Comparative Study of Automated Text Summarization Systems *

Laura Alonso* Irene Castellón* Salvador Climent‡
Maria Fuentes§ Lluís Padró † Horacio Rodríguez†

* Departament de Lingüística General
Universitat de Barcelona

‡ Estudis d'Humanitats i Filologia
Universitat Oberta de Catalunya

§ Departament d'Informàtica i Matemàtica Aplicada
Universitat de Girona

† TALP Research Center
Universitat Politècnica de Catalunya

Abstract

We present a comparative study of Automated Text Summarization Systems. We describe the factors to be taken into account for evaluating those systems. Then, we outline three alternative classifications of TS systems. Finally, we describe some TS systems according to their characterising features, performance and obtained results, and we place them in each of the proposed classifications.

* **Acknowledgements:** This research has been conducted thanks to projects HERMES (TIC2000-0335-C03-02) and PETRA (TIC2000-1735-C02-02), and it has also been partially funded by a grant associated to the X-TRACT project, PB98-1226 of the Spanish Research Department and by the project INTERLINGUA (IN3-IR226).

Contents

1 Introduction

The field of Text Summarization (TS) has experienced an exponential growth in the last years. That's why many comparative studies can be found in the literature. Among the most interesting, Paice (1990), Zechner (1997), Sparck-Jones (1999), Hovy and Marcu (1998), Tucker (1999), Radev (2000), Maybury and Mani (2001). Also the SUMMAC[1] and DUC contests provide a good overview of current working systems (SUMMAC, 1998; DUC, 2002).

Most of the comparative studies are divided in two parts: first, an account of factors influencing summaries is given. Then, a classification of summarization systems is proposed, and some systems are classified. In this study, we present the factors affecting summarization in section 2. In section 3 we outline three possible classifications of summarization systems, and in section 4 an extensive description of summarization systems is provided, giving the characteristics of each of them and the way they would be classed in each of the three classifications presented. To finish, we give a hint of some unsolved problems in TS, with a special focus in multidocument summarization.

2 Some considerations on Summary Aspects

Summarization can be decomposed into three phases: analysing the input text to obtain text representation, transforming it into a summary representation, and synthesising an appropriate output form to generate the summary text.

Effective summarising requires an explicit, and detailed, analysis of context factors, as is apparent when we recognise that what summaries should be like is defined by what they are wanted for. Sparck-Jones (1999) distinguishes three main aspects of summaries: input, purpose and output.

2.1 Input Aspects

- *Document Structure*: heterogeneous documental information can be found in the source, for example, labels that mark headers, chapters, sections, lists, tables, etc.
- *Domain*: domain-sensitive systems are restricted to a single domain, with varying degrees of portability. General purpose systems are not dependant on information about domains, although they can exploit it.
- *Specialization level*: a text may be broadly characterised as ordinary, specialised, or restricted, in relation to the presumed subject knowledge of the source text readers. This aspect can be considered the same as the *domain* aspect discussed above.
- *Scale*: different summarising strategies have to adopted to handle different text lengths.
- *Restriction on the language*: the language of the input can be general language or restricted to a sublanguage within a domain, purpose or audience. It may be necessary to preserve the sublanguage in the summary.
- *Media*: Multimedia vs textual.
- *Genre*: some systems exploit typical genre-determined characteritics of texts, for example, the pyramidal organization of newspaper articles, the argumentative development of a scientific article, etc. Some summarizers are independent of the type of document to be summarized while others are specialized on some type of documents: agency news, broadcast fragments, meeting recordings, e-mails, web pages, etc.
- *Unit*: the input to the summarization process can be a *single document* or *multiple documents*, both simple text or multimedia information such as imagery audio, or video. Text summarization has traditionally focused mainly on text input.
- *Language*: systems can be language-dependant or not.

[1]The SUMMAC contest was only carried out once, in 1998.

2.2 Purpose Aspects

- *Situation*: systems can be embedded in a larger system (MT, IR, QA) or with no precise context specification.
- *Audience*: summaries can be adapted to the needs of specific users, for example, prior knowledge. *Background* summaries assume that the reader's prior knowledge is poor, and so extensive information is supplied, and *just-the-news* are those kind of summaries conveying only the newest information. Additionally, briefings try to collect representative information of a collection of related documents.
- *Usage*: summaries can be sensitive to determined uses: retrieving source text, previewing a text, refreshing the memory of an already read text, sorting...

2.3 Output Aspects

- *Content*: a summary may include all aspects of a source text or it may focus on some specific ones, which can be determined by queries, subjects, etc. *Generic* summaries are text-driven, while *user-focused* (or query-driven) ones rely on a specification of a user's information need. The two basic computational approaches are top-down, using information extraction techniques, and bottom-up, using information retrieval. Top-down is used in query-driven summaries, when criteria of interest are encoded as a search specification, and this specification is used by the system to filter or analyse text portions. On the other hand, bottom-up is used in text-driven summaries, when generic importance metrics are encoded as strategies, which are applied over a representation of the whole text.
- *Format*: output can be plain text, or it can be formatted by headers, tags or by organization in fields.
- *Style*: a summary can be *informative*, if it covers the topics in the source text; *indicative*, if it indicates which topics are addressed in the original; *aggregative*, if it supplies information non present in the source text that completes some of its information; or *critical*, if it provides an additional valoration of the summarised text.
- *Production Process*: the resulting summary text can be an *extract*, if it is composed by literal fragments of text, or an *abstract*, if it is generated, although there are intermediate options. The type of summary output desired can be relatively polished, for example, if text is well-connected, or else more fragmentary in nature (e.g., list of phrases).
- *Surrogation*: summaries can stand in place of the source as a surrogate, or they can be linked to the source, or even be presented in the context of the source (e.g., by highlighting source text).
- *Length*: the targeted level of compression (ratio of summary length) crucially determines the result. Traditionally compression rates range from 1% to 30%.

3 Approaches to Text Summarization

There are several ways in which one can characterise different approaches to text summarization. In this section, we present three possible classifications of text summarization systems, but many others can be found in the literature (Hovy and Marcu, 1998; Radev, 2000; Maybury and Mani, 2001). The first classification is based in the level of processing that each system performs, closely following Mani and Maybury (1999), the second is based in the kind of textual information that is exploited, and the third is based in (Tucker, 1999).

3.1 Classification 1: Level of Processing

One useful way is to examine the level of processing of the text. Based on this, summarization can be characterised as approaching the problem at the surface, entity, or discourse level (Mani and Maybury, 1999).

3.1.1 Surface level

Surface-level approaches tend to represent information in terms of shallow features that are then selectively combined together to yield a salience function used to extract information. These features include:

- *Term frequency* statistics provide a thematic representation of text, assuming that important sentences are the ones that contain words that occur frequently. The score sentences increases for each frequent word. Early summarization systems (Luhn, 1958) directly exploit word distribution in the source.

- *Location* relies on the intuition that important sentences are located at positions that are usually genre-dependent. These positions can be determined automatically through training (Lin and Hovy, 1997). The lead method consists of just taking the first sentences. The title-based method assumes that words in titles and headings are positively relevant to summarization. A generalization of these methods is the OPP used by Hovy and Lin in their SUMMARIST system. They exploit machine learning techniques to identify the positions where relevant information is placed within different textual genres.

- *Bias*, presence of terms from the title or headings in the test, the initial part of text, or user's query.

- *Cue words* and *phrases*, which uses meta-linguistic markers (e.g., cues: "in summary", "in conclusion", "our investigation", "the paper describes"; or emphasisers: "significantly", "important", "in particular", "hardly", "impossible"), as well as domain-specific bonus phrases and stigma terms). These phrases can be detected automatically (Kupiec, Pedersen, and Chen, 1995; Teufel and Moens, 1997).

3.1.2 Entity-level

Entity-level approaches build an internal representation of the text by modelling text entities (simple words, compound nouns, named entities, etc.) and their relationships. These approaches tend to represent patterns of connectivity in the text (e.g., graph topology) to help to determine what is salient. Relations between entities include:

- *Similarity*: similar words are those whose form is similar, for example, those sharing a common stem (e.g., "similar" and "similarity"). Similarity can be calculated with linguistic knowledge or by character string overlap.

- *Proximity*: the distance between the text units where entities occur is a determining factor for establishing relations between entities.

- *Cohesion*: cohesion can be defined in terms of *connectivity*. Connectivity accounts for the fact that important text units usually contain entities that are highly connected in some kind of semantic structure. Cohesion can be approached by:

  · *Word co-occurrence*: words can be related if they occur in common contexts. Some applications are presented in: (Baldwin and Morton, 1998; McKeown et al., 1999). (Salton et al., 1997; Mitra, Singhal, and Buckley, 1997) apply IR methods at the document level, treating paragraphs in texts as documents are treated in a collection of documents. Using a traditional IR-based method, a word similarity measure is used to determine the set S_i of paragraphs that each paragraph P_i is related to. After determining relatedness scores S_i for each paragraph, paragraphs with the largest S_i scores are extracted.

In SUMMAC (Mani et al., 1998), in the context of query-based summarization, Cornell's Smart-based approach expands the original query, compares expanded query against paragraphs, and selects top three paragraphs (max 25% of original) that are most similar to the original query.

· *Local salience*: important phrasal expressions are given by a combination of grammatical, syntactic, and contextual parameters (Boguraev and Kennedy, 1997).

· *Lexical similarity*: words can be related by thesaural relationships (synonymy, hypernymy, meronymy relations). Barzilay (1997) details a system where Lexical Chains are used, based on Morris and Hirst (1991). This line has also been applied to Spanish, relying on EuroWordNet relations between words, by Fuentes and Rodríguez (2002). The assumption is that important sentences are those that are crossed by strong chains[2]. This approach provides a partial account of texts, since it focuses mostly on cohesive aspects. An integration of cohesion and coherence features of texts might contribute to overcome this, as Alonso and Fuentes (2002) point out.

· *Co-reference*: referring expressions can be linked, and co-reference chains can be built with co-referring expressions. Both Lexical Chains and Co-reference Chains can be priorised if they contain words in a query (for query-based summaries) or in the title. So, the preference imposed on chain is: query > title > document. Bagga and Baldwin (1998; Azzam, Humphrey, and Gazauskas (1999) use coreference chains for summarization. Baldwin and Morton (1998) exploit co-reference chains specifically for query-sensitive summarization.

Connectedness method (Mani and Bloedorn, 1999) represents map text with graphs. Words in the text are the nodes, and arcs represent adjacency, grammatical, co-reference, and lexical similarity-based relations.

– *Logical relations*, such as agreement, contradiction, entailment, and consistency.

– *Meaning representation-based relations*, establishing relations, such as predicate-argument, between entities in the text.

3.1.3 Discourse-level

Discourse-level approaches model the global structure of the text, and its relation to communicative goals. At this level, the following information can be exploited:

– *Format* of the document (e.g., hypertext markup, document outlines).

– *Threads of topics* can be revealed in the text

– *Rhetorical structure* of the text, representing argumentation or narrative structure. The main idea is that the coherence structure of a text can be constructed, so that the 'centrality' of the textual units in this structure will reflect their importance. A tree-like representation of texts is proposed by the Rhetorical Structure Theory (Mann and Thompson, 1988). (Ono, Sumita, and Miike, 1994) and (Marcu, 1997) attempt to use this kind of discourse representation in order to determine the most important textual units. They propose an approach to rhetorical parsing by discourse markers and semantic similarities in order to hypothesize rhetorical relations. These hypotheses are used to derive a valid discourse representation of the original text.

[2]Lexical chains have also been used in other NLP tasks, such as automatic extraction of interdocument links (Green, 1997).

3.2 Classification 2: Kind of Information

Summarization systems can be classified by the kind of information they deal with. According to this, we can distinguish between those exploiting lexical aspects of texts, those working with structural information and those trying to achieve deep understanding of texts.

3.2.1 Lexical

These approaches exploit the information associated to words in the texts. Some of them are very shallow, relying on the frequency of words, but some others apply lexical resources to obtain a deeper representation of texts. Beggining by the most shallow, the following main trends can be distinguished. A common assumption of these approaches is that repeated information could be a good indicator of importance:

- *Word Frequency* approaches assume that the most frequent words in text are the most representative of its content, and consequently fragments of text containing them are more relevant. Most systems apply some kind of filter to leave out of consideration those words that are very frequent but not indicative, for example, by the *tf*idf* metric or by excluding the so-called *stop words*, words with grammatical but no content meaning.

- *Domain Frequency* tries to determine the relevance of words by first assigning the document to a particular domain. Domain specific words have a previous relevance score, which serves as a comparison ground to adequately evaluate their frequency in a given text.

- *Concept Frequency* abstracts from mere word-counting to concept-counting. By use of an electronic thesaurus or WordNet, each word in the text is associated to a more general concept, and frequency is computed on concepts instead of particular words.

- *Cue words and phrases* can be considered as indicators of relative relevance or non-relevance of fragments of text in respect to the others.

- *Chains* can be built from lexical items which are related by conceptual similarity according to a lexical resource (*lexical chains*) or by identity, if they co-refer to the same entity (*co-reference chains*). The fragments of text crossed by most chains or by most important chains or by most important parts of chains can be considered the most representative of the text.

3.2.2 Structural Information

A second direction in TS tries to exploit information from the texts as structured entities. Since texts are structured in different dimensions (documental, discursive, conceptual), different kinds of structural information can be exploited. Beginning by the most shallow:

- *Documental Structure* exploits the information that texts carry in their format, for example, headings, sections, etc.

- *Textual Structure*: Some positions in text systematically contain the most relevant information, for example, the beginning paragraph of news stories. These positions are usually genre- or domain-dependant.

- *Conceptual structure* the chains mentioned in lexical approaches can be considered as a kind of conceptual structure.

- *Discursive Structure* can be divided in two main lines: linear or narrative and hierarchical or rhetoric. The first tries to account for *satisfaction-precedence*-like relations among pieces of text, the second explains texts as trees where fragments of text are related with each other by virtue of a set of rhetorical relations, mostly asymmetric.

3.2.3 Deep Understanding

Some approaches try to achieve understanding of the text in order to build a summary. Two main lines can be distinguished:

- *Top-down* approaches try to recognise pre-defined knowledge structures to texts, for example, templates or frames.
- *Bottom-up* approaches try to represent texts as highly conceptual constructs, such as scene. Others apply fragmentary knowledge-structures to clue parts of text, and then build a complete representation out of these small parts.

3.3 Classification 3: Richard Tucker 1999

This classification is taken from (Tucker, 1999). It considers four main directions in TS: summarising from attentional networks, sentence by sentence, from informational content and from discourse structure.

The classes proposed here are even less disjunct than those in the two previous classifications, thus every system can be considered as an instance of more than one of the classes. This shows the inadequacy of a taxonomic perspective on summarization systems, due to the heterogeneous kinds of knowledge and techniques that summarization systems tend to incorporate.

3.3.1 Attentional Networks

The approaches to summarization in this direction try to grasp what a text is 'about' by identifying concepts that are in some sense central to the text, on the basis of the occurrence of the same or related concepts in diferent parts of the source representation. *Aboutness* is represented as the links between these occurrences.

Frequency-based approaches exploit the frequency with which the concepts occur in the representation. In systems based in word frequency, attentional networks are only represented implicitly. Some systems account for frequency significance by applying IR techniques, such as the *tf*idf* measure. Others apply corpus-based statistical natural language processing, such as collocation or proper noun identification. Sill others try to abstract from individual words to achieve concept frequency, for example, by using lexicons or thesauri (Hovy and Lin, 1999).

On the other hand, some systems identify and exploit the *cohesive links* holding between parts of the source text. These links can be represented as graph-like structures (Skorokhod'ko, 1971), in form of lexical chains.

3.3.2 Sentence by Sentence

Some summarising systems decide for each sentence in the source text whether it is important for summarising, rather independently of the text as a whole. To do that, they rely on relevance or irrelevance marks that can be found in sentences, for example, *cue words*.

However, it must be noted that most of the systems applying sentence-by-sentence relevance ranking do not rely entirely in this method, but use it in combination with other methods that tend to consider the text as a whole.

3.3.3 Informational Content

Some approaches to summarization have tried to understand the text, that is to say, to achieve a representation of some or all of its meaning whereupon reasoning can be applied. This approach requires deeper analysis of the source text but allows the production of sofisticated summaries, for example, by applying NLGeneration techniques. However, these methods tend to be highly domain-dependant, because of the huge amount of information they require for processing.

3.3.4 Discourse Structure

Discourse structure is used by many systems in a limited way, for example, by trying to grasp a text's 'aboutness'. In contrast, some other methods apply discourse theories to the analysis of the source text in order to obtain a representation of their discourse structure. However, much of the work in this area has been largely theoretical.

3.4 Combined Systems

The predominant tendency in current systems is to integrate some of the techniques mentioned so far. Integration is a complex matter, but it seems the appropriate way to deal with the complexity of textual objects. In this section, we are going to present some examples of combination of different techniques.

There are several systems where different methods are combined. Among the most interesting are: Kupiec, Pedersen, and Chen (1995), Teufel and Moens (1997), Hovy and Lin (1999), Mani and Bloedorn (1999) where title-based method is combined with cue-location, position, and word-frequency based methods.

As the field progresses, summarization systems tend to use more and deeper knowledge. For example, IE techniques are becoming widely used. Many systems do not rely any more in a single indicator of relevance or coherence, but take into account as many of them as possible. So, the tendency is that heterogeneous kinds of knowledge are merged in increasingly enriched representations of the source text(s).

These enriched representations allow for adaptability of the final summary to new summarization challenges, such as multidocument, multilingual and even multimedia summarization. In addition, such a rich representation of text is a step forward generation or, at least, pseudo-generation by combining fragments of the original text. Good examples of this are McKeown et al. (2002), Lin and Hovy (2002), Daumé III et al. (2002), Lal and Rueger (2002) or Harabagiu and Lacatusu (2002), among others.

4 Summarization Systems

Tables 1 and 2 show how existing summarization systems would be classified according to each of the classifications presented in the previous section. Files with a complete description of some of these systems (marked with an asterisk) can be found in the Annex.

5 Problems to be solved

(Paice, 1990) pointed out that the main shortcomings of summarization systems up to the 1990s was their low representativity of the content in the source text and their lack of coherence. The field has evolved much since then, but the crucial problems of the area are still to be solved.

Much of the work in this area has treated the problem of text summarization from a predominant information-theoretic perspective. Therefore, texts have been modelled as mathematical objects, where relevance and redundancy could be defined in purely statistical terms. This approach seems specially valuable to produce a satisfactory representation of the content of a text. However, it fails in producing coherent texts, acceptable for human users.

The shortcomings of purely statistical approaches to text summarization are addressed from two different perspectives:

– Applying *machine learning* techniques. They have been used mainly for two purposes: classifying a sentence from a source text into relevant or non-relevant (Kupiec, Pedersen, and Chen, 1995; Aone, Okurowski, and Gorlinsky, 1998; Mani and Bloedorn, 1998; Lin, 1999; Hirao et al., 2002) and transforming a source sentence considered relevant into a summary sentence (Jing and McKeown, 2000; Knight and Marcu, 2000; Harabagiu and Lacatusu, 2002). Input for learning algorithms are usually texts with their corresponding abstracts. Therefore, the main shortcoming of this approach is to obtain large quantities of <text, abstract> tuples for a variety of textual genres.

– Resorting to *linguistic (symbolic) or world knowledge*. Understanding of texts, mainly through IE extraction techniques, seems a desirable way of producing quality summaries. Until recently, such techniques had only been applied for very restricted domains (McKeown and Radev, 1995). However, recent systems tend to incorporate IE extraction modules that perform a partial understanding of text, either by modelling the typical context of relevant pieces of information (Lal and Rueger, 2002; Kan and McKeown, 1999), or by applying general templates to find, organize and use the typical content of a kind of text or event (Harabagiu and Lacatusu, 2002; Daumé III et al., 2002). This use of IE techniques has produced the very good results, as is reflected in the high ranking of Harabagiu and Lacatusu (2002) in DUC 2002. A combination of deeper knowledge with surface clues of text seems to yield good results, too (Lin and Hovy, 2002).

Multidocument summarization is one of the major challenges in current summarization systems. It consists of producing a single summary of a collection of documents dealing with the same topic. The work has been mostly determined by the corresponding DUC task. Therefore, it has mainly focused in collections of news articles with a given topic. Remarkable progresses have been achieved in avoiding redundancy by clustering techniques, mainly based on the work in Carbonell and Goldstein (1998).

As for multilingual summarization, not much work has been done yet, but the roadmap for the DUC contests (Baldwin et al., 2000) contemplates this challenge in the near future of the area. When dealing with MDS new problems arise: lower compression factors implying a more aggressive condensation, anti-redundancy, temporal dimension, more challenging coreference task (inter-document), etc. Clustering of similar documents plays now a central role (Carbonell and Goldstein, 1998; Radev, Jing, and Budzikowska, 2000; Hatzivassiloglou et al., 2001). Selecting the most relevant fragments from each cluster and assuring coherence of the summaries coming from different documents are other important problems that are currently under development in MDS systems.

Last but not least, evaluation of summaries is a major issue, because objective judgements are needed to assess the progress achieved by different approaches. Some contests have been carried out to evaluate summarization systems with common, public procedures: the SUMMAC contest and the series of DUC contests. Specially the last has provided sets of criteria to evaluate summary quality in many different dimensions: informational coverage (precision and recall), suitability to length requirements, grammatical and discursive coherence, etc.

References

Allport, D. 1988. The ticc: parsing interesting text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 211–218.

Alonso, Laura and Maria Fuentes. 2002. Collaborating discourse for text summarisation. In *Proceedings of the Seventh ESSLLI Student Session.*

Angheluta, R., R. De Busser, and M-F. Moens. 2002. The use of topic segmentation for automatic summarization. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Aone, C., M. Okurowski, and J. Gorlinsky. 1998. Trainable scalable summarization using robust NLP and machine learning. In *COLING-ACL*, pages 62–66.

Aone, Chinatsu, Mary Ellen Okurowski, James Gorlinsky, and Bjornar Larsen. 1997. A scalable summarization system using robust NLP. In *Proceeding of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 66–73.

Azzam, Saliha, Kevin Humphrey, and Robert Gazauskas. 1999. Using coreference chains for text summarisation. In Amit Bagga, Brek Baldwin, and Sara Shelton, editors, *Proceedings of the ACL'99 Workshop on Coreference and Its Applications*, pages 77 – 84, University of Maryland, College Park, Maryland, USA, June. ACL.

Bagga, Amit and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation (LREC'98)*, pages 536–566, Granada.

Baldwin, Breck, Robert Donaway, Eduard Hovy, Elizabeth Liddy, Inderjeet Mani, Daniel Marcu, Kathleen McKeown, Vibhu Mittal, Marc Moens, Dragomir Radev, Karen Sparck Jones, Beth Sundheim, Simone Teufel, Ralph Weischedel, and Michael White. 2000. An evaluation road map for summarization research. TIDES, TIDES.

Baldwin, Breck and Thomas S. Morton. 1998. Dynamic coreference-based summarization. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing*, Granada, Spain, June.

Banko, M., V. Mittal, and M. Witbrock. 2000. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, ACL.*

Banko, Michele, Vibhu Mittal, Mark Kantrowitz, and Jade Goldstein. 1999. Generating extraction-based summaries from hand-written summaries by aligning text spans. In *Proceedings of PACLING-9*, Waterloo, Ontario, July.

Barzilay, Regina. 1997. *Lexical Chains for Summarization.* Ph.D. thesis, Ben-Gurion University of the Negev.

Barzilay, Regina and Michel Elhadad. 1997. Using lexical chains for text summarization. In Inderjeet Mani and Mark Maybury, editors, *Intelligent Scalable Text Summarization Workshop (ISTS'97)*, pages 10–17, Madrid. ACL/EACL.

Barzilay, Regina, Noemie Elhadad, and Kathy McKeown. 2001. Sentence ordering in multidocument summarization. In *HLT'01.*

Barzilay, Regina, Kathy McKeown, and Michel Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of ACL 1999.*

Benbrahim, M. and K. Ahmad. 1994. Computer-aided lexical cohesion analysis and text abridgement. Technical Report Computing Sciences Report CS-94-11, University of Surrey.

Berger, Adam and Vibhu Mittal. 2001. Ocelot: A system for summarizing web pages. In *Proceedings of the 23rd Annual Conference on Research and Development in Information Retrieval (ACM SIGIR)*, Athens.

Boguraev, Branimir and Christopher Kennedy. 1997. Salience-based content characterisation of text documents. In *Proceedings of ACL'97 Workshop on Intelligent, Scalable Text Summarisation*, pages 2–9, Madrid, Spain.

Boros, E., P.B. Kantor, and D.J. Neu. 2001. A clustering based approach to creating multi-document summaries. In *Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001*, New Orleans.

Brandow, Ronald, Karl Mitze, and Lisa F. Rau. 1995. Automatic condensation of electronic publications by sentence selectio. *Information Processing and Managemen*, 31(5):675–68.

Brunn, M., Y. Chali, and B. Dufou. 2002. The university of lethbridge text summarizer at DUC 200. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Caldwell, N. H. M. 1994. An investigation into shallow processing for summarisation. Technical Report Computer science tripos part II project, University of Cambridge Computer Laboratory.

Carbonell, Jaime G. and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR*, pages 335–336.

Carroll, J., G. Minnen, Y. Canning, S. Devlin, and J. Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology.*

Conroy, John M. and Dianne P. O'Leary. 2001. Text summarization via Hidden Markov Models. In *SIGIR 2001.*

Conroy, John M., Judith D. Schlesinger, Dianne P. O'Leary, and Mary Ellen Okurowski. 2001. Using HMM and Logistic Regression to generate extract summaries for DUC. In *Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001*, New Orleans, Louisiana.

Copeck, T., S. Szpakowicz, and N. Japkowic. 2002. Learning how best to summarize. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Cullingford, R. E. 1981. Sam. In Schank and Riesbeck, editors, *Inside Computer Understanding.* Lawrence Erlbaum Assoc., Hillsdale, NJ.

Daumé III, H., A. Echihabi, D. Marcu, D.S. Munteanu, and R. Soricut. 2002. GLEANS: A generator of logical extracts and abstracts for nice summaries. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Daumé III, Hal and Daniel Marcu. 2002. A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.*

DeJong, G. 1982. An overview of the frump system. In W. G. Lehnert and M. H. Ringle, editors, *Strategies for natural language processing.* Hillsdale, NJ: Lawrence Erlbaum, pages 149 – 176.

Dersy, J. 1996. Producing summary content indicators for retrieved texts. Master's thesis, University of Cambridge Department of Engineering.

DUC. 2002. DUC–document understanding conference. http://duc.nist.gov/.

Edmunson, H. P. 1969. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16(2):264 – 285, April.

Elhadad, Noemie and Kathleen R. McKeown. 2001. Towards generating patient specific summaries of medical articles. In *NAACL'01 Automatic Summarization Workshop*.

Farzindar, A., G. Lapalme, and H. Saggion. 2002. Summaries with SumUM and its expansion for document understanding conference (DUC 2002). In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Fuentes, Maria and Horacio Rodríguez. 2002. Using cohesive properties of text for automatic summarization. In *JOTRI'02*.

Gladwin, P., S. Pulman, and K. Sparck-Jones. 1991. Shallow processing and automatic summarising: a first study. Technical Report 223, University of Cambridge Computer Laboratory.

Goldstein, Jade, Vibhu Mittal, Mark Kantrowitz, and Jaime Carbonell. 1999. Summarizing text documents: Sentence selection and evaluation metrics. In *SIGIR-99*.

Green, Stephen J. 1997. *Automatically generating hypertext by computing semantic similarity*. Ph.D. thesis, University of Toronto.

Hahn, U. 1990. Topic parsing: Accounting for text macro structures in full-text analysis. *Information Processing and Management*, 26(1):135–170.

Harabagiu, S.M. and F. Lacatusu. 2002. Generating single and multi-document summaries with GISTEXTER. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Hatzivassiloglou, V., J. Klavans, M. Holcombe, R. Barzilay, M.Y. Kan, and K.R. McKeown. 2001. Simfinder: A flexible clustering tool for summarization. In *NAACL'01 Automatic Summarization Workshop*.

Hatzivassiloglou, Vassileios, Judith Klavans, and EleazarEskin. 1999. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *EMNLP/VLC'99*, Maryland.

Hermjakob, Ulf. 1997. *Learning Parse and Translation Decisions From Examples With Rich Context*. Ph.D. thesis, University of Texas at Austin.

Hirao, T., Y. Sasaki, H. Isozaki, and E. Maeda. 2002. Ntt's Text Summarization system for DUC-2002. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Hoey, Michael. 1991. *Patterns of Lexis in Text*. Describing English Language. Oxford University Press.

Hovy, Eduard, 2000. *Handbook of Computational Linguistics*, chapter 28: Text Summarization. Oxford University Press.

Hovy, Eduard and Chin-Yew Lin. 1999. Automated Text Summarization in SUMMARIST. In Mani and Maybury, editors, *Advances in Automatic Text Summarization*.

Hovy, Eduard and Daniel Marcu. 1998. Automated Text Summarization. COLING-ACL. tutorial.

Jing, Hongyan. 2000. Sentence simplification in automatic text summarization. In *ANLP-2000*.

Jing, Hongyan. 2001. *Cut-and-Paste Text Summarization*. Ph.D. thesis, Graduate School of Arts and Sciences, Columbia University.

Jing, Hongyan and Kathleen McKeown. 2000. Cut and paste based text summarization. In *1st Conference of the North American Chapter of the Association for Computational Linguistics*.

Kan, Min-Yen, Judith L. Klavans, and Kathleen R. McKeown. 2001. Domain-specific informative and indicative summarization for information retrieval. In *Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001*, New Orleans.

Kan, Min-Yen and Kathleen McKeown. 1999. Information extraction and summarization: Domain independence through focus types. Technical report, Computer Science Department, Columbia University, New York.

Karamuftuoglu, M. 2002. An approach to summarization based on lexical bonds. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Knight, Kevin and Daniel Marcu. 2000. Statistics-based summarization - step one: Sentence compression. In *The 17th National Conference of the American Association for Artificial Intelligence AAAI'2000*, Austin, Texas.

Kraaij, W., M. Spitters, and A. Hulth. 2002. Headline extraction based on a combination of uni- and multidocument summarization techniques. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Kraaij, W., M. Spitters, and M. van der Heijden. 2001. Combining a mixture language model and naive bayes for multi-document summarisation. In *Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001*, New Orleans, Louisiana.

Kupiec, Julian, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73. ACM Press.

Lal, P. and S. Rueger. 2002. Extract-based summarization with simplification. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Lehmam, Abderrafih. 1999. Text structuration leading to an automatic summary system: Rafi. *Information Processing and Managemen*, 35(2):181–191.

Lehnert, W. G. 1982. Plot units: a narrative summarization strategy. In W. G. Lehnert and M. H. Ringle, editors, *Strategies for natural language processing*. Hillsdale, NJ: Lawrence Erlbaum, pages 375 – 412.

Lin, Chin-Yew. 1999. Training a selection function for extraction. In *ACM-CIKM*, pages 55–62.

Lin, Chin-Yew and Eduard Hovy. 1997. Identifying topics by position. In *Proceedings of the Applied Natural Language Processing Conference (ANLP-97)*, pages 283–290, Washington, DC.

Lin, Chin-Yew and Eduard Hovy. 2001. NeATS: A multidocument summarizer. In *Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001*, New Orleans.

Lin, Chin-Yew and Eduard Hovy. 2002. NeATS in DUC 2002. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Lin, Chin-Yew and Eduard H. Hovy. 2000. The automated acquisition of topic signatures for Text Summarization. In *COLING-00*, Saarbrücken.

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159 – 165.

Mani, I., D. House, G. Klein, L. Hirschman, L. Obrst, T. Firmin, M. Chrzanowski, and B. Sundheim. 1998. The tipster SUMMAC text summarization evaluation: Final report. Technical report, DARPA.

Mani, Inderjeet and Eric Bloedorn. 1998. Machine learning of generic and user-focused summarization. In *AAAI*, pages 821–826.

Mani, Inderjeet and Eric Bloedorn. 1999. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1-2):35–67.

Mani, Inderjeet and Mark T. Maybury, editors. 1999. *Advances in automatic text summarisation.* MIT Press.

Mann, William C. and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organisation. *Text*, 3(8):234–281.

Marcu, Daniel. 1997. From discourse structures to text summaries. In Mani and Maybury, editors, *Advances in Automatic Text Summarization*, pages 82 – 88.

Maybury, Mark T. and Inderjeet Mani. 2001. Automatic summarization. ACL/EACL'01. tutorial.

McKeown, K., S.-F. Chang, J. Cimino, S. Feiner, C. Friedman, L. Gravano, V. Hatzivassiloglou, S. Johnson, D. Jordan, J. Klavans, A. Kushniruk, V. Patel, and S. Teufel. 2001. Persival, a system for personalized search and summarization over multimedia healthcare information. In *ACM+IEEE Joint Conference on Digital Libraries (JCDL 2001)*.

McKeown, K., D. Evans, A. Nenkova, R. Barzilay, V. Hatzivassiloglou, B. Schiffman, S. Blair-Goldensohn, J. Klavans, and S. Sigelman. 2002. The columbia multi-document summarizer for DUC 2002. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

McKeown, Kathleen, Judith Klavans, Vassileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *AAAI 99*.

McKeown, Kathleen R. and Dragomir R. Radev. 1995. Generating summaries of multiple news articles. In *ACM Conference on Research and Development in Information Retrieval SIGIR'95*, Seattle, WA.

Minel, Jean-Luc, Jean-Pierre Desclés, Emmanuel Cartier, Gustavo Crispino, Slim Ben Hazez, and Agata Jackiewicz. 2001. Résumé automatique par filtrage sémantique d'informations dans des textes. présentation de la plate-forme filtext. *Revue Technique et Science Informatique*.

Mitra, M., A. Singhal, and C. Buckley. 1997. Automatic Text Summarization by paragraph extraction. In Inderjeet Mani and Mark Maybury, editors, *Intelligent Scalable Text Summarization Workshop (ISTS'97)*, pages 39 – 46, Madrid. ACL/EACL.

Mittal, V., M. Kantrowitz, J. Goldstein, and J. Carbonell. 1999. Selecting text spans for document summaries: Heuristics and metrics. In *AAAI 1999*.

Mittal, Vibhu and Adam Berger. 2000. Query-relevant summarization using faqs. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, Hong Kong.

Morris, Jane and Graeme Hirst. 1991. Lexical cohesion, the thesaurus, and the structure of text. *Computational linguistics*, 17(1):21–48.

Muresan, S., E. Tzoukermann, and J. Klavans. 2001. Combining linguistic and machine learning techniques for email summarization. In *ACL-EACL'01 CoNLL Workshop*.

Myaeng, Sung Hyon and Myung-Gil Jang. 1999. Integrating digital libraries with cross-language ir. In *Proceedings of the 2nd Conference on Digital Libraries*.

Ono, K., K. Sumita, and S. Miike. 1994. Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, pages 344 – 348, Kyoto, Japan.

Otterbacher, J.C., A.J. Winkel, and D.R. Radev. 2002. The michigan single and multi-document summarizer for DUC 2002. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Paice, Chris D. 1981. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In R. N. Oddy, C. J. Rijsbergen, and P. W. Williams, editors, *Information Retrieval Research*. London: Butterworths, pages 172 – 191.

Paice, Chris D. 1990. Constructing literature abstracts by computer. *Information Processing & Management*, 26(1):171 – 186.

Pollock, J. J. and A. Zamora. 1975. Automatic abstracting research at chemical abstracts service. *Journal of Information and Computer Sciences*, 15(4):226–23.

Preston, K. and S. Williams. 1994. Managing the information overload. physics in business. Institute of Physics.

Radev, Dragomir, Sasha Blair-Goldensohn, and Zhu Zhang. 2001. Experiments in single and multi-document summarization using MEAD. In *First Document Understanding Conference*, New Orleans, LA, September.

Radev, Dragomir R. 2000. Text Summarization. ACM SIGIR. tutorial.

Radev, Dragomir R., Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. 2001. Interactive, domain-independent identification and summarization of topically related news articles. In *5th European Conference on Research and Advanced Technology for Digital Libraries*, Darmstadt.

Radev, Dragomir R., Weiguo Fan, and Zhu Zhang. 2001. Webinessence: A personalized web-based multi-document summarization and recommendation system. In *NAACL Workshop on Automatic Summarization*, Pittsburgh.

Radev, Dragomir R., Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization*, Seattle, Washington.

Rau, Lisa F., Paul S. Jacobs, and Uri Zernik. 1989. Information extraction and text summarisation using linguistic knowledge acquisition. *Information Processing & Management*, 25(4):419 – 428.

RIPTIDES. 2002. RIPTIDES: Rapidly Portable Translingual Information Extraction and Interactive Multidocument Summarization. http://www.cs.cornell.edu/Info/People/cardie/tides/.

Rush, J. E. and et al. 1971. Automatic abstracting and indexing. ii. production of abstracts by application of contextual inference and syntactic coherence criteria. *Journal of the American Society for Information Science*, 22(4):260 – 274.

Saggion, Horacio and Guy Lapalme. 2002. Generating informative and indicative summaries with SumUM. *Computational Linguistics*. Special Issue on Automatic Summarization.

Salton, Gerard, James Allan, and Chris Buckley. 1994. Automatic structuring and retrival of large text files. *CACM*, 37(2):97–108.

Salton, Gerard, Amit Singhal, M. Mitra, and C. Buckley. 1997. Automatic text structuring and summarization. *Information Processing and Management*, 33(3):193 – 207.

Schank, R. and R. Abelson. 1977. *Scripts, Plans, Goals, and Understanding*. Lawrence Erlbaum, Hillsdale, NJ.

Schiffman, Barry, Inderjeet Mani, and Kristian J. Concepcion. 2001. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In *EACL'01*.

Schlesinger, J.D., J.M. Conroy, M.E. Okurowski, H.T. Wilson, D.P. O'Leary, A. Taylor, and J. Hobbs. 2002. Understanding machine performance in the context of human performance for multi-document summarization. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Skorokhod'ko, E. F. 1971. Adaptive method of automatic abstracting and indexing. *Information processing*, 71.

Sparck Jones, K., S. Walker, and S. Robertson. 1998. A probabilistic model of information retrieval: Development and status. Technical Report N 446, University of Cambridge Computer Laboratory.

Sparck-Jones, Karen. 1999. Automatic summarising: factors and directions. In Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press.

Strzalkowski, Tomek, Jin Wang, and Bowden Wise. 1998. A robust practical text summarization. In Eduard Hovy and Dragomir Radev, editors, *AAAI Spring Symposium on Intelligent Text Summarisation*, pages 26 – 33, Stanford, California, March 23-25. American Association for Artificial Intelligence, AAAI Press.

SUMMAC. 1998. SUMMAC, the final report. http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/.

SweSum. 2002. http://www.nada.kth.se/ xmartin/swesum/index-eng.html.

Tait, J. L. 1983. Automatic summarizing of english texts. Technical Report 47, University of Cambridge Computer Laboratory.

Taylor, S. L. 1975. *Automatic abstracting by applying graphical techniques to semantic networks*. Ph.D. thesis, Northwestern University.

Teufel, Simone and Marc Moens. 1997. Sentence extraction as a classification task. In Inderjeet Mani and Mark Maybury, editors, *Intelligent Scalable Text Summarization Workshop (ISTS'97)*, pages 58 – 59, Madrid. ACL/EACL.

Tucker, Richard. 1999. *Automatic Summarising and the clasp system*. Ph.D. thesis, University of Cambridge.

Tzoukermann, E., S. Muresan, and J. Klavans. 2001. Gist-it: Summarizing email using linguistic knowledge and machine learning. In *ACL-EACL'01 HLT/KM Workshop.*

van Halteren, H. 2002. Writing style recognition and sentence extraction. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

White, M., D. McCullough, C. Cardie, V. Ng, and K. Wagstaff. 2001. Detecting discrepancies and improving intelligibility: Two preliminary evaluations of riptides. In *Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001*, New Orleans.

White, Michael and Claire Cardie. 2002. Selecting sentences for multidocument summaries using randomized local search. In *ACL Workshop on Automatic Summarization.*

White, Michael, Tanya Korelsky, Claire Cardie, Vincent Ng, David Pierce, and Kiri Wagstaff. 2001. Multi-document summarization via information extraction. In *Proceedings of the First International Conference on Human Language Technology Research.*

Witbrock, M. and V. Mittal. 1999. Ultra-summarization: A statistical approach to generating highly condensed nonextractive summaries,. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR-99).*

Young, S. R. and P. J. Hayes. 1985. Automatic classification and summarisation of banking telexes. In *Second Conference on Artificial Intelligence Applications*, pages 402–408, New York.

Zajic, D., B. Door, and R. Schwartz. 2002. Automatic headline generation for newspaper stories. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Zechner, Klaus. 1997. A literature survey on information extraction and Text Summarization. term paper, Carnegie Mellon University.

Zechner, Klaus. 2001. *Automatic Summarisation of Spoken Dialogues in Unrestricted Domains.* Ph.D. thesis, Carnegie Mellon University.

A Annex: Some Summarization Systems

A.1 ANES

- **Name**: ANES
- **Reference**: Brandow, Mitze, and Rau (1995)
- **Short Description**: it applies IR tecniques (*tf*idf*) to find the most relevant words in the text. Sentences are selected according to the number and distribution of relevant words they have, to their position in the text and to their position in respect to selected sentences.
- **System Features**
  - **Input**
    - document structure is a secondary source of relevance (location in text)
    - domain-insensitive
    - genre-insensitive
    - not multidocument
    - multilingual?
  - **Architecture**
    - not within a larger system
    - no external sources of knowledge
    - no pre-processing
    - no Machine Learning
  - **Output Facilities and Constraints**
    - level of compression?
    - no audience-sensitive
    - no specific usage
    - output format?
    - informative summaries
    - extract
    - surrogates source text
- **Evaluation** (Brandow, Mitze, and Rau, 1995) Summaries of news stories were evaluated by readers, who classified each summary as acceptable or unacceptable by comparison with the source text. Summaries produced by obtaining an initial segment of the source text outperformed Anes summaries.
- **Classification**
  - within Classification 1: surface
  - within Classification 2: lexical
  - within Classification 3: attentional networks
- **Comments**: applies IR tecniques.

A.2 Baldwin and Morton 1998

- **Name**
- **Reference**: Baldwin and Morton (1998)
- **Short description**: Uses co reference between the query and the text for performing indicative, user-focused (query-sensitive) summarization
- **System Features**
  - **Input**:
  - **Architecture**: The system is based on a rich linguistics processing that includes the following tasks:
    - NER
    - Tokenization
    - Sentence segmentation
    - POS tagging
    - Morphological analysis
    - Parsing
    - Argument detection
    - Co-reference resolution: Identity and Part-Whole, including nominal and verbal phrases, acronyms, events
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): entity
  - within classification 2 (kind of information): lexical
  - within classification 3 (Tucker, 1999): sentence by sentence
- **Comments**:

A.3 MULTIGEN

- **Name**: MULTIGEN
- **Reference**: Barzilay, McKeown, and Elhadad (1999), McKeown et al. (1999)
- **Short description**: Multi-document Summarization using Information Fusion and Reformulation
- **System Features**
  - **Input**: News articles presenting different descriptions of the same event.
  - **Architecture**:
    - identify similarities and differences across documents by statistical techniques (McKeown and Radev, 1995)
    - extract sets of similar sentences: THEMES
    - shallow syntactic analysis
    - order sets of similar sentences (Reformulation). Two different forms of implementing ordering are included: majority ordering and chronological ordering.
    - generation: Sentence generation begins with phrases, with paraphrases rules derived from corpus analysis. MULTIGEN takes profit of the experience of Columbia's group in NL Generation for building high quality summaries (not extracts but abstracts).
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): entity
  - within classification 2 (kind of information): structural
  - within classification 3 (Tucker, 1999): informational content
- **Comments**: MULTIGEN has been extended in several directions. See Columbia MDS (Barzilay, Elhadad, and McKeown, 2001) PERSIVAL (McKeown et al., 2001) and CENTRIFUSER (Kan, Klavans, and McKeown, 2001) among others.

A.4 Columbia MDS

- **Name**: Columbia MDS
- **Reference**: Barzilay, McKeown, and Elhadad (1999), Barzilay, Elhadad, and McKeown (2001), Hatzivassiloglou, Klavans, and EleazarEskin (1999), Hatzivassiloglou et al. (2001), McKeown et al. (2002)
- **Short description**: Enhanced version of MULTIGEN. Complex system that can be applied to different sources. It can be considered a sort of meta-summarizer.
- **System Features**
    - **Input**: Four different types of input that are identified in a way that the most appropriate summarizer is applied in each case. The system can deal with simple event, biography, multi-event and others.
    - **Architecture**: There is a pre-processing phase followed by a router that depending on the kind of input triggers the appropriate summarizer. For simple events the summarizer used is the conventional MULTIGEN, for biographies, DEMS (Schiffman, Mani, and Concepcion, 2001) with the bio configuration, for multi-event and others, DEMS with the default configuration.
    - **Output facilities and constraints**:
- **Evaluation**: DUC 2002, consistently among the top three systems (second or third). For extracts, it ranked second precisionwise and third recallwise. For abstracts, it ranked second coveragewise and third precisionwise.
- **Classification**
    - within classification 1 (level of processing): entity
    - within classification 2 (kind of information): structural
    - within classification 3 (Tucker, 1999): informational content
- **Comments**:

A.5 PERSIVAL

– **Name**: PERSIVAL

– **Reference**: McKeown et al. (2001)

– **Short description**: PERSIVAL (Personalized Retrieval and Summarization of Image, Video and Language). The system builds patient specific (tailored access for both patients and physicians) summaries of medical articles contained in a distributed multimedia patient care digital library. It is a Digital Library project.

– **System Features**

  · **Input**: Multimedia collections in the medical domain

  · **Architecture**: Multimedia search triggered by a concept from patient's data. The system includes the annotation and organization of large collections of video data. Video documents are segmented and a storyboard summary is produced. Video are indexed at syntactic and semantic levels. A set of content-based video search tools has been developed. The system includes the use of DEFINDER tool (for looking for definitions).

  · **Output facilities and constraints**:

– **Evaluation**:

– **Classification**

  · within classification 1 (level of processing): entity

  · within classification 2 (kind of information): understanding

  · within classification 3 (Tucker, 1999): informational content

– **Comments**:

A.6 CENTRIFUSER

- **Name**: CENTRIFUSER
- **Reference**: Kan, Klavans, and McKeown (2001)
- **Short description**: Multi-document Summarizer. CENTRIFUSER meets the needs of browsers and searchers in highly structured domains.
- **System Features**
  - **Input**:
  - **Architecture**: The system uses SIMFINDER (Hatzivassiloglou, Klavans, and EleazarEskin, 1999; Hatzivassiloglou et al., 2001), a flexible clustering tool for summarization (used also in MULTIGEN). This tool detects text similarity over short passages exploring linguistic features combinations via Machine Learning techniques. Among the primitive linguistic features we can find word co-occurrence, shared proper nouns, linked noun phrases, WN synonyms and semantically similar verbs. Composite features consist of pairs of simple features. An automatic feature detection system is applied and then the well-known ILP system, RIPPER, is performed. After clustering, the system uses key-terms for selecting one sentence or paragraph from each cluster (using the centroid method of Radev, Jing, and Budzikowska (2000)). The selected sentences are finally reordered by reformulation (in a similar way as in MULTIGEN).
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): discourse
  - within classification 2 (kind of information): understanding
  - within classification 3 (Tucker, 1999): informational content
- **Comments**:

A.7 Banko et al. 1999, Mittal et al. 1999

- **Name**:
- **Reference**: Banko et al. (1999), Mittal et al. (1999)
- **Short description**: Extraction-based summarization from hand-written summaries, i.e. going from abstracts to extracts, of single documents, by aligning text spans.
- **System Features**
  - **Input**:
  - **Architecture**: A *tl*tf* (term length * term frequency) measure is used for weighting the relevance of terms and NE. Mittal et al. (1999) focuses on the selection of spans for document summaries. Sentences from the original document are ranked according to their salience using two parameters for tuning the process: i) granularity, e.g. paragraph, sentence, etc. and ii) metric for ranking. Features at discourse level include:
    - length of the span
    - density of NEs
    - complexity of NPs
    - punctuation
    - thematic phrases
    - anaphora density

    There are also features at subdocument level (sentence, phrase and word). These include:
    - word length
    - communicative actions
    - thematic phrases
    - use of honorifics, auxiliary verbs, negation, prepositions, etc.
    - type of sentence (interrogative, evaluative, etc.)
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): discourse
  - within classification 2 (kind of information): understanding
  - within classification 3 (Tucker, 1999): informational content
- **Comments**: Related work includes Headline production (Banko, Mittal, and Witbrock, 2000; Berger and Mittal, 2001) and Ultrasummarization (Witbrock and Mittal, 1999).

A.8 FociSum

- **Name**: FociSum
- **Reference**: Kan and McKeown (1999), Kan, Klavans, and McKeown (2001)
- **Short description**: Summarizing long documents. Domain specific informative and indicative summarization for Information Retrieval. Closely related to CENTRIFUSER.
- **System Features**
  - **Input**:
  - **Architecture**: Summarization of long documents presents interesting characteristics that do not occurs in conventional summarization systems (usually applied to summarize news, articles, Web pages and so). In long documents summarization sentences to be extracted occurs in distant locations. So, coherence properties are of less importance here. Focisum is an hybrid system that merges: i) Information Extraction techniques (template-based), ii) Sentence extraction (including both sentence-based and lead-based strategies) and iii) based on the dynamically determined foci of the text (in this context focus is the topic). Foci are built from NE and multiword terms.
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): entity
  - within classification 2 (kind of information): understanding
  - within classification 3 (Tucker, 1999): attentional networks
- **Comments**:

A.9 Conroy et al. 2001

– **Name**:

– **Reference**: Conroy et al. (2001), Conroy and O'Leary (2001), Schlesinger et al. (2002)

– **Short description**: Statistical approaches to summarisation

– **System Features**

  · **Input**:

  · **Architecture**: Two different techniques are used in Conroy et al. (2001):

    * HMM, using as features the position in the sentence, the number of tokens and the number of pseudo-query terms.
    * Logistic Regression (LRM), using as features the number of query terms occurring in the sentence, the number of tokens (sentence length), the distance to the query terms and the position of the sentence.

  Conroy and O'Leary (2001) use pivoted GR matrix decomposition. A token-sentence matrix is built and from it the columns giving good coverage of the tokens are selected. Two different approaches are used for this process: a greedy election and a pivoted QR factorisation. Schlesinger et al. (2002) merged the LRM and HMM by including all the features of the LRM in the HMM. An additional feature was the conditional probability that a sentence is a summary sentence given that the previous sentence is. A post-process is run on extracted sentences to remove sentence starting discourse markers and boilerplate, to improve cohesiveness. An extensive investigation was carried out to account for human performance in multi-document summarization,. Conclusions were that single-document summaries could be used as a base for multi-document, but had to be enriched, possibly wiht discourse structure. Sentence pruning techniques were also found useful.

  · **Output facilities and constraints**:

– **Evaluation**: participated in DUC'01 and DUC'02. In the latter, it was ranked among the first systems, but did not beat the baselines.

– **Classification**

  · within classification 1 (level of processing): surface

  · within classification 2 (kind of information): lexical

  · within classification 3 (Tucker, 1999): informational content

– **Comments**:

A.10 SUMMARIST

- **Name**: SUMMARIST
- **Reference**: Hovy and Lin (1999), Hovy (2000), Lin and Hovy (2000)
- **Short description**: Extractive single document summarisation system
- **System Features**
  - **Input**:
  - **Architecture**: The system proceeds in three steps: Topic identification, Interpretation and Summary generation.
    * Topic identification implies previous acquisition of Topic Signatures and then the identification of a text span as belonging to a topic characterised by its signature. Topic Signatures are tuples of the form <Topic, Signature> where Signature is a list of weighted terms: <t1,w1>, <t2,w2>, ..., <tn,wn>. Topic signatures can be automatically learned ([Lin, 1997], [Lin, Hovy, 2000]). Topic identification, then, includes text segmentation (using TextTiling) and comparison of text spans with existing Topic Signatures.
    * The topic identified are fused during the interpretation (2nd step) of the process. The fused topics are then reformulated (expressed in new terms).
    * The last step is a conventional extractive task.
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): surface
  - within classification 2 (kind of information): lexical
  - within classification 3 (Tucker, 1999): attentional networks
- **Comments**:

A.11 NeATS

– **Name**: NeATS

– **Reference**: Lin and Hovy (2001), Lin and Hovy (2002)

– **Short description**: Multi-document summarizer presented in DUC-2001

– **System Features**

  · **Input**:

  · **Architecture**: NeATS proceeds in the following steps:

    1. extracting and ranking passages
       * Identification of key concepts for each topic group
       * Computing of unigram, bigram, trigram Topic Signatures
       * Removing words or phrases occurring in less than the half of texts
       * Saving signatures in a tree
       * Webclopedia query formation
       * Sentence-level IR giving to a ranked list of sentences
    2. Filtering for content: remove all sentences that are not within the first 10 sentences of a document, decrease ranking score of sentences containing stigma words.
    3. Enforcing cohesion and coherence by pairing each sentence with the lead sentence of the document
    4. Filtering for length: include sentences (paired with the corresponding lead sentence) that are most different from the included ones, until targeted length is satisfied.
    5. Ensuring chronological coherence

  · **Output facilities and constraints**:

– **Evaluation**: in DUC 2002, it was the system with highest precision and F1 measure, although it performed low in recall.

– **Classification**

  · within classification 1 (level of processing): entity

  · within classification 2 (kind of information): structural

  · within classification 3 (Tucker, 1999): informational content

– **Comments**:

A.12 Cut-and-Paste

- **Name**: Cut-and-Paste
- **Reference**: Jing (2000), Jing and McKeown (2000)
- **Short description**: Sentence Reduction for automatic text summarization. The system relates the phrases occurring in a summary written by a professional summarizer and the phrases occurring in the original document.
- **System Features**
  - **Input**:
  - **Architecture**: 6 editing operations (learned from the performance of human summarizers) are used for sentence reduction:
    - removing extraneous phrases
    - combining a reduced sentence with other reduced sentences
    - syntactic transformations
    - substitution with paraphrases
    - substitution with more general or more specific descriptors
    - reordering
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): surface
  - within classification 2 (kind of information): structural
  - within classification 3 (Tucker, 1999): informational content
- **Comments**:

A.13 Myaeng and Jang 1999

- **Name**:
- **Reference**: Myaeng and Jang (1999)
- **Short description**: Single document summarizer based on statistical techniques
- **System Features**
  - **Input**:
  - **Architecture**: The system uses two similarity measures for determining if a sentence belongs to the major content: a similarity between the sentence and the rest of the document and a similarity between the sentence and the title of the document. Two statistical techniques are applied, a Bayesian model based on 14 features (signature terms and positional information) and the Dempter-Shafer combination rule.
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): surface
  - within classification 2 (kind of information): lexical
  - within classification 3 (Tucker, 1999): informational content
- **Comments**:

A.14 Knight and Marcu 2000

- **Name**:
- **Reference**: Knight and Marcu (2000)
- **Short description**: This system is not a full summarizer but a sentence compressor. Sentence compressing is presented as a fundamental component of any high-quality non extractive summarizer
- **System Features**
  - **Input**:
  - **Architecture**: The system follows a statistical approach. Sentence compression is considered as a process of translation from a source language (full text) into a target language (summary). The process is accomplished following two different approaches: a conventional noise channel model and decision trees (using C4.5). The probabilistic models are trained on a corpus of <full text, summary> pairs.
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): surface
  - within classification 2 (kind of information): lexical
  - within classification 3 (Tucker, 1999): sentence by sentence
- **Comments**: an enhancement of this approach was carried out later on, applying the same technique to rhetorical parse trees, with a scope beyond the sentence (Daumé III and Marcu, 2002).

A.15 Kraaij et al. 2001

- **Name**:
- **Reference**: Kraaij, Spitters, and van der Heijden (2001)
- **Short description**: Probabilistic single document extractive summarizer.
- **System Features**
  - **Input**:
  - **Architecture**: The system follows a probabilistic approach. Two different statistical models are applied and their results are combined for selecting the sentences that have to be included in the summary. The former is a content-based language model (unigrams + smoothing) and the latter is based on non-content features (being or not the first sentence, containing cue phrases, sentence length, etc.)
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): surface
  - within classification 2 (kind of information): lexical
  - within classification 3 (Tucker, 1999): informational content
- **Comments**:

A.16 OCELOT

- **Name**: OCELOT
- **Reference**: Mittal and Berger (2000), **?**)
- **Short description**: Summarizing of Web pages. Gist of Web document based on probabilistic models.
- **System Features**
    - **Input**:
    - **Architecture**: OCELOT is one of the applications of a general probabilistic approach that models summarisation as a translation process between two languages, the language of full text and the language of summaries. Berger in his thesis applies conventional stochastic translation methods for summarizing. Three different examples of application are provided and OCELOT is one of them.
    - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
    - within classification 1 (level of processing): surface
    - within classification 2 (kind of information): lexical
    - within classification 3 (Tucker, 1999): informational content
- **Comments**:

A.17 Strzalkowski 1998

- **Name**:
- **Reference**: Strzalkowski, Wang, and Wise (1998)
- **Short description**: Query-based single document non-extractive summarizer
- **System Features**
  - **Input**:
  - **Architecture**: The system proceeds in two steps, Analysis and Generation. Analysis phase consists of three tasks: Feature extraction, feature synthesis and rule induction. As result a set of themes is identified. The system uses both simple and composite features. Simple features include word co-occurrence, noun phrases (detected with linkIT), WN synonyms and common semantic classes for verbs (following Levin's, see [Klavans, Kan, 1998]). Generation phase includes the performance of a content planner (based on the intersection of themes obtained in the previous phase and on a sentence planner) and a sentence generator.
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): entity
  - within classification 2 (kind of information): structural
  - within classification 3 (Tucker, 1999): informational content
- **Comments**:

A.18 Muresan et al. 2001, Tzoukermann et al. 2001

- **Name**:
- **Reference**: Muresan, Tzoukermann, and Klavans (2001), Tzoukermann, Muresan, and Klavans (2001)
- **Short description**: e-mail summarization combining Machine Learning and linguistic information.
- **System Features**
  - **Input**:
  - **Architecture**: The basic process consists on learning the salient NPs occurring in the text. The following features are used for the learning task:
    * for the head of the NP:
      - head-tf*idf (relevance)
      - head-focc (position of the first occurrence of head)
    * for the whole NP
      - np-tf*idf
      - np-focc
      - np-length-words
      - np-length-chars
      - sentence-position
      - paragraph-position
      - all constituents in the NP equally weighted

    Different ML methods have been applied including decision trees (C4.5) and rule induction (Ripper). The linguistic process include:
    * inflectional morphology processing
    * removing unimportant modifiers
    * removing common words
    * removing empty words
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): entity
  - within classification 2 (kind of information): understanding
  - within classification 3 (Tucker, 1999): attentional networks
- **Comments**:

A.19 Carbonell and Goldstein 1998, Goldstein et al. 1999

- **Name**:
- **Reference**: Carbonell and Goldstein (1998), Goldstein et al. (1999)
- **Short description**: CMU approach to both SDS and MDS combines criteria of query relevance and novelty.
- **System Features**
  - **Input**:
  - **Architecture**: The base of the system is the MMR (Maximal Marginal Relevance) metric. Important issues are the diversity-based re-ranking for reordering documents (in MDS), the relevant passage extraction, the anti-redundancy measures, the way of combining criteria of relevance and novelty (relevant novelty vs. declining relevance to users"s query). In the case of SDS the system ranks sentences from the original document according to their salience or their likelihood of being part of the summary. For doing so, a weighted score of both linguistic and statistical features is used. The weights are optimised according to application genres. Among linguistic features we can find: name, place, honorifics, quotations, thematic phrases, etc. Statistical features include cosine, *tf*idf*, pseudo-relevance feedback, query expansion, user interest profiles, etc. In the case of MDS different types of summaries can be produced using:
    - Common sections of documents
    - Common sections + unique sections of documents
    - Centroid
    - Centroid + outliers
    - Common sections + unique sections + time weighting factor
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): surface
  - within classification 2 (kind of information): lexical
  - within classification 3 (Tucker, 1999): informational content
- **Comments**:

A.20 Boros et al. 2001

- **Name**:
- **Reference**: Boros, Kantor, and Neu (2001)
- **Short description**: Multi-document summarization system
- **System Features**
  - **Input**:
  - **Architecture**: The system proceeds through the following steps i) From a document set a finite number of topics are extracted, ii) topics are ordered by importance, iii) a unique sentence is extracted from the collection for covering each topic; salience of sentences is computed using *tf*idf*, iv) sentences are clustered (several clustering techniques both hierarchical and non-hierarchical are experimented) and, finally, v) the summary is produced.
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): surface
  - within classification 2 (kind of information): lexical
  - within classification 3 (Tucker, 1999): informational content
- **Comments**:

A.21 MEAD

- **Name**: MEAD
- **Reference**: Radev, Jing, and Budzikowska (2000), Radev, Blair-Goldensohn, and Zhang (2001), Otterbacher, Winkel, and Radev (2002)
- **Short description**: Centroid-based multi-document summarization
- **System Features**
  - **Input**:
  - **Architecture**: MEAD begins identifying all the articles related to an emerging event (using the CIDR Topic Detection and Tracking system). CIDR produces a set of clusters. From each cluster a centroid is built. Then the sentences closest to the each of the centroids are selected to be included in the summary. CBSU (Centroid-based sentence utility) scores the degree of relevance of a particular sentence to the general topic of the entire cluster. CSIS (Cross-sentence informational subsumption) measures the overlap between the informational content of the sentences. CSIS is a similar measure than MMR. The difference is that CSIS is multi-document and query-independent while MMR is single-document and query-based. More recent versions of MEAD use a linear combination of three features: the centroid score and it assigns higher scores to sentences closer to the beginning of the document and to longer sentences.
  - **Output facilities and constraints**:
- **Evaluation**: DUC 2001, in DUC 2002 they had format problems (SGML tags)
- **Classification**
  - within classification 1 (level of processing): surface
  - within classification 2 (kind of information): lexical
  - within classification 3 (Tucker, 1999): informational content
- **Comments**:

A.22 eSseNSe, NewsInESSence, WebInESSence

– **Name**: eSseNSe, NewsInESSence, WebInESSence

– **Reference**: Radev et al. (2001), Radev, Fan, and Zhang (2001)

– **Short description**: eSseNSe is basically a system for clustering documents after/before retrieval, summarization single/multi-document, personalization and recommendation of documents. From it two systems applied respectively to news (NewsInESSence) and Web pages (WebInESSence) have been derived.

– **System Features**

  · **Input**:

  · **Architecture**: These systems are based on the CST (Cross-Document Structure Theory). CST (that is related to RST for single documents) proposes a taxonomy of the informational relationships between documents in clusters of related documents. In NewsInESSence the aim is finding, visualizing and summarizing a topic-based cluster of news stories. A user selects a single news story from a news Web site. The system searches for other live sources of news for other stories related to this one and presents summaries

  · **Output facilities and constraints**:

– **Evaluation**:

– **Classification**

  · within classification 1 (level of processing): surface

  · within classification 2 (kind of information): lexical

  · within classification 3 (Tucker, 1999): attentional networks

– **Comments**:

A.23 Schiffman et al. 2001

- **Name**:
- **Reference**: Schiffman, Mani, and Concepcion (2001)
- **Short description**: Multi-document summarizer producing Biographical Summaries combining linguistic knowledge with corpus statistics.
- **System Features**
  - **Input**:
  - **Architecture**: A number of modules co-operate for producing the summaries:
    - Sentence tokenizer
    - Alembic POS tagger
    - Nametag NER
    - Cass parser
    - Cross-document co-reference
    - Appositives
    - Relative clause weighting
    - Sentential description, following [Sagion, Lapalme, 2000]
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): entity
  - within classification 2 (kind of information): understanding
  - within classification 3 (Tucker, 1999): informational content
- **Comments**:

A.24 RIPTIDES

- **Name**: RIPTIDES
- **Reference**: White et al. (2001), White et al. (2001)
- **Short description**: user directed document summarizer combining the application of techniques of Information Extraction, Extraction-based Summarization and Natural Language Generation. The former reference refers to single-document summarization while the latter to multi-document summarization.
- **System Features**
  - **Input**:
  - **Architecture**: The system proceeds in the following steps:
    1. User information needs are acquired from the system
    2. Scenario templates are filled by an IE system
    3. IE output templates are merged into an event-oriented structure where comparable facts are grouped. For doing so SimFinder is used.
    4. Importance scores are assigned to slot/sentences based on a combination of document position, document recency and group/cluster membership.
    5. Content selection
    6. Summary generation
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): entity
  - within classification 2 (kind of information): understanding
  - within classification 3 (Tucker, 1999): informational content
- **Comments**:

A.25 DiaSumm

- **Name**: DiaSumm
- **Reference**: Zechner (2001)
- **Short description**: Automatic Summarization of Spoken Dialogues in Unrestricted Domains
- **System Features**: Dealing with non textual documents implies that additional problems have to be faced. If the input comes from ASR (with or without confidence scores), speech disfluencies have to be detected and removed. Besides, sentence boundaries have to be detected and inserted. Topic segmentation plays a more important role in this situation. In addition, in the case of multi-party dialogs, relations between moves have to be identified (e.g. linking of question/answering pairs).
  - **Input**: Spoken dialogues
  - **Architecture**: DiaSumm is organised in the following modules:
    1. speech disfluency detection and removal
    2. identification and insertion of sentence boundaries
    3. identification and linking of Question-Answer regions
    4. topical segmentation
    5. information condensation (using MMR)
  - **Output facilities and constraints**:
- **Evaluation**:
- **Classification**
  - within classification 1 (level of processing): surface
  - within classification 2 (kind of information): lexical
  - within classification 3 (Tucker, 1999): discourse structure
- **Comments**:

A.26 GISTEXTER

- **Name**: GISTEXTER
- **Reference**: Harabagiu and Lacatusu (2002)
- **Short description**: produces multidocument extracts and abstracts by template-driven IE. Templates are chosen by their adequacy to the topic of the document or collection of documents. Single document summaries by sentence extraction and compression.
- **System Features**
  - **Input**: collections of documents dealing with the same topic.
  - **Architecture**: for single documents, the most relevant sentences are extracted and compressed by rules that are learned from a corpus of human-written abstracts and their source texts (no further detail of these processes is given). For multi-document summarization, the system:
    * the IE system CICERO extracts relevant information by applying templates that are determined by the topic of the collection. Each template keeps a record of the text snippets where the information has been extracted from. If one of these snippets contains an anaphoric element, its co-reference chain is also recorded. If no template is provided for a given topic, a template is generated ad-hoc, based on the topical relations of the words in WordNet.
    * the *dominant event* of the collection is determined, and templates are classed depending on how central the dominant event is in the template and in the document where the template is extracted from.
    * within each class, templates are ordered by their representativeness. Highly representative templates are those that have the same slot fillers in the same slots as the majority of templates. Also those templates related to text snippets crossed by co-reference chains are more representative.
    * the summary is made from the text snippets recorded by the most representative template in the class of templates most closely related to the dominant event in the collection, in their order of appearance in the text. If they contain an anaphoric element, sentences containing the antecedent are also included. If the summary is too long, the linguistic form of dates and locations is shortened, unimportant coordinated phrases are dropped or, finally, the last sentence is dropped until the targeted length is achieved. If the summary is too short, the same process is applied to the most representative templates to the other classes of templates, in order of closeness to the dominant event.
  - **Output facilities and constraints**:
- **Evaluation**: participated in DUC 2002 and was ranked among the first. The best coverage rates for single and multi-document summarization, only surpassed by one system as to precision in multi-document summarization.
- **Classification**
  - within classification 1 (level of processing): entity/discourse
  - within classification 2 (kind of information): understanding
  - within classification 3 (Tucker, 1999): informational content
- **Comments**: the mentioned reference does not provide much detail on some of the modules of the system.

A.27 GLEANS

- **Name**: GLEANS
- **Reference**: Daumé III et al. (2002)
- **Short description**: IE-based multi-document summarizer, makes explicit the main entities and relations in a document collection. It produces headlines, extracts and a reduced form of abstract.
- **System Features**
  - **Input**:
  - **Architecture**: summarization in four steps:
    - documents are parsed (Hermjakob, 1997), the main constituents of each sentence are identified, some anaphoric expressions are resolved, and finally mapped into a canonical representation that explicits their main entities and relations
    - each collection of documents is classified by its content into *person*, *single event*, *multiple event* or *natural disaster*
    - given the collection type and the canonical representation of the documents, the core entities and relations are extracted, by choosing the most salient words in the collection.
    - a headline is created, based on the type of collection and teh core entities and relations. For *multiple event* collections, a short abstract can also be generated with the mechanisms to generate headlines.
    - an abstract is generated by applying a library of canonical schemas obtained from manual analysis of abstracts in a training corpus. These schemas determine which sentences of a source text fulfill the requirements of a canonical summary, and extract them. Chronological coherence, redundancy and dangling discourse references are treated.
    - in a post-process, dangling discourse markers are removed, decisions are made on which anaphoric expressions to use for each entity and temporal expressions are represented in a canonical form.
  - **Output facilities and constraints**:
- **Evaluation**: performance in DUC 2002 not high: low coverage, but improved when document collections were correctly classified. Specially bad on headline generation.
- **Classification**
  - within classification 1 (level of processing): entity/discourse
  - within classification 2 (kind of information): understanding
  - within classification 3 (Tucker, 1999): informational content
- **Comments**:

A.28 Angheluta et al. 2002

– **Name**:

– **Reference**: Angheluta, Busser, and Moens (2002)

– **Short description**: adapts a hierarchical topic segmetnation algorithm to text summarization

– **System Features**

  · **Input**:

  · **Architecture**: thematic structures in texts are detected using generic text structure cues:

    * lexical chains are built following Barzilay and Elhadad (1997) but using only WordNet synonymy relations.
    * the topic of each sentence is determined, by general topicality mechanisms of English (initial position, persistency).
    * topics are distinguished from subtopics, because the first spread throughout the whole text, while the second have local scope.
    * for single document summarization, the number of levels of the topic hierarchy is restricted by the targeted summary length, so that only sentences in higher levels are included.
    * for multiple document summarization, headline-kind summaries are produced by listing non-redundant topic terms. For longer summaries, open-class words of every sentence in the collection are clustered.

    Key terms are associated to each topic, and a tree-like table of content is produced.

  · **Output facilities and constraints**: oriented to tables of contents, lacks cohesion for texts.

– **Evaluation**: DUC 2002, average scores, bad for short abstracts

– **Classification**

  · within classification 1 (level of processing): entity

  · within classification 2 (kind of information): lexic

  · within classification 3 (Tucker, 1999): attentional networks

– **Comments**:

A.29 University of Lethbridge

- **Name**: University of Lethbridge
- **Reference**: Brunn, Chali, and Dufou (2002)
- **Short description**: single- and multidocument lexical chain summarizer by extraction. It filters out chain candidates in subordinate clauses.
- **System Features**
  - **Input**:
  - **Architecture**: for multidocument summaries, the procedure is the same as for single document (below), but all segments in the collection are pooled together, assigning a time stamp to each.
    - topic segmentation of the text
    - removing unimportant nouns from text (nouns in subordinate clauses).
    - lexical chaining
    - sentence extraction
    - surface repairs: add previous sentence to a sentence containing a dangling anaphora, remove short sentences or sentences with question or quotation marks.
    -
  - **Output facilities and constraints**:
- **Evaluation**: DUC 2002, but no results reported in reference
- **Classification**
  - within classification 1 (level of processing): entity
  - within classification 2 (kind of information): lexical
  - within classification 3 (Tucker, 1999): attentional networks
- **Comments**:

A.30 Lal and Rüger 2002

- **Name**:
- **Reference**: Lal and Rueger (2002)
- **Short description**: single-document, extract-based summarizer, applies anaphora resolution and text simplification.
- **System Features**
  - **Input**:
  - **Architecture**: following the approach of Kupiec, Pedersen, and Chen (1995), it works as a Bayesian pattern classifier over sentences trained from an annotated corpus. The features that are taken into account are: length of the sentence, position of the sentence within the paragraph and the paragraph within the document, mean *tf*idf* of named entities, co-reference with named entities in headline, inclusion of highly co-refered named entities. Some dangling anaphors are replaced by their referent. Lexical simplification is performed with tools from the PSET project (Carroll et al., 1998). Background knowledge on people and places, taken from sources on the web, can also be included.
  - **Output facilities and constraints**:
- **Evaluation**: DUC 2002, performed well except for grammaticality and coherence.
- **Classification**
  - within classification 1 (level of processing): entity/discourse
  - within classification 2 (kind of information): lexical/structural
  - within classification 3 (Tucker, 1999): sentence by sentence
- **Comments**: A demonstration can be found at http://km.doc.ic.ac.uk/pr-p.lal-2002/, and the system can be downloaded as a CREOLE Repository for GATE users.

A.31 SumUM

- **Name**: SumUM
- **Reference**: Saggion and Lapalme (2002), Farzindar, Lapalme, and Saggion (2002)
- **Short description**: generates single-document abstracts of scientific papers, based on shallow syntactic and semantic analysis oriented to conceptual identification and hand-made templates for text-regeneration. It interacts with the user.
- **System Features**
  - **Input**: single-document, scientific or technical articles with the following structure: title, author and affiliation, introduction, main section, references.
  - **Architecture**:
    - transducers identify concepts in text: domain transducers identify author, references, etc., and linguistic transducers identify noun groups and verb groups.
    - concepts are tagged semantically, marking discourse domain relations
    - sentences of indicative and informative type are identified
    - an indicative abstract is composed, by re-generation of text using pre-defined summary templates
    - based on the first, indicative abstract, an informative abstract can be composed, elaborating a specific query of the user
  - **Output facilities and constraints**: an interactive system: the user is presented with a short indicative abstract and a list of topics available for expansion, and an informative abstract can be produced, focusing on the topics chosen by the user.
- **Evaluation**: it was formally adapted to participate in DUC 2002, but with no adaptation to the news domain. It was ranked among the three first in quality, and the second in length-adjusted coverage, most probably due to the efficiency of templates.
- **Classification**
  - within classification 1 (level of processing): entity
  - within classification 2 (kind of information): understanding
  - within classification 3 (Tucker, 1999): informative content
- **Comments**:

A.32 Lexical Bonds

- **Name**: Lexical Bonds
- **Reference**: Karamuftuoglu (2002)
- **Short description**: extractive single-document system based on analysis of lexical bonds between sentences in a text and a classification of sentences into important and unimportant using SVM.
- **System Features**
  - **Input**: single documents
  - **Architecture**: the original design includes a transformation phase that should compact the text extracted in the first phase and resolve anaphoric references, but it is not yet developed. The current architecture is:
    * sentences are splitted and stopwords are removed
    * record of features for every sentence: sentence position, number of words, number of backward, forward and total lexical bonds and lexical links, and information content
      - a lexical link between two sentences is found when a word stem occurs in both of them, a lexical bond is found when there are two or more lexical links between a pair of sentences (Hoey, 1991).
      - the information content of a sentence is the IR function BM25 (Sparck Jones, Walker, and Robertson, 1998), which indicates the importance of the sentence with respect to the document.
    * SVM are used to select sentences according to these features (trained on DUC'02 manually selected extracts)
    * summaries are generated by following lexical bonds from a given sentence. Some constraints are: only sentences in the upper half of the document and selected by SVM are considered.

    The system produces cohesive summaries, but they are very redundant.
  - **Output facilities and constraints**: compactation process is under development.
- **Evaluation**: participated in DUC 2002, with good results in quality.
- **Classification**
  - within classification 1 (level of processing): surface/entity
  - within classification 2 (kind of information): discourse
  - within classification 3 (Tucker, 1999): attentional networks
- **Comments**:

A.33 TNO-TPD summarizer

- **Name**: TNO-TPD summarizer
- **Reference**: Kraaij, Spitters, and van der Heijden (2001), Kraaij, Spitters, and Hulth (2002)
- **Short description**: extractive multi-document summarizer. Sentences are selected according to a statistical language model and applying a bayesian classifier.
- **System Features**
  - **Input**:
  - **Architecture**:
    - an unigram language model of a cluster of documents determines content-based salience of each sentence
    - each sentence is assigned values for some surface features: sentence position, length, presence of positive or negative cue phrases, and the mentioned content score.
    - sentences are classified by a Naive Bayes classifier into summary and non-summary sentences.
    - redundancy is reduced by applying MMR (Carbonell and Goldstein, 1998)
    - to generate headlines, the most frequent word in the highest ranked sentence for every document and the titles is considered a *trigger word*. Then, the sentences in the whole cluster are ranked according to their importance. The highest ranked noun phrase that contains the trigger word is chosen as the headline.
  - **Output facilities and constraints**:
- **Evaluation**: participated in DUC 2002 in the multi-document extract and abstract tracks, with "disappointing performance". In addition, a self-evaluation applying relative utility (Radev, Jing, and Budzikowska, 2000), which reports better results. An investigation on the individual contribution of each feature was also performed, revealing that *position in the sentence* is highly indicative, while *negative cue phrase* was not well-defined.
- **Classification**
  - within classification 1 (level of processing): surface
  - within classification 2 (kind of information): lexic
  - within classification 3 (Tucker, 1999): attentional networks / sentence by sentence
- **Comments**:

A.34 NTT

- **Name**: NTT
- **Reference**: Hirao et al. (2002)
- **Short description**: single-document extractive summarizer based on classification of sentences by Support Vector Machines (SVM).
- **System Features**
  - **Input**:
  - **Architecture**: each sentence in the document is described with the following features: position, length, weight (*tf*idf* score of the words in the sentence), similarity with the headline and presence of certain prepositions or verbs.
  - **Output facilities and constraints**:
- **Evaluation**: participated in DUC'02, with good results in coverage but low quality.
- **Classification**
  - within classification 1 (level of processing): surface
  - within classification 2 (kind of information): lexical
  - within classification 3 (Tucker, 1999): sentence by sentence
- **Comments**:

A.35 van Halteren 2002

- **Name**:
- **Reference**: van Halteren (2002)
- **Short description**: multi-document, extractive summarizer. Sentences are classified by feature sets used for writing style recognition.
- **System Features**
  - **Input**:
  - **Architecture**: each sentence is described by a set of features: distance between occurrences of the same word, distribution of words, relative position of words, sentence length, sentence position and context of POS tags. A classifier trained for a writing style recognition task exploits these features for sentence scoring and extraction.
  - **Output facilities and constraints**:
- **Evaluation**: participated in DUC 2002, but obtained not so good results.
- **Classification**
  - within classification 1 (level of processing): surface
  - within classification 2 (kind of information): lexical
  - within classification 3 (Tucker, 1999): sentence by sentence
- **Comments**: the system was trained on materials not oriented to the summarization task

System	Level of Processing	Kind of Information	Tucker 1999
Adam Rush and et al. (1971) Pollock and Zamora (1975)	surface	structural	sentencewise
Angheluta, Busser, and Moens (2002)	entity	lexical	att. networks
* Anes Brandow, Mitze, and Rau (1995)	surface	lexical	att. networks
Barzilay and Elhadad (1997)	entity	lexical	att. networks
Boguraev and Kennedy (1997)	entity	lexical	att. networks
Caldwell (1994)	entity	lexical	att. networks
* CENTRIFUSER Elhadad and McKeown (2001)	discourse	understanding	info. content
* Columbia MDS McKeown et al. (2002)	entity/discourse	understanding	info. content
Copeck, Szpakowicz, and Japkowic (2002)	surface	lexical	att. networks
* Cut-and-Paste Jing (2001)	surface	structural	info. content
Dersy (1996)	entity	lexical	att. networks
* DiaSumm Zechner (2001)	surface	lexical	discourse structure
DimSum Aone et al. (1997)	surface	lexical	att. networks
Edmunson (1969)	surface	structural	sentencewise
FilText Minel et al. (2001)	surface	structural	info. content
* FociSum Kan and McKeown (1999)	entity	understanding	att. networks
Frump DeJong (1982)	entity	understanding	info. content
GISTexter Harabagiu and Lacatusu (2002)	discourse/entity	understanding	info. content
Gladwin, Pulman, and Sparck-Jones (1991)	entity	lexical	att. networks
* GLEANS Daumé III et al. (2002)	entity/discourse	understanding	info. content
Hirao et al. (2002)	surface	structural/lexical	att. networks
* Karamuftuoglu (2002)	surface	structural	att. networks
* Kraaij, Spitters, and Hulth (2002)	surface	lexical	att. networks
* Lal and Rueger (2002)	entity/discourse	understanding	info. content
Lehnert (1982)	entity	understanding	info. content
* Univ. of Lethbridge 2002 Brunn, Chali, and Dufou (2002)	entity	structural/lexical	att. networks
Luhn (1958)	surface	lexical	att. networks
Marcu (1997)	discourse	structural	disc. structure

Table 1: Classification of summarization systems - 1

System	Level of Processing	Kind of Information	Tucker 1999
* MEAD Radev, Blair-Goldensohn, and Zhang (2001) Otterbacher, Winkel, and Radev (2002)	surface	lexical	att. networks
* MultiGen McKeown et al. (1999) Barzilay, McKeown, and Elhadad (1999)	entity	structural	info. content
* NeATS Lin and Hovy (2001); Lin and Hovy (2002)	entitity	structural	info. content
NewsInEssence Radev et al. (2001)	surface	lexical	att. networks
Ono, Sumita, and Miike (1994)	discourse	structural	disc. structure
NetSumm Preston and Williams (1994)	surface	lexical	att. networks
Paice (1981)	surface	structural	sentencewise
* PERSIVAL McKeown et al. (2001)		understanding	info. content
Rafi Lehmam (1999)	surface	structural	att. networks
* RIPTIDES RIPTIDES (2002); White and Cardie (2002)	entity/discourse	understanding	info. content
Sam Schank and Abelson (1977) Cullingford (1981)	entity	understanding	info. content
Schlesinger et al. (2002)	surface	lexical	att. networks
Scisor Rau, Jacobs, and Zernik (1989)	entity	understanding	info. content
Scrabble Tait (1983)	entity	understanding	info. content
Skorokhod'ko (1971)	entity	lexical	att. networks
Smart Salton, Allan, and Buckley (1994) Mitra, Singhal, and Buckley (1997)	entity	lexical	att. networks
* SUMMARIST Hovy and Lin (1999)	surface	lexical	att. networks
SUMMONS McKeown and Radev (1995)	entity	understanding	info. content
SumUM Farzindar, Lapalme, and Saggion (2002)	discourse	structural	discourse structure
* SweSum SweSum (2002)	surface	lexical	att. networks
Taylor (1975)	entity	understanding	info. content
Tele-Pattan Benbrahim and Ahmad (1994)	entity	lexical	att. networks
Tess Young and Hayes (1985)	entity	understanding	info. content
TICC Allport (1988)	entity	understanding	info. content
TOPIC Hahn (1990)	discourse	structural	disc. structure
van Halteren (2002)	surface	lexical	att. networks
WebInEssence Radev, Fan, and Zhang (2001)	surface	lexical	att. networks
Zajic, Door, and Schwartz (2002)	surface	lexical	att. networks

Table 2: Classification of summarization systems-2

On-line Demos	
Centrifuser on-line demo	English multi-document (specific-topic medical documents) http://centrifuser.cs.columbia.edu/centrifuser.cgi
Copernic downloadable demo	English, French, German single document (many formats, Web pages incl.) http://www.copernic.com/desktop/products/summarizer/download.html
Extractor downloadable demo	English, French, Spanish, German, Japanese, Korean single document (many formats, Web pages incl.) http://www.dbi-tech.com/download_trial_dbiExtractor.asp [3]
Intelligent Miner for Text (IBM) no straightforward downloading	supposedly many languages single document (many formats, Web pages incl.) evaluation version at the "How to buy" Section, form at: http://www-3.ibm.com/software/data/iminer/fortext/summarize/summarize.html
Island InText no straightforward downloading	English Single document form at: http://www.islandsoft.com/orderform.html
Inxight Summarizer / LinguistX / Xerox PARC no straightforward downloading	supposedly any language single document (Web pages incl.) form at: http://www.inxight.com/products/core/summarizer/contact_sales.php
Kmaritime on-line demo	Korean http://nlplab.kmaritime.ac.kr/demo/
Lal and Rüger (2002) on-line demo project home at	English Single document http://rowan.doc.ic.ac.uk:8180/summarizer/demo.html http://www.doc.ic.ac.uk/ srueger/pr-p.lal-2002/home.html
Language Computer no straightforward acces	supposedly English no indications about language form at: http://www.languagecomputer.com/demos/summarization/index.html
MEAD / NewsInEssence / CLAIR on-line demo downloadable demo	English, Chinese multi-document http://www.clsp.jhu.edu/ws2001/groups/asmd/ http://www.newsinessence.com/nie.cgi
MS-Word Autosummarize	supposedly any language single document included in MS-Word
Pertinence Summarizer on-line demo	English, French, Spanish, German, Italian, Portuguese, Japanese, Chinese, Korean, Arabic, Greek, Dutch, Norwegian and Russian single document http://www.pertinence.net/register_en.html
Sinope Summarizer Personal Edition downloadable demo	English, Dutch, German single document http://www.sinope.nl/en/sinope/index.html
Summ-It on-line demo	probably English only pasted text http://www.mcs.surrey.ac.uk/SystemQ/summary/
Surfboard downloadable demo	probably English only single Web pages http://www.glu.com/binaries/surfboard/surfboard.dmg.gz
SweSum on-line demo	Danish, English, French, German, Spanish, Swedish Single document (Web pages or pasted text) (in two categories: Newspaper and Academic) http://www.nada.kth.se/ xmartin/swesum/index-eng.html
TextWise no straightforward access	probably English only single document or e-mail form at: http://www.textwise.com/solutions/crs/demo.html

Table 3: Some on-line demos of summarization systems, both commercial and academic