that this NP would be parsed correctly, but that would result in some shorter noun phrases which are adjacent to one another being incorrectly amalgamated into single long noun phrases.

The incorrectly marked reference NP was due to a non-referring `it' being marked as a reference. There is no LOB tag to symbolise this kind of `it' (called `prop' it by Quirk et al [12]) which occurs in sentences such as `It is raining', so we are unable at present to distinguish the reference and non-reference uses of `it'.

4.2.2 Unseen text

Again the RNP/NRNP parser is very successful in correctly identifying reference noun phrases and non-reference noun phrases, with 93% of them correctly identified as shown in Table 3:

	Correctly Identified	Incorrectly Identified
Reference NPs	34%	1%
Non-reference NPs	62%	3%
TOTALS	96%	4%

Table 3: Results for Identification of Reference NPs and Non-Reference NPs for the unseen text

Unlike the LOB a1 text, which has been correctly tagged by hand, the unseen text was tagged by the part-of-speech tagger. The parser is, of course, totally dependent on the tags being correct. The few mistakes made by the parser were mainly due to tagging errors. For instance, the verb `results' was tagged incorrectly as a noun causing it to be included in the spurious noun phrase `less space results'.

4.3 Reference Resolution Results

4.3.1 LOB a1

Figure 2 shows the results for resolving the references in part of LOB a1.

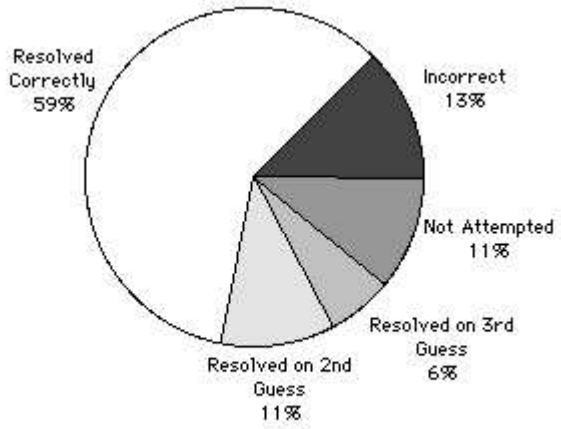

Fig. 2: Reference resolution results for LOB a1 (newspaper articles)

The pie-chart is divided into 5 regions as follows:

Resolved Correctly	The NP resolvers of these references were correctly identified with the first match found.
Resolved on 2nd Estimate	The NP resolvers of these references were the second match found.
Resolved on 3rd Estimate	The NP resolvers of these references were the third match found.
Not Attempted	References which the system does not attempt to resolve as yet.
Incorrect or Not Found	No resolvers were found or those suggested by the system were incorrect.

Note that when this text was parsed, 2% of the non-reference NPs were marked incorrectly. These were not artificially corrected before the reference resolution and, as it turned out, they did not adversely affect the results. The 1% incorrectly marked reference NPs, on the other hand, did adversely affect the result.

The system has correctly resolved 59% of all references present in this part of LOB a1. An example of a correctly resolved anaphoric reference is `he' resolved by `Mr Michael Foot' in the sentence:

Mr Michael Foot has put down a resolution on the subject and he is to be backed by Mr Will Griffiths, MP for Manchester Exchange.

By far the most frequently occurring kinds of reference in this text are `the' demonstratives (32% of all references) and third person male personal references (28%), the majority of which have been resolved correctly.

In some cases the system has found the correct resolver with its second choice (11%) or third

choice (6%). As mentioned earlier, the usefulness of this entirely depends on the kind of application in which the reference resolving system is to be used. If the application is interactive, then the user can be presented with a small set of alternatives from which they can choose. If an application is for a small domain, then extra domain knowledge could allow semantic restrictions to determine which is correct.

11% of the references in this text fall into categories that the system does not yet attempt. These consist of first person personal references (`I' and `we'), adverbial demonstratives (`now'), and comparatives introduced by `another'. The proportion of comparative references in this text is very small (only 2%). First and second person personal references make up 6% of the total, and demonstrative adverbs 4%.

13% of references are labelled `Incorrect or Not Found'. The few that were incorrect include: a non-referring `it' which the system attempted to resolve; and a co-occurrence of a head noun which did not resolve a reference with the same head noun. Those that were `not found' are demonstratives references where the head noun is a new word introduced into the text to refer to something previously described by some group larger than a noun phrase. An example of this is `the crisis' functioning as a reference to a situation described by a group of preceding sentences.

The majority of references in this text, occur in the same sentence as their resolvers (42%), or one sentence back (27%). It is worth noting, however, that the system successfully resolved references when their resolvers are up to six sentences back in the text.

4.3.2 Unseen text

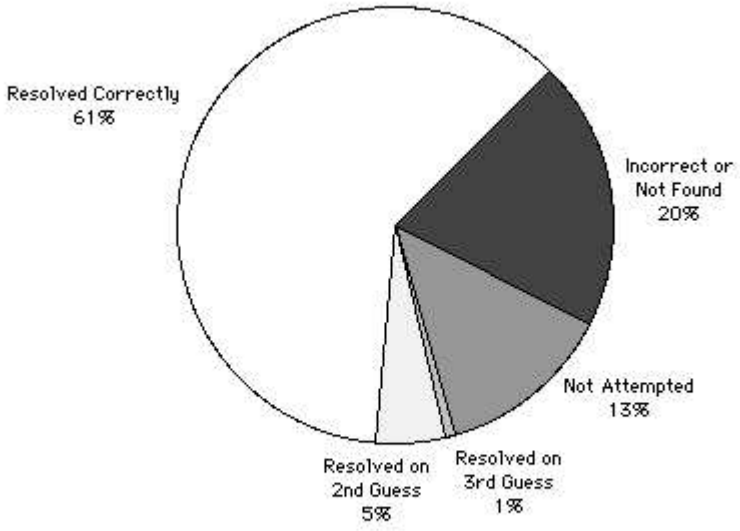

Fig. 3: Reference resolution results for unseen text (New Scientist article)

These results compare very well with the results for LOB a1 with 61% correct compared to 59% in LOB a1. Note that these results include the 4% incorrectly parsed noun phrases mentioned in 4.2.

The percentage of incorrect or not found, (20%), is higher than for LOB a1 (13%). Twice as many are not found rather than being incorrect. The ones that are not found are mainly due to new words being introduced to the text to refer to previous groups of sentences or verb phrases,

e.g. `the cooling' referring to `to carry away huge amounts of heat' five sentences back. Some of the incorrect resolutions have occurred because the system has encountered a new name, For instance Issam Mudawar, which was not present in the system's database of names, so it was not recognised as a name. References to `he' and `his' were then incorrectly resolved by another male name placed further back in the text.

The vast majority of the references in this text (63%) are `the' demonstratives. This is unsurprising for scientific writing. There are very few of each of the other types of references.

The majority of references are again either referring to items in the same sentence (42%), or one sentence back (20%), However a few were six sentences back and one was nine sentences back.

5. Conclusions

When we set out to build a system to resolve all references in any text, regardless of domain, style, or genre, we did not set ourselves an easy task! Reference resolution is not even easy for humans sometimes: `*While deictic reference and ellipted matter must, from a grammatical viewpoint, be recoverable, discourse permits a good deal of vagueness. This seems to be actively cultivated in propaganda and other persuasive material...*': page 1463, Quirk et al. [12].

Good results can be achieved with a purely syntactic algorithm. The reference resolution system described here has two major parts: **reference identification** (performed by the part-of-speech tagger and reference/resolver parser) and **reference resolution** (performed by the reference resolution component). The reference identification part of the system performs very well. Any spurious output it produces tends to be due to the part-of-speech tagger rather than the parser. The limitations of the reference resolver are, on the whole, due to the categories of reference that it does not attempt at the present time, namely: demonstrative adjuncts, comparatives, and first and second person personals. It is also limited to noun phrases resolvers at the present time. We will concentrate our future efforts in these areas.

Acknowledgements

We are grateful to the members of the BT Labs NLP Group for their help, encouragement and advice during this work. In particular we would like to thank Edward Kaneen and Steve Minnis for their work on the part-of-speech tagger.

References

1 . Preston, K. R.: `From Books to Bytes - Managing Information in the Information Age', British Telecommunications Engineering Journal, Vol 15, April 1996.

2. http://www.labs.bt.com/innovate/informat/netsumm/

3. Halliday, M.A.K. and Hasan, R. `Cohesion in English', Longman, Harlow, 1976.

4. Dahl,D.A. `Focusing and reference resolution in PUNDIT', AAAI, 1986.

5. Carbonell, J.G. and Brown, R.D. `Anaphora Resolution: a multi-strategy approach', COLING, Budapest, 1988.

6. McEnery, A.M. and Bench-Capon, T.J.M. `The WEB cohesive tie marking system: an expert system to resolve pronoun reference', Expert Systems for Information Management, Volume 2, No. 3, 1989.

7. Dagan, I and Itai, A. `A statistical filter for resolving pronoun references', Artificial Intelligence and Computer Vision, Y.A. Feldman and A. Bruckstein (Eds), Elsevier, 1991.

8. Hobbs, J.R. `Resolving pronoun references', Readings in Natural Language Processing', B.J.Grosz, K. Sparck Jones, and B.L.Webber (Eds), Morgan Kaufmann, 1986.

9. Leass, H. and Schwall, U. `An anaphora resolution procedure for machine translation'. IBM Germany Science Center, Institute for Knowledge Based Systems, Report 172, 1991.

10. Johansson, S. et al. `The tagged LOB corpus: Users' manual', Norwegian Computing Centre for the Humanities, Bergen, 1996.

11. Church, K.W. `A stochastic parts program and noun phrase parser for unrestricted text', Proceedings of the Second Conference on Applied Natural Language Processing (ACL), 1988.

12. Quirk, R. Greenbaum, S. Leech, G. and Svartvik, J. `A comprehensive grammar of the English Language', Longman, Harlow, 1985.