Integrating NLP Tools to Support Information Access to News Archives

Horacio Saggion*, Emma Barker*, Robert Gaizauskas*

Jonathan Foster**

*Department of Computer Science - University of Sheffield
211 Portobello Street - Sheffield - S1 4DP - UK
{saggion,ejbarker,robertg}@dcs.shef.ac.uk

**Department of Journalism Studies - University of Sheffield
18-22 Regent Street - Sheffield - S1 3NJ - UK
j.foster@shef.ac.uk

Abstract

We describe Cubreporter, a project which investigates the use of advanced natural language processing techniques to enhance access to a news archive for the specific purpose of background writing. We describe the problem of background writing for a breaking news story and the requirement for advanced NLP tools. We focus on the description of the overall functionalities of our prototype and give an account of our methodology for evaluation.

1 Introduction

Cubreporter is a research project which investigates how language technologies might help journalists to access information in a news archive in the context of a background writing task. The function of background material is to support and contextualise a breaking news story. The specific characteristics of the background-writing scenario make recent advances in areas of natural language processing such as question answering and text summarization relevant to this task.

The main research questions we address in this project are: (i) what are the essential components of a background story and how does background information relate to the "foreground" breaking news story? (ii) how can background information for a breaking news story be accurately found in the archive given the initial breaking news story? (iii) how can human language technology assist a journalist to access the vast amount of information in a news archive? in particular can recent advances in NLP technologies, in areas such as question answering, summarisation, and information extraction, offer advantages in gathering background that standard information retrieval cannot? (iv) how is background writing quality affected by the use of human language technology?

To address these questions we have designed and implemented a prototype that incorporates a standard information retrieval engine as a baseline, as well as a question answering system and document summarization technology. Information extraction technology is also used to extract structured representations of events which are in turn used to populate a database to support similar event search. These information access technologies are embedded in a browser-based graphical user interface which allows users to combine them flexibly in an iterative information seeking process.

Here we give an overview of the project and describe our work on the background gathering task and the tools used to support it. The main contributions of the work are: (i) a descriptive theory characterising the nature of background in the news and its relation to the foreground news story; (ii) a design for an information access platform that integrates information retrieval, summarisation, question answering and information extraction capabilities within a single system operating over a text archive of significant size; (iii) a methodology for comparative evaluation of different combinations of language technologies for the task of background writing, allowing an assessment of the relative utility of more sophisticated natural language processing tools versus traditional information retrieval tools for the task of background writing.

The rest of the paper is organised as follows. In the following section we describe the task of writing background news. In Section 3, we describe the structure of the news archive. Section 4 gives an overview of the different NLP processes involved in the project. In Section 5, we describe our methodology for extrinsic evaluation. Section 6 closes with an account of work in progress and future developments. It should be noted that while we focus on the specific task of journalistic background writing, investigative intelligence gathering in response to a new event is by no means exclusive to the news-producing community and work described here is also relevant to information seeking professionals working in commercial, policing, military and scientific domains.

2 Writing Backgrounds

Our work to date has involved the study of journalists who either work for or with materials produced by the Press Association, the major UK domestic newswire

service which provides copy to all major national daily newspapers. While background figures in a number of ways, including simple descriptive phrases interjected into the current story (e.g. *former Chancellor of the Exchequer*) and fact sheets listing similar or relevant occurrences (e.g. a listing of previous train crashes), we shall focus on the most significant form of background material only, the so-called "backgrounder". Backgrounders are coherent documents, typically written when a news editor deems a particular story worthy of dedicated background material, but which can be read on their own, out of their production context. They are usually not released till sometime after a news story has broken as time is needed both to determine whether a story merits a backgrounder, but also for the research to be carried out to assemble the material. Their function is not to continue to report details of new events, but rather to provide text that supports and contextualises these events.

There has been no prior work, so far as we are aware, on gathering information for background writing. Attfield & Dowell (03) propose a general model of journalistic information gathering. However, the backgrounder task is different from other types of news writting and deserves special attention.

Interviews with journalists, observation during a controlled task and text analysis of a sizeable set of archived background stories show that backgrounds are composed of four types of material: (1) accounts of similar events in the past (e.g. other train crashes, scandals of similar nature, etc.); (2) accounts of events which have led up to the current event (e.g. a chronology of company takeovers, store openings, price cuts and profit warnings in the months leading up to a supermarket's announcement of low annual profits); (3) profiles of persons or organisations or locations (usually role players in the new event) comprising some highly structured factual information about the role player, for example date and place of birth, career appointments, spouse etc; accounts of the role player in events leading up to the event and accounts of the role player in similar events to the current event; and (4) comment (quotes) on any of the preceding by notable individuals.

Interestingly, these information gathering requirements are similar to those addressed in recent NLP challenge tasks. For example, finding profiles of people or organisations is a task dealt with in recent TREC Question Answering evaluations (Voorhees 04) and Document Understanding Conferences (Over & Yen 04) and can be supported by solutions proposed in these contexts. Finding events similar to one reported in breaking news can be implemented with information extraction technology: text in the archive could be mapped off-line into structured representations which could be stored in a database for on-line searching (Milward & Thomas 00). Question answering technology can be used to support fact gathering as well as fact checking in a background writing context. Consider as an illustration the news about the "kidnapping of UK-born Margaret Hassan". Of considerable importance for the UK public are answers to the following (among other) questions: *How many British citizens are living in Iraq?* and *Where was Margaret Hassan born?*. Techniques used on factoid question answering are relevant here.

3 The News Archive

Through our collaboration with the PA we have obtained access to 11 years of newswire copy from 1994 to 2004. The archive contains more than 8.5 million stories totalling 20GB of data. The raw corpus has been processed and encoded in XML following a strict Document Type Definition (DTD) specification which captures all meta-data delivered by the Press Association and which includes elements such as story date, category, topic, and structural information such as headlines, bylines, and paragraphs. One example story is shown in Figure 1. The archive is organised per dates following the logical organisation of the PA wire where years are composed of months, months are composed of days, and there are a number of stories per day. Stories in the PA archive are classified into a number of topics or news categories from a controlled vocabulary representing the subject matter of the story (e.g., *Courts*, *Politics*). Within the same topic, stories are further identified by a number of free-text keywords that the journalists would assign which are called *catch-lines*.

When a "news event" occurs, a reporter writes a *snap*, a line of text summarising the news and "moves" it to the wire. From that point on, stories follow an installment pattern where each installment carries an updated account of the story. Installments have names such as *snapfull*, one or two paragraph long text expanding the snap, *lead*, copy that summarises the major aspects of the story, and so on. These installment types reflect their position and significance in the publishing cycle of major newspapers.

Subscribers to the PA have access to the archive through the PA Digital Text Library and Mediapoint systems which have a number of functionalities for information access including: text, keyword, and topic search. Output to a specific query is presented as a ranked list of documents and associated 'lead' paragraphs. Access to the full document is done by following a link.

4 Advanced NLP Technology

One of the objectives of the project is to carry out experimentation in order to investigate the research questions identified in the introduction. In the future it might be possible to deploy, at least partially, the technology produced in this project in a real application to give journalists cutting-edge natural language

```
<?xml version="1.0"  encoding="ISO-8859-1"?>
<!DOCTYPE HSA SYSTEM "../../../../dtds/HSA.DTD">
<HSA  DATE="20042004"  DAY="20"  YEAR="2004"  MONTH="04"
ID="HSA7041"  PRIORITY="4" CATEGORY="NRG" COUNT="76"
MSGINFO="PA"  TOPIC="1 ROYAL Cockle Morecambe"
TIMEDATE="201407 APR 04">
<HEADLINE>COCKLE PICKERS RESCUED FROM NOTORIOUS SANDS
</HEADLINE>
<BODY>
<PARAGRAPH NRO="1"> Four cockle-pickers have been rescued
by lifeboatmen  after getting trapped on
the sands at Morecambe Bay.</PARAGRAPH>
<PARAGRAPH NRO="2"> A group of ten cocklers,
who were not Chinese,
were returning to Hest Bank on a tractor
which got stuck as the tide swept in.</PARAGRAPH>
<PARAGRAPH NRO="3"> Some of the group were washed off the
tractor  but managed to get to a rocky
outcrop called Priest Skier. The rest were rescued by
the RNLI Morecambe hovercraft.</PARAGRAPH>
</BODY>
</HSA>
```

Figure 1: Corpus Encoding

processing capabilities for information access. Cubreporter comprises an off-line *corpus processing* subsystem and an on-line *information access* subsystem (see Figure 2).

The off-line subsystem produces a *text index* for document retrieval, *generic summaries* at fixed length for each story, *generic multi-document summaries* for sets of known related stories, and *logical forms* for database population. The database is an entity-event-relation relational repository which stores the information resulting from a process of semantic interpretation of each story. The database contains tables to record references to entities (such as people and organisations), events, locations, and temporal information. Relations are a set of fixed logic relations including logical subject and object, apposition, qualification, etc. A table of attributes stores the different values that qualify entities and events such as adjectives, adverbials, and quantifiers.

The on-line system provides question answering, keyword search, similar-event, and further ad hoc summarization capabilities.

4.1 Off-line processing

The whole archive is processed with tools adapted from the GATE Java library (Cunningham *et al.* 02). We perform tokenisation, sentence boundary identification, part-of-speech tagging, morphological analysis, and named entity recognition, keeping the results of the analysis for use by various language processing components. A text index is produced for the processed documents using Lucene[1], a Java-based open source tool for indexing and searching. The text of each story at textual and paragraph level as well as each metadata field are indexed. Search can be performed in any of the fields alone or in combination with boolean operators. Further linguistic processing of the archive is carried out with SUPPLE (Gaizauskas *et al.* 05), a freely-available parser, integrated in GATE, and with an in-house discourse interpreter. SUPPLE uses a feature-based context-free grammar in order to produce syntactic representations and logical forms. The grammar in use consists of a sequence of subgrammars for: noun phrases (NP), verb phrases (VP), prepositional phrases (PP), relative clauses (R) and sentences (S). The semantic rules produce unary predicates for entities and events and binary predicates for attributes and relations. Predicate names are: (i) the citation forms obtained during lemmatisation; (ii) forms used to code syntactic information (e.g. *lsubj* for the logical subject of a given verb); (iii) specific predicates are used to encode, for example, named entity information (e.g. *name* for the name of a person). The document semantics is further analysed by a discourse interpreter which maps entities into a discourse model and performs coreference resolution based on an ontology we are adapting for the purpose of this project. The results of this semantic discourse analysis is transformed into records that are used to populate the database. For example for the headline presented in Figure 1 an event of type `rescue` and two entities `cockle pickers` and `sand` would be created. A `patient` relation would be created between the event and the entity `cockle pickers` and a `from` relation would be created between the event and entity `sand`. We are currently looking at standard classification systems such as the Subject Code three level system for describing content produced by the International Press Telecommunications Council (`http://www.iptc.org`). This system is used to describe news content and seems appropriate for the creation of a Cubreporter ontology for news events and actors.

Summaries at fixed compression rate and ranked sentences (for on-line summary access) are computed for each story in the archive using an in-house sin-

[1] http://jakarta.apache.org/lucene

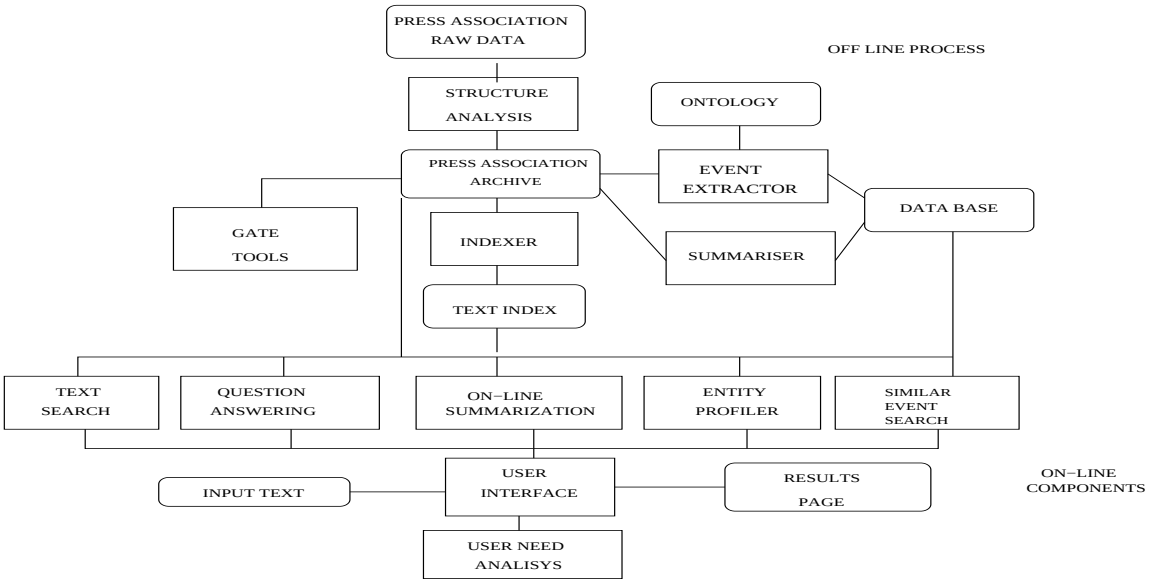

Figure 2: System Components

gle document summariser (Saggion 02). Sentences are ranked based on sentence-summary worthiness score obtained by combining scores for various features including sentence position, similarity of the sentence to the document headline, term distribution, named entity distribution, etc. Individual scores are combined using weights experimentally obtained from training corpus.

Off-line multi-document summarisation is carried out on a set of story-related documents. The tool extends the single document summariser by implementing a centroid-based summarisation system (Saggion & Gaizauskas 04a) which computes the similarity of each sentence to a cluster centroid and combines this value with single document summarization features. An n-gram similarity metric has been implemented to filter out redundant information, using a similarity threshold adjusted over training data. The weights used to combine the different features are trained over corpora.

4.2 On-line Processing

Access to the archive is through a user interface which is designed with the input text as its focus. The user enters a text which can be a sequence of keywords, a well formed natural language question, or a short snap-like text such as the initial report of a breaking news story. The system first carries out full text analysis of the fragment, and depending on the result of the analysis, additional options are made available including:

- access to full documents and summaries;
- answers and contexts to specific questions;
- profiles of persons, organisations, and locations;
- events similar to those described in the input.

Access to full documents and summaries In a pure document search situation – when the input text is a list of keywords – the journalist is presented with a results page containing access to full documents and to the previously computed story summaries, and installment multi-document summaries. The documents are ranked either by date or relevance – for the latter the standard `tf * idf` Lucene's default scoring mechanism is used. In addition query-focused summaries, tailored to the user's input text, are computed dynamically, in such a way that extracted sentences will be related to the user's assumed information need. Such summaries can be very effective when trying to identify the relevance of a document with respect to a query (Tombros *et al.* 98); generic sentence-summary worthiness features are combined with a query-based feature in a scoring function to obtain such summaries. A set of user-selected documents can be multi-document summarised on-line.

Question Answering Question Answering (QA) functionalities are used to provide the journalist with short, text units that answer their specific, well-formed natural language questions. We make use of a logic-based question answering system which given the logical form produced by the text analysis module, scores answer candidates in the database based on syntactic and semantic criteria (Gaizauskas *et al.* 03). Briefly, the scoring mechanism operates as follows. When the text analysis module finds a question, it produces an analysis which includes the expected answer type (EAT) and depending on the question, a special attribute created to refer to the attribute-value to be extracted from the answer entity. Each candidate answer gets a preliminary score according to (1) its semantic proximity to the EAT using WordNet and (2) the number of relations the candidate answer has with

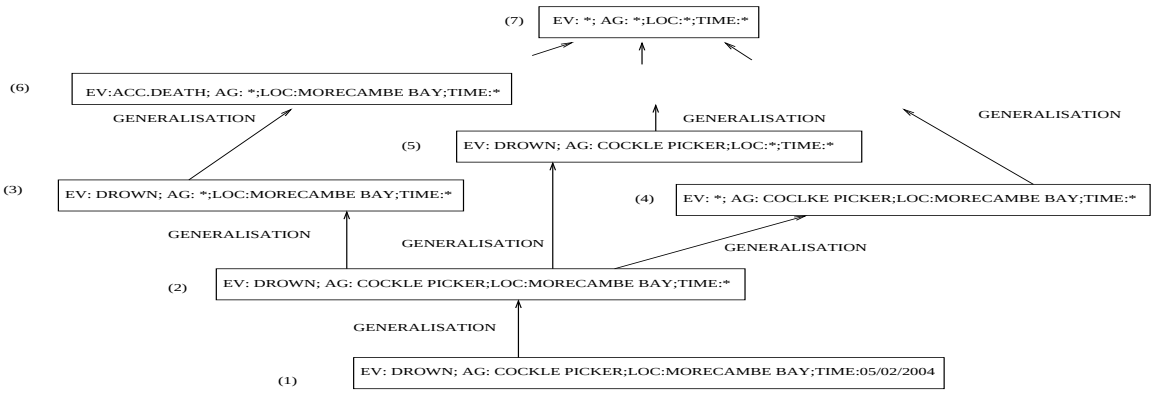

Figure 3: Similar : Generalisation

elements of the question. An overall score is computed for each entity as a function of its preliminary score plus a similarity value between the question and the sentence where the entity comes from (e.g., similar to word overlap).

Entity Profiles If the analysis of the input text recognises entities such as persons or organizations the system will return a profile. Tools for creating entity profiles are adapted from our definitional question answering system, developed for the TREC QA track, which uses pattern matching techniques against definition patterns to identify text fragments conveying profile information (Saggion & Gaizauskas 04b). Passages extracted from the collection are then filtered with the assistance of a similarity metric to avoid repetition of information.

Similar Event Search Our research into background news writing has shown that users are likely to be interested in past events similar to the new event that is the focus of the breaking news story. One strategy for extracting similar events is to use the IR component with the snap as a query in the hope that stories describing similar events will be returned at high ranks. However, this may be problematic as by definition the breaking story, to be "news", must be new and hence different in significant respects from previous events. Here, we propose a novel approach based on searching the database of extracted semantic representations of texts. Given a snap-like input text, a structured representation of the input is produced which includes a list of event-like representations which is then used to query the database.

Consider for example the following snap:

> *Eighteen Chinese Cockle pickers drowned in Morecambe Bay last night.*

The analysis of this text fragment produces the following template-like representation:

> Event: drown
> Agent: cockle picker
> Location: Morecambe Bay
> Time: 05/02/2004

In order to obtain similar events, this initial representation is transformed into a series of successively more general queries (see Figure 3). One possibility consists in replacing the time of the event by a wildcard: this will result in retrieving from the database "previous drownings in Morecambe Bay involving cockle pickers" (representation (2) in Figure 3). A further refinement would replace the agent of the event by a wildcard, resulting in a statement like "previous drownings in Morecambe Bay" (representation (3)). Yet, another possibility would be to replace the location of the event by a wildcard producing a statement like "previous drownings involving cockle pickers" (representation (5)). Yet another possibility would be to replace the actual arguments by generalisations: for example the type of event (drowning) could be replaced by other accidental deaths (using the information provided by the ontology as in (6) in Figure 3). Generalisations can be applied until all arguments have been replaced so as to effectively obtain "all events in the database."

The output of this process is a list of sentences from which each matching event was derived. For example, representation (3) in Figure 3 might return the following sentence:

> *Today a man drowned while he was walking his dog in Morecambe Bay ...*

We are currently investigating methods for presenting the results to the user, e.g., ranking and clustering.

4.3 Prototype Implementation

The off-line processing results in a Lucene inverted index, summaries and structured semantic representations of each story in the archive. The summaries and semantic representations are held in relational tables in a mySQL database. The user interacts with the system through a web client which communicates with a

web server (Tomcat). Both the text index and the relational database are accessed by the server as needed during on-line processing and web content is dynamically created for return to the client. A user database records details of users for security purposes and to allow search histories to recorded and revisited in subsequent sessions. Session management in the server allows multiple concurrent access to the archive.

5 Evaluation

A key criterion in evaluating our project is that of quality of the background stories which can be created by using the prototype. In order to measure in a scientific experiment whether new information technologies offer better access to information than conventional text search engines for the purpose of background information gathering, one has to articulate a theory of what constitutes a good background story. In order to address this issue we are following two complementary directions. First, we are investigating whether independent assessors can consistently rank and categorise backgrounds according to their quality. Secondly, we are working on a descriptive theory of background based on the semantic relation of content units in the background in relation to the breaking news event. This theory will help to predict the quality of a background as a function of its content and form. A pilot study has been conducted to investigate the first issue. The data for this study consisted of a collection of student assignments that were evaluated in terms of their respective quality by three independent journalist evaluators. Preliminary results reported elsewere (Barker & Gaizauskas 05) indicate reasonably high agreement among evaluators. Given the positive results of this experience, our plan is to construct a broader and more controlled corpus which will include different types of background written by professional journalists. In order to develop a theory of background, a set of relations is needed which indicate not only the relation between background and breaking news event, but also the relations between the different content units of the background.

We propose to adopt a framework such as that of Wolf and Gibson (Wolf & Gibson 04) or Marcu (Marcu 00) who have shown that it is possible to specify a set of discourse relations for text segments that are easy to code. Given such a descriptive framework a corpus of backgrounds will be annotated and experiments will be carried out to test the quality of background with respect to the descriptive theory.

While no extrinsic evaluation of the overall Cubreporter system has yet been carried out, some of the components have been evaluated. For example:

- the generic multi-document text summariser had a very good performance in DUC 2004, it was the second best system in task 2;
- the profile-based multi-document text summariser performed reasonably well in task 5 of DUC 2004 coming among the top nine participants. We have recently implemented a new method for extracting biographical information from text and obtained improved performance (Saggion & Gaizauskas 05);
- the QA system has participated in TREC/QA and in particular the definitional component placed fourth in 2004;
- in spite of the fact that our parser has never been formally evaluated, it has contributed to many successful information extraction projects in the past. We are currently assessing two approaches to evaluation: one is the evaluation of the logical forms produced by the parser using a resource such as Suzanne (Sampson 95), the other is to develop test suites for testing a range of grammatical phenomena and to support regression testing during grammar development.

While advanced NLP tools are far from perfect, they have the potential of offering improved access to news archives as compared with existing information access technologies, an hypothesis we are trying to validate.

6 Conclusion and Future Work

From a theoretical point of view, this work contributes with an in-depth examination of the background gathering task and with a methodological framework for extrinsic evaluation of information access systems.

From a technical point of view our work to date contributes to the creation, adaptation, and integration of NLP technology to support the task of background gathering. Much work has been done on specification and design of a web-based user interface and on in-house intrinsic evaluation of the different NLP components.

Current work involves the full integration of the NLP modules to carry out evaluation and testing of our research hypotheses.

Acknowledgements

We would like to thank four anonymous reviewers for their comments and suggestions. We acknowledge the support of the UK Engineering and Physical Sciences Research Council, research grant: R91465.

References

(Attfield & Dowell 03) S. Attfield and J. Dowell. Information seeking and use by newspaper journalists. *Journal of Documentation*, 59(2):187–204, 2003.

(Barker & Gaizauskas 05) E. J. Barker and R. Gaizauskas. Evaluating Cub Reporter: proposals for extrinsic evaluation of journalists using language technologies to access a news archive in background

research. In *Proceedings of the COLIS 2005 Workshop on Evaluating User Studies in Information Access*, 2005. To appear.

(Cunningham *et al.* 02) H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

(Gaizauskas *et al.* 03) Robert Gaizauskas, Mark A. Greenwood, Mark Hepple, Ian Roberts, Horacio Saggion, and Matthew Sargaison. The University of Sheffield's TREC 2003 Q&A Experiments. In *Proceedings of the 12th Text REtrieval Conference*, 2003.

(Gaizauskas *et al.* 05) R. Gaizauskas, M. Hepple, H. Saggion, and M. Greenwood. SUPPLE: A Practical Parser for Natural Language Engineering Applications. In *International Workshop on Parsing Technologies*, 2005. Accepted.

(Marcu 00) D. Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, Mass, 2000.

(Milward & Thomas 00) D. Milward and J. Thomas. From information retrieval to information extraction. In *Proceedings of the ACL Workshop on Recent Advances in Natural Language Processing and Information Retrieval*, 2000. Available at: http://www.cam.sri.com/html/highlight.html.

(Over & Yen 04) P. Over and J. Yen. Introduction to DUC-2004: An intrinsic evaluation of generic news text summarization systems. In *Proceedings of the HLT/NAACL 2004 Document Understanding Workshop (DUC-2004)*, 2004. Available at: http://www-nlpir.nist.gov/ projects/duc/pubs/ 2004slides/duc2004.intro.pdf.

(Saggion & Gaizauskas 04a) H. Saggion and R. Gaizauskas. Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of Document Understanding Conference*, Boston, MA, May 6-7 2004. NIST.

(Saggion & Gaizauskas 04b) Horacio Saggion and Robert Gaizauskas. Mining on-line sources for definition knowledge. In *Proceedings of FLAIRS 2004*, Florida, USA, 2004. AAAI.

(Saggion & Gaizauskas 05) Horacio Saggion and Robert Gaizauskas. Experiments on Statistical and Pattern-based Biographical Summarization. In *Proceedings of the 12th Portuguese Conference on Artificial Intelligence - TeMA Workshop*, 2005. Accepted.

(Saggion 02) Horacio Saggion. Shallow-based Robust Summarization. In *Automatic Summarization: Solutions and Perspectives*, ATALA, December, 14 2002.

(Sampson 95) G. Sampson. *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Clarendon Press, Oxford, 1995.

(Tombros *et al.* 98) A. Tombros, M. Sanderson, and P. Gray. Advantages of Query Biased Summaries in Information retrieval. In *Intelligent Text Summarization. Papers from the 1998 AAAI Spring Symposium. Technical Report SS-98-06*, pages 34–43, Standford (CA), USA, March 23-25 1998. The AAAI Press.

(Voorhees 04) E. Voorhees. Overview of TREC 2003. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*, NIST Special Publication 500-255, 2004. Available at: http://trec.nist.gov/ pubs/trec12/papers/ OVERVIEW.12.pdf.

(Wolf & Gibson 04) F. Wolf and E. Gibson. A response to Marcu (2003). Discourse structure: trees or graphs?, 2004. Available at: http://web.mit.edu/fwolf/www/discourse-annotation/Wolf_Gibson-coherence-representation.pdf.