

Coreference Resolution of Named Entities and Noun Phrases in Web Pages

Nick Weaver

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

N.Weaver@dcs.shef.ac.uk

Yorick Wilks

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

Y.Wilks@dcs.shef.ac.uk

Abstract

An approach for intra-document coreference resolution of named entities and noun phrases is proposed. This approach is a knowledge-poor, integrated approach to coreference resolution which relies on syntactic, discourse and semantic information (using WordNet). Our approach is also intended to exploit the structural features of web pages for the purposes of discourse analysis. This research is in its preliminary stages, implementation and evaluation having not yet been completed.

1 Introduction

This paper proposes an approach for coreference resolution of named entities and noun phrases in web pages. The application of information extraction to the web is of paramount importance for the development of the semantic web (Ciravegna et. al, 2003). Coreference resolution is an important sub-module in information extraction since it enables partially filled template data objects about the same entities and entity relationships described at different discourse positions to be merged to create a network of related data objects (Kameyama 1997).

Web pages usually contain short, choppy sentences, textual fragments and a semi-structured style as opposed to the fuller and more grammatical style found in free texts. These features are notoriously difficult for automatic systems to process. Unlike free texts, web pages frequently have information crammed into them in an abbreviated, informal and graphically appealing style

that focuses on presentation and layout rather than content itself. The Internet's lack of regulation, its informality, poor grammatical quality and multi-media features combine to create a disordered style which is very hard to break down into a systematic form that could be easily processed by an automatic system. Attempts to adapt information extraction techniques and their sub-processes (including coreference resolution) to the textual style of the Internet have proved largely unsuccessful for these reasons (Soderland 1997). The methodology for intra-document coreference resolution proposed in this paper is designed to be able to cope with the style of web pages and even to exploit some of the structural features unique to web pages. The principle of using a knowledge-poor approach, for example in the use of the output of a part of speech tagger, is key to this design philosophy. Also key is the exploitation of the structured style of web pages for the purposes of discourse analysis. As a consequence, it is hoped that the approach to coreference resolution proposed in this paper will be well suited to the problematic domain of the Internet.

We have decided in our research to focus upon the problem of named entity and noun phrase coreference resolution. This is due to the great deal of research which has already been done in the area of pronominal coreference resolution, and also because a module for pronominal coreference resolution (Dimitrov 2002) already exists in GATE, the General Architecture for Text Engineering produced by our research group (Cunningham et al, 2002). The kind of general discourse analysis proposed in our research is also better suited to nominals and named entities than to pronominal anaphora, as will be explained later. The difficulty of the named entity and noun phrase resolution task is also of interest from the point of view of new research.

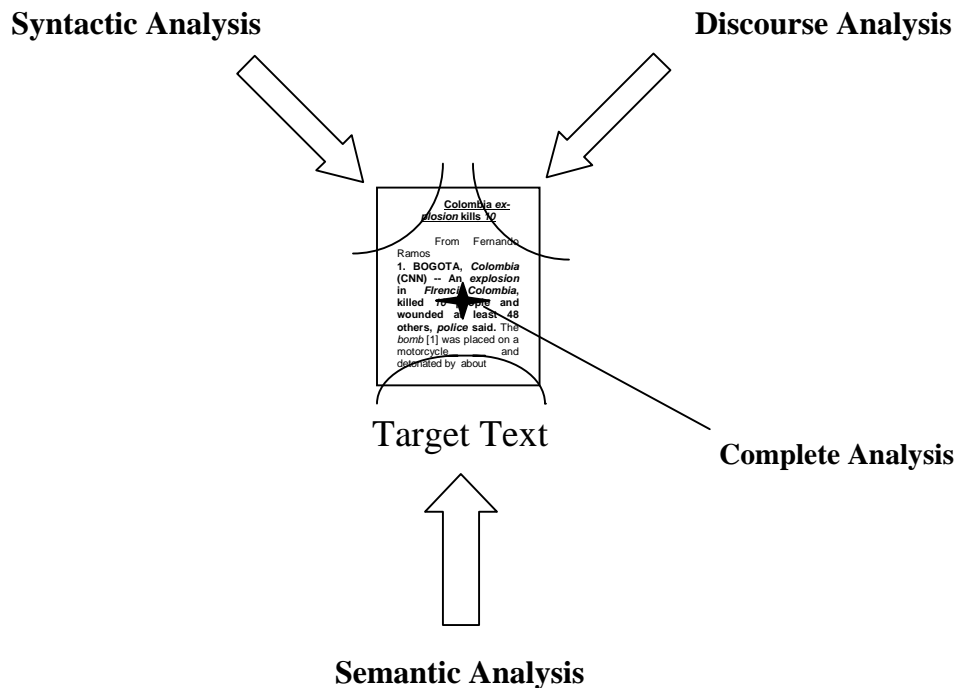


Figure 1: Integrated Tri-Strategy Approach to Coreference Resolution

2 Approach Strategy

The underpinning philosophy of our approach is that coreference resolution should be performed using shallow linguistic and semantic information. As a consequence, the system should be domain independent, which is necessary for a system which functions on web pages, whose domain varies widely. Shallow linguistic analysis also suits the fragmented linguistic style of web pages.

In the absence of detailed world knowledge and rich semantic/linguistic analysis of the target text, it is proposed in this research that an integrated, tri-directional approach to the analysis of web pages be adopted. Three types of analysis are proposed:

1. Shallow semantic analysis
2. Shallow linguistic or syntactic analysis
3. General discourse analysis

It is proposed that this approach can provide an approximation of detailed linguistic and semantic analysis. Figure 1 is a pictorial representation of how it is believed that our tri-directional knowledge poor strategy approximates to a knowledge rich approach. A complete

analysis of text would be represented by a complete shading in of the text-representative icon. Our tri-strategy approach approximates to this full analysis by analyzing from three different linguistic directions. The maximum horizon for textual analysis is therefore the central point in the text icon. The arcs represent the level of saturation our three-way, knowledge poor analysis achieves. Thus, a balanced approach towards the horizon of complete analysis (the central point) is achieved by the three separate analytical directions of our multi-strategy, knowledge poor approach.

Shallow linguistic analysis is based on the use of a part of speech tagger. It is unlikely that a full parse of the informal textual style of web pages will be possible with any degree of robustness. Shallow semantic analysis will be based on placing lexical terms into their semantic categories and then finding associations between them in WordNet. The discourse analysis proposed in this paper is classed as general because it is less concerned with the high granularity linguistic analysis for specific focus identification, normally associated with pronominal anaphora resolution. A more general, constraint-orientated discourse analysis is proposed, based on discourse segmentation, although some specific focus detection heuristics are used.

As part of our research, various source web texts have been annotated for coreference and analyzed for patterns and structures indicative of coreference. Heuristics in the above three categories have been drawn from these patterns, and are outlined in the rest of this paper. Mitkov (1999) identifies preference and constraint factors as the two main aspects of any coreference resolution system, and our heuristics fall into both of these categories.

3 Preference Factors

3.1 Semantic Comparison of Noun Phrases Using WordNet

WordNet is an electronic lexical database containing almost 80,000 noun word forms organized into some 60,000 lexicalized concepts (Miller 1998). These concepts are related according to two basic concepts: synonymy and hyponymy. Synonymy occurs as a relation between words when these words, although different lexemes, have the same meaning. Sparck Jones (1986) expresses these synonymous relations as ‘synsets’ which are ‘runs’ of lexemes that possess this synonymous relationship. Examples of a synset might be {shot, pellet} or {shot, injection}. The point that is illustrated by these two examples is that noun phrases can have different senses, in this case the noun phrase ‘shot’. Thus, a variety of synsets are possible for different senses of a word. These various senses of words, and their synsets, are contained within the WordNet database. It is proposed for our methodology that synsets taken from WordNet could be used to detect semantic similarity between nominals in a document and thereby indicate coreference relations.

The hyponymy relation present in WordNet can also be used to indicate coreference. An example of this relationship occurs between the noun phrases ‘bird’ and ‘robin’. The noun phrase ‘robin’ is a hyponym (subordinate) of the noun phrase ‘bird’, and ‘bird’ is a hypernym (superordinate) of ‘robin’ (Miller). Hyponymy relations are also used to express coreference in texts.

WordNet is a lexical database and not a knowledge base or ontology. It is also large and comprehensive. These two features should allow WordNet to be used in any domain for shallow semantic analysis of texts. This would therefore be an ideal information source for our web page coreference resolution methodology.

By putting lexical terms into their semantic classes and then comparing them using WordNet according to the synonymy and hyponymy relations, a shallow skim of the text can be conducted, without the need for linguistic analysis. It is proposed that this shallow approach to semantic analysis will be effective for web

pages whose domain conceptualization is relatively simple when compared, for example, with academic or technical texts.

3.2 Resolution of Named Entities

Rules are proposed for the resolution of named entities. The most obvious rule involves the detection of abbreviations of named entities. An example of an abbreviated named entity which occurs in our source web page texts is ‘Associated Press’ and ‘A.P.’ Variations on a name can also be resolved by searching for whole word substrings of the longest proper name string in the text; for example, ‘Bill Clinton’ and ‘Clinton’.

Named entity to named entity resolution is a relatively straightforward area of coreference resolution which requires the basic manipulation of strings and substrings. Although uninteresting from the point of view of our research, named entity resolution is an important part of any comprehensive coreference resolution system. A named entity resolution system is implemented in GATE using simple pattern-matching heuristics (D.Maynard, K. Bontcheva and H. Cunningham, 2003). This can be used for the task of named entity resolution for our system, given that it is intended that our system will be integrated into the GATE architecture.

3.3 Named Entity to Noun Phrase Coreference

A more interesting area than named entity to named entity resolution is the area of named entity to noun phrase coreference resolution. This type of coreference resolution can be performed by use of focusing, semantic and syntactic factors. An outline of proposed focusing and syntactic factors for coreference resolution is given below. The use of semantic information to associate named entities and nouns is more complex than in the case of noun phrase to noun phrase resolution. The problem is that it is harder to classify named entities than noun phrases semantically (where we can rely on WordNet). Named entity information might be useful for their semantic classification, such as a list of personal English names, or a list of all the countries in the world, but this information will not be comprehensive. A list of nouns likely to have proper names and also a list of nouns *unlikely* to have proper names might also be useful. For example certain nouns specifying entities which often have proper names such as ‘persons’, ‘hotels’ or ‘films’ are more likely to have proper names. Nouns designating abstract concepts such as ‘feeling’, ‘relationship’, ‘participation’ are unlikely to specify entities which have a proper name. Beyond the use of semantic information, reliance on syntactic and dis-

course based factors will be necessary for named entity to noun phrase resolution.

3.4 Syntactic Indicators of Coreference

Another key feature of this work is the detection of certain types of phrases, words and syntactic structures which specifically indicate coreference between noun phrases. Predicate nominals are an important phrasal type indicating coreference with a significant likelihood (Dimitrov 2002). A predicate nominal completes a reference to the subject of a clause and occurs after a copular verb such as ‘is’, ‘seems’, ‘looks’, ‘appears’ and so on. The ‘is a’ phrase is the principal type of predicate nominal, for example:

“George Bush is the President of the United States”

Here, ‘the President of the United States’ is a predicate nominal and corefers with the subject ‘George Bush’.

Syntactic indicators of coreference include apposition and parallelism. An example of apposition taken from one of our source texts is “the Monitor program, a show on the arts”, where ‘the Monitor program’ and ‘show’ are considered to be coreferent because of this syntactic structure. This is an example of how syntactic factors can be used to perform named entity to noun phrase resolution.

3.5 Focus Based Indicators of Coreference

Discourse theory seeks to formulate general principles of discourse structure and interpretation, and to integrate methods of anaphora resolution into a computational model of discourse interpretation (Lappin and Leass 1994). Researchers in this area include Grosz (1986), Grosz and Sidner (1986), Brennan et al. (1987) and Webber (1988). Discourse theory is highly influential on our approach for coreference resolution in web pages.

Discourse analysis deals with two major concepts of textual structure. These are ‘focus’ and ‘discourse segment’. The term ‘focus’ describes the way in which a part of text is ‘about’ a certain topic or entity, that it is considered to be ‘focusing upon’ a certain concept or topic. The fact that a part of text has a specific focus makes it a ‘discourse segment’. The ‘discourse segment’ is therefore defined as that portion of a text which has a specific focus.

Candace Sidner is one of the major researchers into focus based methodologies for anaphora resolution. Her

research is a major source for our approach to focus based coreference resolution. In her paper, ‘Focusing in the Comprehension of Definite Anaphora’ (1983), she outlines the main features of her focus based approach to coreference resolution. Any new term introduced in discourse is potentially a new focus, and candidates are either rejected or accepted according to various syntactic factors, such as pronominal confirmation (a candidate which has anaphoric pronouns is confirmed as the focus), and various types of constructions such as cleft and there-insertion sentence constructions, which are deemed to be reliable indicators of a new focus. Tense changes are also proposed as a way of detecting focus shifts. She also defines two distinct types of focus: actor focus and discourse focus, which correspond roughly to the main subject and theme of the sentences in a discourse segment.

Similar factors are proposed as being indicative of focus in our approach. Focus shift factors (see below) are the basic parameters within which focus identification occurs. Within discourse segments thus delimited, subjects and themes in consecutive sentences can be assumed to have a higher probability of coreference, corresponding to actor and discourse focus continuation. Pronominal anaphora will also be used as an indication of noun phrases or named entities that are in focus. It should be noted that discourse analysis is not as certain an indicator of coreference as syntactic or semantic information. This is why the emphasis of our use of discourse analysis is as a constraint factor, providing a soft window of text accessibility for the resolution algorithm.

Focus indicators can also be used to resolve title named entities to noun phrases in the succeeding paragraph or sentences. The title of a document is likely to be the main topic or focus of the document, and the first paragraph or sentences in the text are likely to be an introduction to the main topic of the text. It is likely therefore that there will be coreference between the title and the focus concepts of the first paragraph or sentences of the text. This identification of this discourse structure in web page texts is likely to be especially useful, since web pages often exhibit a high quantity of short paragraphs with titles or links associated with them.

4 Constraint Factors

4.1 Discourse Segmentation

Of greater importance in our approach than fine-grained discourse analysis (which is more appropriate

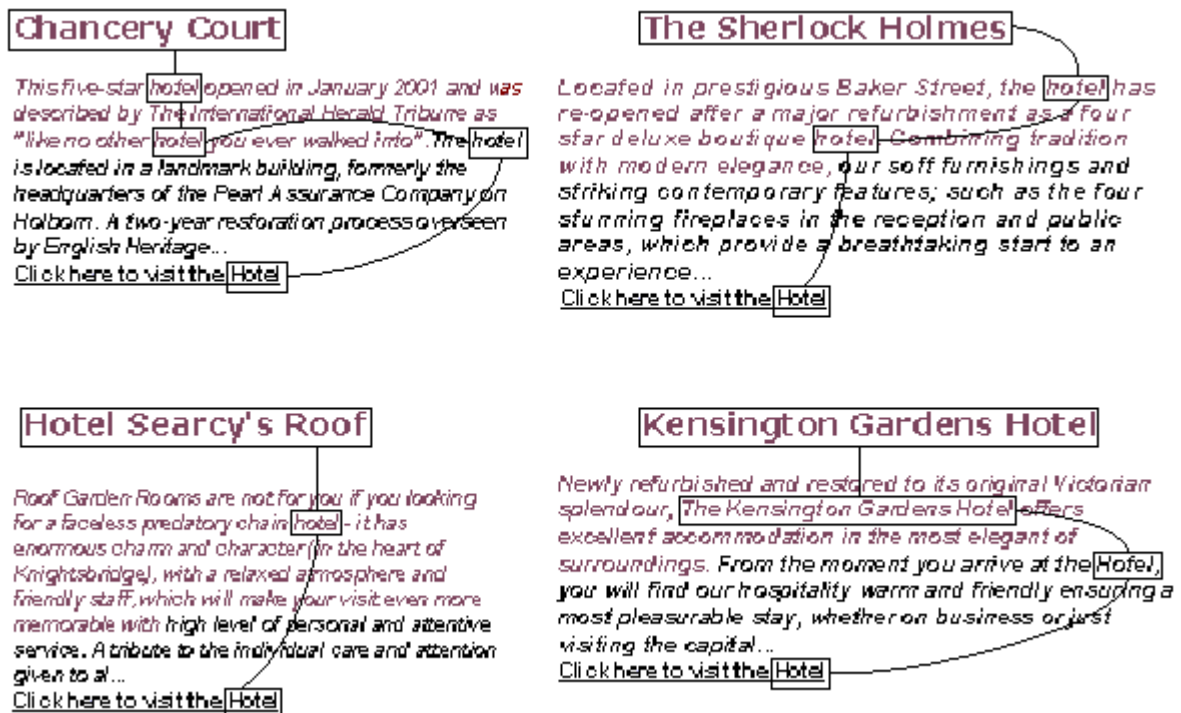


Figure 2: Disambiguating Noun Phrases Using Web Page Structure

for pronominal resolution) is a more general approach to discourse analysis based on discourse segmentation. Discourse segmentation acts as a constraint factor, by delimiting the region of text that is accessible for our coreference resolution algorithm.

Discourse segmentation is performed by the detection of focus shift indicators. These include indicators of new focus (certain phrasal structures, tense changes, introduction of new terms, patterns of pronominal anaphora), linguistic cue phrases (Reichman 1985) and semantic cohesion patterns (lexical chains) (Morris and Hirst 1991, Stairdmand 1996, Barzilay and Elhadad 1997, Hirst and St-Onge 1998).

Discourse segmentation can be used to eliminate false positives for nominal coreference resolution. In one of our source documents, the word 'film' is used in different discourse segments to refer to distinct movies. Semantic comparison will clearly associate all instances of the noun phrase 'film' in this textual extract. However, discourse segmentation indicates that the noun phrases occur in different parts of discourse, and are therefore likely to have distinct references. Thus we see that our general approach to discourse analysis acts not as a positive indicator of focus and corresponding coreference but as a constraint factor and delimiter for possible coreference patterns.

This approach can be seen as a variation of Grosz's methodology for semantic network partitioning according to focus factors (Grosz 1977). Our approach is a shallow version of Grosz's methodology because it is not a knowledge base which is being partitioned but rather the search space for shallow semantic comparison in the target text.

4.2 Discourse Segmentation Based on Web Page Structure

Web page structure constraint rules are an extension of discourse segmentation constraint rules. They are based on the assumption that textual structuring, which is common on web pages, corresponds to some degree with coherence patterns and discourse segments. The exploitation of web page structure as a discourse indicator is one of the key innovations of this research, and a central aspect of our approach to coreference resolution being 'tailor-made' for the web.

For example, one of our source documents contains many uses of the noun phrase 'hotel'. However, different sections of the page concern different hotels. We can use this textual structure as an indicator of discourse structure, and thereby delimit the resolution of the noun phrases 'hotel' throughout the document. Four separate

noun phrase coreference chains are therefore created instead of just one. This is shown in Figure 2.

Note also in this example the use of other preference and constraint factors for coreference resolution. Firstly, the four coreference chains extend to the title named entities of each section. This is an example of title named entity to focus noun phrase resolution heuristic previously described. This web page is a good example of the sort of subtitle and short paragraph textual structure that is common on web pages, a structure which our heuristic can exploit. It should also be noted that the second member of the coreference chain in the ‘Hotel Searcy’s Roof’ entitled section should not be part of the chain because it is in fact a false positive. Syntactic and phrasal constraint factors might be able to eliminate this false positive by detecting the presence of the word ‘not’. Syntactic and phrasal constraint factors are described in the next section.

An algorithm for performing discourse segmentation based on web page structure is therefore proposed. One of the innovations of our approach proposal is to take Reichman’s (1985) use of ‘cue phrases’ for locating focus shifts in task-orientated and conversational dialogues, and adapt this methodology to web page texts. Web pages are semi-structured and it is proposed that this structure has significant correspondence to discourse structure. It follows that web page structure can be exploited as an indicator of discourse segmentation. HTML markers, which are often used to structure web pages, can be interpreted in the same manner as cue phrases, as being indicative of focus shifts and segment boundaries. A more specific discourse theory for web page structure and markers is likely to be developed in the course of our research.

4.3 Syntactic Constraint Factors

As well as being positive indicators of coreference, certain types of words, phrases and syntactic structures can indicate that nominals are *not* coreferent, thus eliminating false positives which might otherwise be indicated as coreferent by preference factors. Words that are factors in indicating coreference incompatibility include ‘but’, which indicates disjunction between two clauses and therefore the increased likelihood of logical distinction between their constituent noun phrases, and the word ‘not’ (see above example). It is also proposed that prepositional attachment between two potential coreferents makes them logically incompatible. Prepositional attachments acting as modifiers to nouns can also be used as a distinguishing factor. For example, in one of our source web pages, two instances of the noun phrase ‘war’ can be disambiguated by way of the semantic incompatibility of their prepositional attachments, ‘in Iraq’ and ‘against terrorism’. Given that the lexemes ‘Iraq’ and ‘terrorism’ are semantically incom-

patible, the modified noun phrase can be disambiguated for coreference.

The use of syntactic structures as filters for nominal and named entity coreference resolution is based on similar use of syntactic filters for pronominal resolution in the work of Lappin and Leass (1994), and Boguraev et al. (1996). The innovation in our approach is to adapt syntactic filters for pronominal resolution to the nominal and proper name resolution task.

5 Criticisms of the Proposed Approach

Various criticisms can be raised concerning the proposed approach for web page coreference resolution. One of the principal ideas of this research is that an approach which is ‘tailor-made’ for the web can be developed. Such an approach will be knowledge-poor and involve a shallow ‘skim’ of the text, so that we do not have to perform any detailed linguistic or semantic analysis. Such detailed analysis would be impossible for the fragmented linguistic style of the web and its domain ubiquity. The other central principle of a web-approach to coreference resolution is the proposition that web page structure corresponds to discourse structure.

The question may be asked as to whether the web is a specific textual genre or rather an amalgamation of many different texts with their own individual styles. To an extent the latter is true, because the web contains many documents taken from other textual genres such as newspaper texts and academic journals. However, it is proposed that a significant number of web pages which are not merely transcriptions from other non-web, textual sources do exhibit a specific ‘web’ style. This is typically a style involving many subtitles (which are often links), short paragraphs and the interleaving of the text with graphical icons and pictures. Further research is being undertaken to attempt some sort of statistical clarification of how predominant and unique this ‘web’ page style is.

Another criticism might be that web page structure, if it does exist, does not correspond to discourse structure. From our analysis of source texts, there is strong evidence that the short paragraphs, links and titles of web page texts do correspond closely to topicality and therefore are representative of discourse structure to a significant extent. This is not to say that web page structure is the sole indicator of discourse structure, and our approach also facilitates the detection of traditionally specified linguistic indicators of discourse structure (cue phrases, tense changes etc). The combination of web page structure and traditional indicators of discourse segments constitutes our overall approach to discourse analysis for coreference resolution on web pages.

Finally, it may be said that web page structure varies widely from page to page and would therefore be hard

to exploit computationally because of this variety. However, different types of web pages exhibit different types of structure with some regularity. Home pages typically exhibit a structure involving multiple links and short paragraphs. Semi-structured timetable or weather report pages possess a structure which is closer to that of a table. For this reason it will be necessary for our system to identify certain types of web pages, and to classify a target web page for our algorithm according to this type. Our algorithm can then employ a different variation of discourse analysis for the target page according to the web page type it has been classified as. So for example, free text dominated web pages would require the use of little or no web page structure based discourse analysis, whereas tabular, semi-structured web pages would require discourse analysis that adheres to this tabular semantic structure. The multiplicity of possible web page structure types can therefore be arranged into a coherent order, from which basis the discourse analysis stage of the resolution algorithm can proceed.

6 Conclusion and Further Research

A knowledge-poor, tri-directional approach for coreference resolution of named entities and noun phrases has been outlined. The main innovations of this approach are as follows:

1. The use of HTML markers and web page structure as indicators of discourse structure.
2. An approach to coreference resolution which performs shallow analysis from three primary analytical angles: syntactic analysis, semantic analysis and discourse analysis. It is proposed that this tri-directional, knowledge-poor approach can approximate to a knowledge-rich approach.
3. An approach designed to cope with web pages due to its knowledge-poverty and exploitation of web page structure.
4. The use of syntactic and discourse based strategies traditionally associated with pronominal resolution for named entity and noun phrase resolution.

The rules and heuristics outlined above express various types of linguistic features which are commonly indicative of coreference or can act as guidelines for these patterns. The main task of our remaining research is to translate these linguistic rules into computational rules which use the input of WordNet, a part of speech tagger and web page structure. Implementation and evaluation can then follow.

At this stage, evaluation has not yet been performed except through our reconnaissance analysis of a collection of web pages. Thorough evaluation shall proceed once implementation of the system has been completed and a test corpus has been created. In view of the target

text being web pages, there may be some difficulty in obtaining such a test corpus for evaluation. It may be the case that we will have to manually annotate this test corpus ourselves. We can then compare our system annotations against this standard. Alternatively, we could use a standard test corpus for coreference, such as DARPA data and MUC annotated corpora. In order to assess the performance of our system on web pages, we would then have to determine the distinguishing coreference pattern features of web pages compared to these standard corpora, and then evaluate the performance of our system on these distinguishing features. We can also test our system directly on standard corpora. A combined evaluation would therefore be produced, which tests our system on both general textual coreference features, and also on coreference patterns which are unique to the web.

It should be noted that the above rules and heuristics are in their preliminary stages. Our account of them constitutes a general, preliminary outline of the approach to coreference resolution which will be developed in our research. Further research will endeavor to clarify, evaluate and further develop these rules, and also convert them into computational rules. It may also be the case that we will have to develop some sort of detailed discourse theory specifically for web pages.

Finally, named entity and nominal coreference resolution might be useful for performing cross document coreference resolution. Cross document coreference resolution requires information about the identity of candidates in order to decide whether they are coreferent or not. Nominal and named entity coreference chains would be highly informative as to the identity of candidates. It is therefore possible to extend our system for intra-document nominal and named entity coreference resolution to cross-document coreference resolution. A system that performs both inter- and intra-document coreference resolution would be highly suited to the web, which involves multiple interlinked documents and where, as a result, coreference occurs as both an inter- and intra-document phenomenon.

References

- Barzilay R, Elhadad M, 'Using Lexical Chains for Text Summarization', ACL, 1997
- Boguraev B, Kennedy C, 'Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser', Proceedings of the 16th International Conference on Computational Linguistics, 1996
- Brennan S, Friedman M, Pollard C, 'A Centering Approach to Pronouns', Proceedings of the 25th Annual

Meeting of the Association for Computational Linguistics, 1987

Cardie C, Wagstaff K, 'Noun Phrase Coreference as Clustering', Proceedings of the Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora, 1999

Ciravegna F, Dingli A, Guthrie D, Wilks Y, 'Mining Web Sites Using Unsupervised Adaptive Information Extraction', Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary, April 2003

Cunningham H, Maynard D, Bontcheva K, Tablan V, 'GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications', Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, 2002

M. Dimitrov, K. Bontcheva, H. Cunningham, D. Maynard, 'A Light-weight Approach to Coreference Resolution for Named Entities in Text', Proceedings of the Fourth Discourse Anaphora and Anaphor Resolution Colloquium (DAARC), Lisbon, September 2002

Grosz JB, 'The Representation and Use of Focus in a System for Understanding Dialogs', SRI Internationals, California, 1977

Grosz JB, 'Focusing and Description in Natural Language Dialogues' in 'Elements of Discourse Understanding', Cambridge University Press, 1981

Grosz JB, Joshi A, Weinstein S, 'Towards a Computational Theory of Discourse Interpretation', Harvard University and University of Pennsylvania, unpublished, 1986

Grosz B, Sidner C. 'Attention, Intentions and the Structure of Discourse', Computational Linguistics, 12(3), 175-204, 1986

Grosz BJ, Joshi AK, Weinstein S, 'Centering: A Framework for Modelling the Local Coherence of Discourse', ACL, 1995

Hirst G, St-Onge D, 'Lexical Chains as Representation of Context for the Detection and Correction of Malapropisms', Wordnet: An Electronic Lexical Database, ed. Fellbaum, MIT Press, 1998

Kameyama M, 'Recognising Referential Links: an Information Extraction Perspective', Proceedings of the ACL'97/EAL'97 Workshop on Operational Factors in

Practical, Robust Anaphora Resolution, Madrid, Spain, 1997

Lappin S, Leass H, 'An Algorithm for Pronominal Anaphora Resolution', Computational Linguistics 20(4), 1994

Maynard D, Bontcheva K and Cunningham H. 'Towards a semantic extraction of named entities', Recent Advances in Natural Language Processing, Bulgaria, 2003

Miller GA, 'Nouns in WordNet', ch1. 'WordNet, an Electronic Lexical Database', ed. Christiane Felbaum, MIT Press, 1998

Mitkov R, 'Anaphora Resolution: the State of the Art', University of Wolverhampton, 1999

Morris, J. Hirst, G, 'Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text', Computational Linguistics, March 1991, Vol. 17, No 1, 21-48.

Reichman R, 'Getting Computers to Talk Like You and Me', MIT Press, 1985

Sidner CL, 'Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse', Technical Report No. 537 MIT, Artificial Intelligence Laboratory, 1979

Sidner CL, 'Focusing in the Comprehension of Definite Anaphora', in 'Readings in Natural Language Processing', 1983

Soderland S, 'Learning to Extract Text-Based Information from the World Wide Web', Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, 1997

Sparck Jones K, 'Synonymy and Semantic Classification', Edinburgh: Edinburgh University Press, 1986

Stairmand MA, 'A Computational Analysis of Lexical Cohesion with Applications in Information Retrieval', PhD thesis, Center for Computational Linguistics, UMIST, Manchester, 1996

Webber BL, 'Discourse Deixis: Reference to Discourse Segments', Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics, 113-121, 1988