FASIL Email Summarisation System

Angelo Dalli, Yunqing Xia, Yorick Wilks
NLP Research Group
Department of Computer Science
University of Sheffield
{a.dalli, y.xia, y.wilks}@dcs.shef.ac.uk

Abstract

Email summarisation presents a unique set of requirements that are different from general text summarisation. This work describes the implementation of an email summarisation system for use in a voice-based Virtual Personal Assistant developed for the EU FASiL Project. Evaluation results from the first integrated version of the project are presented.

1 Introduction

Email is one of the most ubiquitous applications used on a daily basis by millions of people world-wide, traditionally accessed over a fixed terminal or laptop computer. In the past years there has been an increasing demand for email access over mobile phones. Our work has focused on creating an email summarisation service that provides quality summaries adaptively and quickly enough to cater for the tight constrains imposed by a real time text-to-speech system.

This work has been done as part of the European Union FASiL project, which aims to aims to construct a conversationally intelligent Virtual Personal Assistant (VPA) designed to manage the user's personal and business information through a voice-based interface accessible over mobile phones.

As the quality of life and productivity is to improved in an increasingly information dominated society, people need access to information anywhere, anytime. The Adaptive Information Management (AIM) service in the FASiL VPA seeks to automatically prioritise and present information that is most pertinent to the mobile users and adapt to different user preferences. The AIM service is comprised of three main parts: an email summariser, email categoriser, calendar scheduling/PIM interaction and an adaptive prioritisation service that optimizes the sequence in which information is presented, keeping the overall duration of the voice-based dialogue to a minimum.

2 Email Characteristics

Email Summarisation techniques share many characteristics with general text summarisation techniques while catering for the unique characteristics of email:

1. short messages usually between 2 to 800 words in length (after thread-filtering)
2. frequently do not obey grammatical or conventional stylistic conventions
3. are a cross between informal mobile text or chat styles and traditional writing formats
4. display unique thread characteristics with 87% containing three previous emails or less (Fisher and Moody, 2001)

All these four main characteristics combined together mean that most document summarisation techniques simply do not work well for email. The voice-based system also required that summaries be produced on demand, with only a short pause allowed for the summariser to output a result – typically a maximum of around 1 second per email.

Another main constraint imposed in the FASiL VPA was the presence of two integer parameters – the preferred and maximum length of the summary. The maximum length constraint had to be obeyed strictly, while striving to fit in the summary into the preferred length. These performance and size constraints, coupled with the four characteristics of email largely determined the design of the FASiL Email Summariser.

2.1 Short Messages

Email is a form of short, largely informal, written communication that excludes methods that need large amounts of words and phrases to work well.

The main disadvantage is that sometimes the useful content of a whole email message is simply a one word in case of a yes/no answer to a question or request. The summariser exploits this characteristic by filtering out threads and other commonly repeated text at the bottom of the email text such as standard email text signatures. If the resulting text is very short and falls within the preferred length of the summary, the message can be output in its entirety to users. The short messages also make it easier to achieve relevancy in the summaries.

Inadvertently context is sometimes lost in the summary due to replies occurring in threaded emails. Also, emails containing lots of question-answer pairs can get summarised poorly due to the fixed amount of space available for the summary.

2.2 Stylistic Conventions and Grammar

Email messages often do not follow formal stylistic conventions and are may have a substantial level of spelling mistakes, abbreviations and other features that make text analysis difficult.

A simple spellchecker using approximate string matching and word frequency/occurrence statistics was used to match misspelled names automatically.

Another problem that was encountered was the identification of sentence boundaries, since more than 10% of the emails seen by the summariser frequently had missing punctuation and spurious line breaks inserted by various different email programs. A set of hand-coded heuristics managed to produce acceptable results, identifying sentence boundaries correctly more than 90% of the time.

2.3 Informal and Formal Styles

Email can often be classified into three categories: informal short messages – often sent to people whom are directly known or with whom there has been a prolonged discussion or interaction about a subject, mixed formal/informal emails sent to strangers or when requesting information or replying to questions, and formal emails that are generally electronic versions of formal letter writing.

The class of emails that cause most problems for summarisation purposes are the first two classes of e-mails. One of the main determining factors for the style adopted by people in replying to emails is the amount of time that lapses between replies. Generally email gets more formal as the time span between replies increases.

Informal email can also be recognised by excessive use of anaphora that need to be resolved properly before summarisation can take place. The summariser thus has an anaphora resolver that is capable of resolving anaphoric references robustly.

Linguistic theory indicates that as the formality of a text increases, the number of words in the deictic category will decrease as the number of words in the non-deictic category increase (and vice-versa). Deictic (or anaphoric) word classes include words that have variable meaning whose meaning needs to be resolved through the surrounding (usually preceding) context. Non-deictic word classes are those words whose meaning is largely context-independent, analogous to predicates in formal logic.

2.4 Threaded Emails

Many emails are composed by replying to an original email, often including part or whole of the original email together with new content, thus creating a thread or chain of emails. The first email in the thread will potentially be repeated many times over, which might mislead the summarisation process. A thread-detection filtering tool is used to eliminate unoriginal content in the email by comparing the contents of the current email with the content of previous emails. A study of over 57 user's incoming and outgoing emails found that around 30% of all emails are threaded. Around 56% of the threaded emails contained only one previous email – i.e. a request and reply, and 87% of all emails contained only three previous emails apart from the reply (Fisher and Moody, 2001).

Some reply styles also pose a problem when combined with threads. Emails containing a list of questions or requests for comments are often edited by the replying party and answers inserted directly inside the text of the original request, as illustrated in Figure 1.

```
> … now coming back to the issue
> of whether to include support for
> location names in the recogniser
> I think that we should include
> this - your opinions appreciated.
I agree with this.
```

Figure 1 Sample Embedded Answer

Figure 1 illustrates the main two difficulties faced by the summariser in this situation. While the threaded content from the previous reply should be filtered out to identify the reply, the reply on its own is meaningless without any form of context. The summariser tries to overcome this by identifying this style of embedded responses when the original content is split into chunks or is only partially included in the reply. The text falling before the answer is then treated as part of the reply. Although this strategy gives acceptable results in some cases, more research is needed into finding the optimal strategy to extract the right amount of context from the thread without either destroying the context or copying too much from the original request back into the summary.

3 Summarisation Techniques

Various summarisation techniques were considered in the design of the FASiL email summariser. Few operational email-specific summarisation systems exist, so the emphasis was on extracting the best-of-breed techniques from document summarisation systems that are applicable to email summarisation.

3.1 Previous Work

Many single-document summarisation systems can be split according to whether they are extractive or non-extractive systems. Extractive systems generate summaries by extracting selected segments from the original document that are deemed to be most relevant. Non-extractive systems try to build an abstract representation model and re-generate the summary using this model and words found in the original document.

Previous related work on extractive systems included the use of semantic tagging and co-reference/lexical chains (Saggion et al., 2003; Barzilay and Elhadad, 1997; Azzam et al., 1998), lexical occurrence/structural statistics (Mathis et al., 1973), discourse structure (Marcu, 1998), cue phrases (Luhn, 1958; Paice, 1990; Rau et al., 1994), positional indicators (Edmunson, 1964) and other extraction methods (Kuipec et al., 1995).

Non-extractive systems are less common – previous related work included reformulation of extracted models (McKeown et al., 1999), gist extraction (Berger and Mittal, 2000), machine translation-like approaches (Witbrock and Mittal, 1999) and generative models (De Jong, 1982; Radev and McKeown, 1998; Fum et al., 1986; Reihmer and Hahn, 1988; Rau et al., 1989).

A sentence-extraction system was decided for the FASiL summariser, with the capability to have phrase-level extraction in the future. Non-extractive systems were not likely to work as robustly and give the high quality results needed by the VPA to work as required. Another advantage that extractive systems still pose is that in general they are more applicable to a wider range of arbitrary domains and are more reliable than non-extractive systems (Teufel, 2003).

The FASiL summariser uses named entities as an indication of the importance of every sentence, and performs anaphora resolution automatically. Sentences are selected according to named entity density and also according to their positional ranking.

3.2 Summariser Architecture

The FASiL Summariser works in conjunction with a number of different components to present real-time voice-based summaries to users. Figure 2 shows the overall architecture of the summariser and its place in the FASiL VPA.

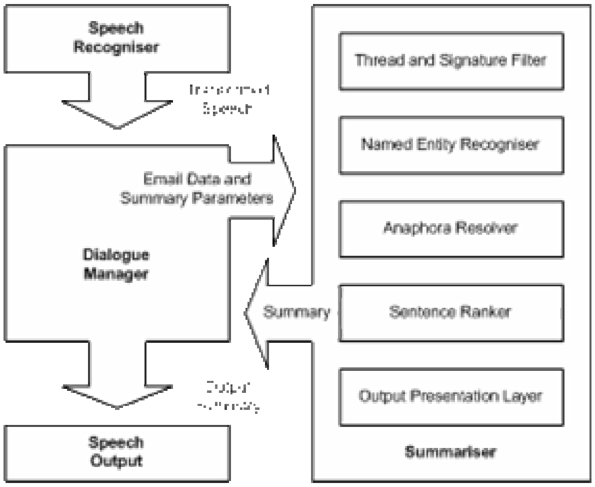

Figure 2 Summariser and VPA Architecture

An XML-based protocol is used to communicate with the Dialogue Manager enabling the system to be loosely coupled but to have high cohesion (Sommerville, 1992).

3.3 Named Entity Recognition

One of the most important components in the FASiL Summariser is the Named Entity Recogniser (NER) system.

The NER uses a very efficient trie-like structure to match sub-parts of every name (Gusfield, 1997; Stephen, 1994). An efficient implementation enables the NER to confirm or reject a word as being a named entity or not in O(*n*) time. Named entities are automatically classified according to the following list of 11 classes:

- Male proper names (M)
- Female proper names (F)
- Places (towns, cities, etc.) (P)
- Locations (upstairs, boardroom, etc.) (L)
- Male titles (Mr., Esq., etc.) (Mt)
- Female titles (Ms., Mrs., etc.) (Ft)
- Generic titles (t)
- Date and time references (TIME)
- Male anaphors (Ma)
- Female anaphors (Fa)
- Indeterminate anaphors (a)

The gazetteer list for Locations, Titles, and Anaphors were compiled manually. Date and time references were compiled from data supplied in the IBM International Components for Unicode (ICU) project (Davis, 2003). Place names were extracted from data available online from the U.S. Geological Survey Geographic Names Information System and the GEOnet Names Server (GNS) of the U.S. National Imagery and Mapping Agency (USGS, 2003; NIMA, 2003).

An innovative approach to gathering names for the male and female names was adopted using a small custom-built information extraction system that crawled Internet pages to identify likely proper names in the texts. Additional hints were provided by the presence of anaphora in the same sentence or the following sentence as the suspected proper name. The gender of every title and anaphora was manually noted and this information was used to keep a count of the number of male or female titles and anaphors associated with a particular name. This information enabled the list of names to be organised by gender, enabling a rough probability to be assigned to suspect words (Azzam et al., 1998; Mitkov, 2002).

An Internet-based method that verified the list and filtered out likely spelling mistakes and non-existent names was then applied to this list, filtering out incorrectly spelt names and other features such as online chat nicknames (Dalli, 2004).

A list of over 592,000 proper names was thus obtained by this method with around 284,000 names being identified as male and 308,000 names identified as female. The large size of this list contributed significantly to the NER's resulting accuracy and compares favourably with previously compiled lists (Stevenson and Gaizauskas, 2000).

3.4 Anaphora Resolution

Extracting systems suffer from the problem of dangling anaphora in summaries. Anaphora resolution is an effective way of reducing the incoherence in resulting summaries by replacing anaphors with references to the appropriate named entities (Mitkov, 2002). This substitution has the direct effect of making the text less context sensitive and implicitly increases the formality of the text.

Cohesion problems due to semantic discontinuities where concepts and agents are not introduced are also partially solved by placing emphasis on named entities and performing anaphora resolution. The major cohesion problem that still has not been fully addressed is the coherence of various events mentioned in the text.

The anaphora resolver is aided by the gender-categorised named entity classes, enabling it to perform better resolution over a wide variety of names. A simple linear model is adopted, where the system focuses mainly on nominal and clausal antecedents (Cristea et al., 2000). The search scope for candidate antecedents is set to the current sentence together with the three preceding sentences as suggested in (Mitkov, 1998) as empirical studies show that more than 85% of all cases are handled correctly with this window size (Mitkov, 2002). Candidate antecedents being discarded after ten sentences have been processed without the presence of anaphora as suggested in (Kameyama, 1997).

3.5 Sentence Ranking

After named entity recognition and anaphora resolution, the summariser ranks the various sentences/phrases that it identifies and selects the best sentences to extract and put in the summary. The summariser takes two parameters apart from the email text itself: a preferred length and a maximum length. Typical lengths are 160 characters preferred with 640 characters maximum, which compares to the size a mobile text message.

Ranking takes into account three parameters: named entity density and importance of every class, sentence position and the preferred and maximum length parameters.

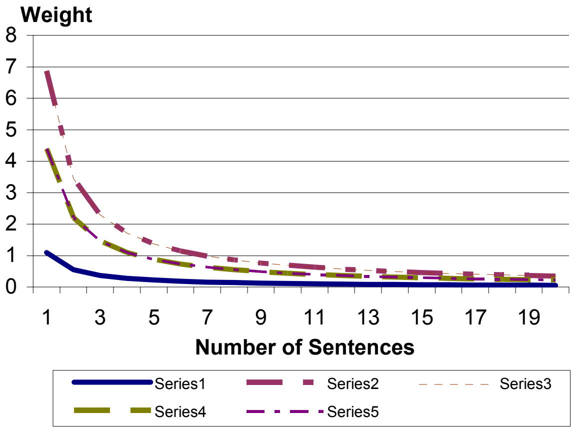

Figure 3 Positional sentence weight for varying summarisation parameters

Positional importance was found to be significant in email text since relevant information was often found to be in the first few sentences of the email.

Figure 3 shows how the quadratic positional weight function γ changes with position, giving less importance to sentences as they occur further from the start (although the weight is always bigger than zero). Different kinds of emails were used to calibrate the weight function. Series 1 (bottom) represents a typical mobile text message length summary with a very long message. Series 4 and 5 (middle) represent the weight function behaviour when the summary maximum length is long (approximately more than 1,000 characters), irrelevant of the email message length itself. Series 2 and 3 (top) represent email messages that fall within the maximum length constraints.

The following ranking function *rank*(*j*), where *j* is the sentence number, is used to rank and select excerpts:

$$rank(j) = (1-(\alpha+\beta))\sum_{i=0}^{\tau}(\tau_c(j,i)\omega(i)) + \alpha\left(\frac{j_{\max}}{j+1}\right) + \beta\left(\frac{\rho-(\lceil length(j)-\rho\rceil)}{\rho}\right)$$

where α and β are empirically determined constants, ρ is the preferred summary length, and $j_{\max}$ is the number of sentences in the email. The NER function τ_c represents the number of words of type i in sentence j and $\omega(i)$ gives the weight associated with that type. In our case τ equals 10 since there are 11 named entity classes. The NER weights $\omega(i)$ for every class have been empirically determined and optimized. A third parameter γ is used to change the values of α and β according to the maximum and preferred lengths together with the email length as shown in Figure 3.

The first term handles named entity density, the second the sentence position and the third biases the ranking towards the preferred length. The sentences are then sorted in rank order and the preferred and maximum lengths used to determine which sentences to return in the summary.

4 Experimental Results

The summariser results quality was evaluated against manually produced summaries using precision and recall, together with a more useful utility-based evaluation that uses a fractional model to cater for varying degrees of importance for different sentences.

4.1 Named Entity Recognition Performance

The performance of the summariser depends significantly on the performance of the NER. Speed tests show that the NER consistently processes more than 1 million wps on a 1.6 GHz machine while keeping resource usage to a manageable 300-400 Mb of memory.

Precision and recall curves were calculated for 100 emails chosen at random, separated into 10 random sample groups from representative subsets of the three main types of emails – short, normal and long emails as explained previously. The samples were manually marked according to the 11 different named entity classes recognised by the NER to act as a comparative standard for relevant results. Figures 4 and 5 respectively show the NER precision and recall results.

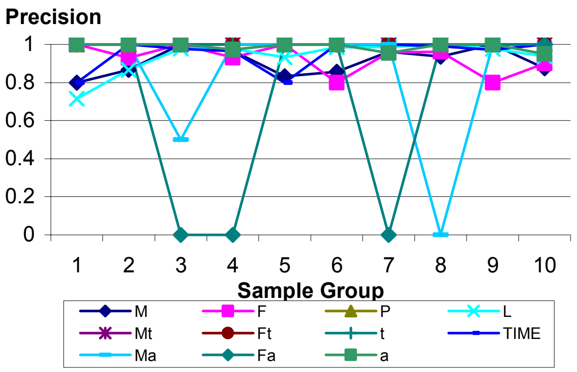

Figure 4 Precision by Named Entity Class

It is interesting to note that the NER performed worst at anaphora identification with an average precision of 77.5% for anaphora but 96.7% for the rest of the named entity classes.

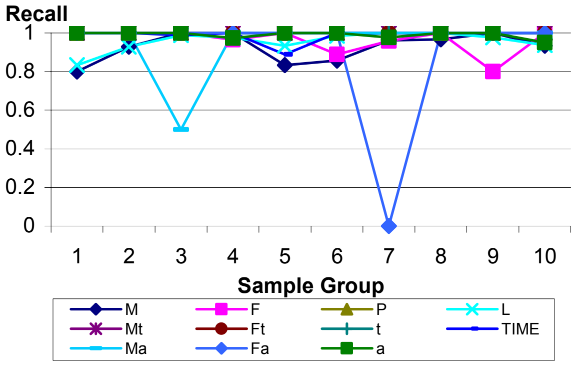

Figure 5 Recall by Named Entity Class

Figure 6 shows the average precision and recall averaged across all the eleven types of named entity classes, for the 10 sample email groups. An average precision of 93% was achieved throughout, with 97% recall.

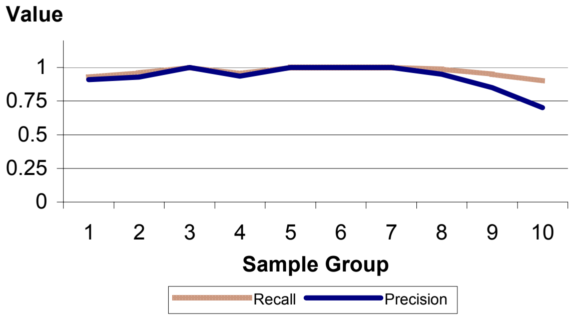

Figure 6 Average Precision and Recall

It is interesting to note that the precision and recall curves do not exhibit the commonly observed inverse trade-off relationship between precision and recall (Buckland and Gey, 1994; Alvarez, 2002). This result is explained by the fact that the NER, in this case, can actually identify most named entities in the text with high precision while neither over-selecting irrelevant results nor under-selecting relevant results.

4.2 Summariser Results Quality

Quality evaluation was performed by selecting 150 emails at random and splitting the emails up into 15 groups of 10 emails at random to facilitate multiple person evaluation. Each sentence in every email was then manually ranked using a scale of 1 to 10. For recall and precision calculation, any sentence ranked ≥ 5 was defined as relevant. Figure 7 shows the precision and recall values with 74% average precision and 71% average recall.

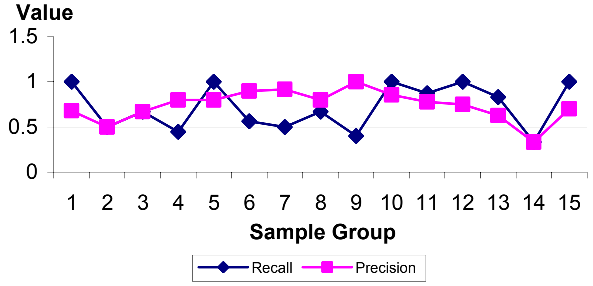

Figure 7 Summaries Recall and Precision

A utility-based evaluation was also used to obtain more intuitive results than those given by precision and recall using the methods reported in (Jing et al., 1998; Goldstein et al., 1999; Radev et al., 2000). The average score of each summary was compared to the average score over infinity expected to be obtained by extracting a combination of the first [1..N] sentences at random. The summary average score was also compared to the score obtained by an averaged pool of 3 human judges. Figure 8 shows a comparison between the summariser performance and human performance, with the summariser averaging at 86.5% of the human performance, ranging from 60% agreement to 100% agreement with the gold standard.

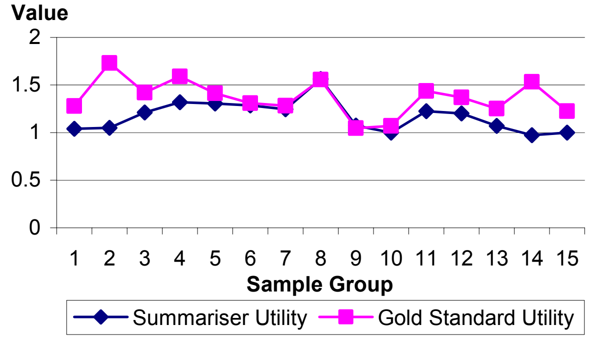

Figure 8 Utility Score Comparison

In Figure 8 a random extraction system is expected to get a score of 1 averaged across an infinite amount of runs. The average sentence compression factor for the summariser was 42%, exactly the same as the human judges' results. The selected emails had an average length of 14 sentences, varying from 7 to 27 sentences.

5 Conclusion and Future Work

The FASiL Email Summarisation system represents a compact summarisation system optimised for email summarisation in a voice-based system context.

The excellent performance in both speed and accuracy of the NER component makes it ideal for re-use in projects that need high quality real-time identification and classification of named entities.

A future improvement will incorporate a fast POS analyser to enable phrase-level extraction to take place while improving syntactic coherence. An additional improvement will be the incorporation of co-reference chain methods to verify email subject lines and in some cases suggest more appropriate subject lines.

The FASiL summariser validates the suitability of the combined sentence position and NER-driven approach towards email summarisation with encouraging results obtained.

Acknowledgments

This research is funded under the EU FASiL Project, an EU grant in Human Language Technology (IST-2001-38685) (Website: www.fasil.co.uk).

References

Alvarez, S. 2002. 'An exact analytical relation among recall, precision, and classification accuracy in information retrieval.' Boston College, Boston, Technical Report BCCS-02-01.

Azzam, S., Humphreys, K. and Gaizauskas, R. 1998. 'Coreference resolution in a multilingual information extraction', *Proc. Workshop on Linguistic Coreference*. Granada, Spain.

Barzilay, R. Elhadad, M. 1997. 'Using Lexical Chains for Text Summarization.', *Proc. ACL Workshop on Intelligent Scaleable Text Summarization*, Madrid, Spain. 10-17.

Berger, L. Mittal, V. 2000. 'OCELOT: A system for summarizing web pages'. Carnegie Mellon University. Just Research. Pittsburgh, Pennsylvania.

Buckland, M. Gey, F. 1994. 'The relationship between recall and precision.' *J. American Society for Information Science*, 45(1):12-19.

Cristea, D., Ide, N., Marcu, D., Tablan, V. 2000. 'An empirical investigation of the relation between discourse structure and coreference.', *Proc. 19th Int. Conf. on Comp. Linguistics (COLING-2000)*, Saarbrücken, Germany. 208-214.

Dalli, A. 2004. 'An Internet-based method for Verification of Extracted Proper Names'. CICLING-2004.

David, C. 2003. *Information Society Statistics: PCs, Internet and mobile phone usage in the EU.* European Community, Report KS-NP-03-015-EN-N.

Davis, M. 2003. 'An ICU overview'. *Proc. 24th Unicode Conference*, Atlanta. IBM Corporation, California.

De Jong, G. 1982. ‘An overview of the FRUMP system.’, in: Lehnert and Ringle eds., *Strategies for Natural Language Processing*, Lawrence Erlbaum Associates, Hillsdale, New Jersey. 149-176.

Edmunson, H.P. 1964. ‘Problems in automatic extracting.’, *Comm. ACM*, 7, 259-263.

Fisher, D., Moody, P. 2001. *Studies of Automated Collection of Email Records.* University of California, Irvine, Technical Report UCI-ISR-02-4.

Fum, D. Guida, G. Tasso, C. 1986. ‘Tailoring importance evaluation to reader’s goals: a contribution to descriptive text summarization.’ *Proc. COLING-86*, 256-259.

Goldstein, J. Kantrowitz, M. Mittal, V. Carbonell, Jaime. 1999. ‘Summarizing Text Documents: Sentence Selection and Evaluation Metrics’, *Proc. ACM-SIGIR 1999*, Berkeley, California.

Gusfield, D. 1997. *Algorithms on Strings, Trees and Sequences.* Cambridge University Press, Cambridge, UK.

Halliday, M.A.K. 1985. *Spoken and written language*. Oxford University Press, Oxford.

Jing, H. Barzilay, R. McKeown, K. Elhadad, M. 1998. ‘Summarization Evaluation Methods: Experiments and Analysis’, *AAAI Spring Symposium on Intelligent Text Summarisation*, Stanford, California.

Kameyama, M. 1997. ‘Recognising referential links: an information extraction perspective.’, *Proc. EACL-97 Workshop on Operational Factors in Practical, Robust, Anaphora Resolution*, Madrid, Spain. 46-53.

Kuipec, J. Pedersen, J. Chen, F. 1995. ‘A Trainable Document Summarizer.’, *Proc. 18th ACM SIGIR Conference*, Seattle, Washington. 68-73.

Luhn, P.H. 1958. ‘Automatic creation of literature abstracts’. *IBM J.* 159-165.

Marcu, D. 1998. ‘To Build Text Summaries of High Quality, Nuclearity is not Sufficient.’ *Proc. AAAI Symposium on Intelligent Text Summarisation*, Stanford University, Stanford, California. 1-8.

Mathis, B.A. Rush, J.E. Young, C.E. 1973. ‘Improvement of automatic abstracts by the use of structural analysis.’, *J. American Society for Information Science*, 24, 101-109.

McKeown, K. Klavens, J. Hatzivassiloglou, V. Barzilay, R. Eskin, E. 1999. ‘Towards Multidocument Summarization by Reformulation: Progress and Prospects.’, *AAAI Symposium on Intelligent Text Summarisation.*

Mitkov, R. 1998. ‘Robust pronoun resolution with limited knowledge.’, *Proc. 17th International Conference on Comp. Linguistics (COLING-1998)*, Montreal, Canada. 869-875.

Mitkov, R. 2002. *Anaphora Resolution*. London, Longman.

National Imagery and Mapping Agency (NIMA). 2003. *GEOnet Names Server (GNS).*

Paice, C. 1990. ‘Constructing literature abstracts by computer: techniques and prospects.’, *Information Processing and Management*, 26:171-186.

Radev, D. McKeown, K. 1998. ‘Generating Natural Language Summaries from Multiple On-Line Sources.’, *Computational Linguistics*, 24(3):469-500.

Radev, D. Jing, H. Budzikowska, M. 2000. ‘Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, user studies.’ in *Automatic Summarisation: ANLP/NAACL 2000 Workshop*, New Brunswick, New Jersey.

Rau, L. Jacobs, P. Zernick, U. 1989. ‘Information extraction and text summarization using linguistic knowledge acquisition.’, *Information Processing and Management*, 25(4):419-428.

Rau, L. Brandow, R. Mitze, K. 1994. ‘Domain-Independent Summarization of News.’, in: *Summarizing Text for Intelligent Communication*, Dagstuhl, Germany. 71-75.

Reimer, U. Hahn, U. 1988. ‘Text condensation as knowledge base abstraction.’ *Proc. 4th Conference on Artificial Intelligence Applications*. 338-344.

Saggion, H. Bontcheva, K. Cunningham, H. 2003. ‘Robust Generic and Query-based Summarisation’. *Proc. EACL-2003*, Budapest.

Sommerville, I. 1992. *Software Engineering*. 4th ed. Addison-Wesley.

Stephen, Graham A. 1994. *String Searching Algorithms*. World Scientific Publishing, Bangor, Gwynedd, UK.

Stevenson, M. Gaizauskas, R. 2000. ‘Using Corpus-derived Name Lists for Named Entity Recognition, *Proc. ANLP-2000*, Seattle.

Teufel, S. 2003. ‘Information Retrieval: Automatic Summarisation’, University of Cambridge. 24-25.

Witbrock, M. Mittal, V. 1999. ‘Ultra Summarization: A Statistical Approach to Generating Non-Extractive Summaries.’, Just Research, Pittsburgh.

United States Geological Survey (USGS). 2003. *Geographic Names Information System (GNIS).* http://geonames.usgs.gov/