Approaches to Text Summarization: Questions and Answers

Laura Alonso, Irene Castellón
Dept. de Lingüística General
Universitat de Barcelona
Gran Via de les Corts Catalanes, 585
08007 Barcelona
{lalonso,castel}@lingua.fil.ub.es

Salvador Climent
Estudis d'Humanitats i Filologia
Universitat Oberta de Catalunya
Av. Tibidabo 39-43
08035 Barcelona
scliment@uoc.edu

Maria Fuentes
Dept. d'Informàtica i Matemàtica Aplicada
Universitat de Girona
Campus Montilivi
17071 Girona
maria.fuentes@udg.es

Lluís Padró, Horacio Rodríguez
TALP Research Center
Universitat Politècnica de Catalunya
Jordi Girona 1-3
08034 Barcelona
{padro,horacio}@lsi.upc.es

Abstract

In this paper a comparative study of Automated Text Summarization (TS) Systems is presented. It describes the factors to be taken into account for evaluating those systems and outlines three alternative classifications. The paper provides extensive examples of working TS systems according to their characterizing features, performance, and obtained results, with a special emphasis on the multilingual aspect of summarization.

Key Words: Automated Text Summarization, Multilingual Systems

1 Introduction

The field of Text Summarization (TS) has experienced an exponential growth in the last years. That is why many comparative studies can be found in the literature, among the most comprehensive, Paice (1990) [111], Zechner (1997) [150], Sparck-Jones (1998) [133], Hovy and Marcu (1998) [62], Tucker (1999) [142], Radev (2000) [118], Mani (2001) [88] and Maybury and Mani (2001) [95]. Given that an upper bound of performance for TS systems is still far from being reached, task-based competitions are the main forum of discussion in the area. As follows, the SUMMAC (1998) [134] and especially DUC (2001, 2002, 2003) [40] contests provide a good overview of current working systems.

In this study, we provide an analysis of current work in TS, with special attention to the future developments of the field, like multilingual summarization. First, we present the factors affecting summarization in Section 2, and provide examples of how working systems handle each of these factors. In Section 3 three possible classifications of summarization systems are outlined, which are applied to concrete systems in Section 4, with a concrete example of multilingual summarization. To finish, we briefly discuss some burning issues in TS.

Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. No 20 (2003), pp. 34-52.
ISSN: 1137-3601.

2 Some considerations on Summary Aspects

Summarization has traditionally been decomposed into three phases [133, 91, 52, 60, 88]:

- *analyzing* the input text to obtain text representation,
- *transforming* it into a summary representation,
- and *synthesizing* an appropriate output form to generate the summary text.

Effective summarizing requires an explicit and detailed analysis of context factors, as is apparent when we recognize that what summaries should be like is defined by what they are wanted for. The parameters to be taken into account in summarization systems have been widely discussed [91, 60, 88]. We will follow Sparck Jones (1998) [133], who distinguishes three main aspects that affect the process of TS: input, purpose and output, with a special focus on multilinguality.

2.1 Input Aspects

The features of the text to be summarized crucially determine the way a summary can be obtained. The following aspects of input are relevant to the task of TS:

Document Structure. Besides textual content, heterogeneous documental information can be found in a source document, for example, labels that mark headers, chapters, sections, lists, tables, etc. If it is well systematized and exploited, this information can be of use to analyze the document. For example, Kan (2002) [65] exploits the organization of medical articles in sections to build a tree-like representation of the source. Teufel and Moens (2002) [141] systematize the structural properties of scientific articles to assess the contribution of each textual segment to the article, in order to build a summary from that enriched perspective.

However, it can also be the case that the information it provides is not the target of the analysis. In this case, document structure has to be removed in order to isolate the textual component of the document.

Domain. Domain-sensitive systems are only capable of obtaining summaries of texts that belong to a pre-determined domain, with varying degrees of portability. The restriction to a certain domain is usually compensated by the fact that specialized systems can apply knowledge intensive techniques which are only feasible in controlled domains, as is the case of the multidocument summarizer SUMMONS [101], specialized in summaries in terrorism domain applying complex Information Extraction techniques. In contrast, general purpose systems are not dependant on information about domains, which usually results in a more shallow approach to the analysis of the input documents.

Nevertheless, some general purpose systems are prepared to exploit domain specific information. For example, the meta summarizer developed at Columbia University [17, 16, 55, 54, 98] applies different summarizers for different kinds of documents: MULTIGEN [17, 99] is specialized in simple events, DEMS [130] (with the bio configuration) deals with biographies, and for the rest of documents, DEMS has a default configuration that can be resorted to.

Specialization level. A text may be broadly characterized as ordinary, specialized, 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.

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.

Scale. Different summarizing strategies have to be adopted to handle different text lengths. Indeed, the analysis of the input text can be performed at different granularities, for example, in determining meaning units. In the case of news articles, sentences or even clauses are usually considered the minimal meaning units, whereas for longer documents, like reports or books, paragraphs seem a more adequate unit of meaning. Also the techniques for segmenting the input text in these meaning units differ: for shorter texts, orthography and syntax, even discourse boundaries [93] indicate significant boundaries, for longer texts, topic segmentation [70, 57] is more usual.

Media. Although the main focus of summarization is textual summarization, summaries of non-textual documents, like videos, meeting records, images or tables have also been undertaken in recent years. The complexity of multimedia summarization has prevented the development of wide coverage systems, which means that most summarization systems that can handle multimedia information are limited to specific domains or textual genres [56, 94]. However, research efforts also consider the integration of information of different media [19], which allow a wider coverage of multimedia summarization systems by exploiting different kinds of documental information collaboratively, like metadata associated to video records [145].

Genre. Some systems exploit typical genre-determined characteristics of texts, such as the pyramidal organization of newspaper articles, or the argumentative development of a scientific article. Some summarizers are independent of the type of document to be summarized, while others are specialized on some type of documents: healthcare reports [43], medical articles [65], agency news [101], broadcast fragments [56], meeting recordings [151], e-mails [105, 3], web pages [120], etc.

Unit. The input to the summarization process can be a *single document* or *multiple documents*, either simple text or multimedia information such as imagery audio, or video [135].

Language. Systems can be language-independant, exploiting characteristics of documents that hold cross-linguistically [117, 113], or else their architecture can be determined by the features of a concrete language. This means that some adaptations must be carried out in the system to deal with different languages. As an additional improvement, some multi-document systems are able to deal simultaneously with documents in different languages [29, 30], which will be developed in Section 2.4.

2.2 Purpose Aspects

Situation. TS systems can perform general summarization or else they can be embedded in larger systems, as an intermediate step for another NLP task, like Machine Translation, Information Retrieval or Question Answering. As the field evolves, more and more efforts are devoted to task-driven summarization, in detriment of a more general approach to TS. This is due to the fact that underspecification of the information needs supposes a major problem for design and evaluation of the systems. As will be discussed in Section 5, evaluation is a major problem in TS. Task-driven summarization presents the advantage that systems can be evaluated with respect to the improvement they introduce in the final task they are applied to.

Audience. In case a user profile is accessible, summaries can be adapted to the needs of specific users, for example, the user's prior knowledge on a determined subject. *Background* summaries assume that the reader's prior knowledge is poor, and so extensive information is supplied, while *just-the-news* are those kind of summaries conveying only the newest information on an already known subject. Briefings are a particular case of the latter, since they collect representative information from a set of related documents.

Usage. Summaries can be sensitive to determined uses: retrieving source text [66], previewing a text [78], refreshing the memory of an already read text, sorting...

2.3 Output Aspects

Content. A summary may try to represent all relevant features 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 the user's information need, like a question or key words.

Related to the kind of content that is to be extracted, different computational approaches are applied. The two basic approaches are top-down, using information extraction techniques, and bottom-up, more similar to information retrieval procedures. 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 analyze text portions. The strategies applied in this approach are similar to those of Question Answering. On the other hand, bottom-up is used in text-driven summaries, when generic importance metrics are encoded as strategies, which are then applied over a representation of the whole text.

Format. The output of a summarization system can be plain text, or else it can be formatted. Formatting can be targeted to many purposes: conforming to a pre-determined style (tags, organization in fields), improving readability (division in sections, highlighting), etc.

Style. A summary can be *informative*, if it covers the topics in the source text; *indicative*, if it provides a brief survey of the topics addressed in the original; *aggregative*, if it supplies information non present in the source text that completes some of its information or elicits some hidden information [141]; or *critical*, if it provides an additional valoration of the summarized 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. The type of summary output desired can be relatively polished, for example, if text is well-formed and connected, or else more fragmentary in nature (e.g., a list of key words).

There are intermediate options, mostly concerning the nature of the fragments that compose extracts, which can range from topic-like passages, paragraph or multiparagraph long, to clauses or even phrases. In addition, some approaches perform editing operations in the summary, overcoming the incoherence and redundancy often found in extracts, but at the same time avoiding the high cost of a NL generation system. Jing and McKeown (2000) [64] apply six re-writing strategies to improve the general quality of an extract-based summary by edition operations like deletion, completion or substitution of clausal constituents.

Surrogation. Summaries can stand in place of the source as a surrogate, or they can be linked to the source [66, 78], or even be presented in the context of the source (e.g., by highlighting source text, [76]).

Length. The targeted length of the summary crucially affects the informativeness of the final result. This length can be determined by a compression rate, that is to say, a ratio of the summary length with respect to the length of the original text. Traditionally, compression rates range from 1% to 30%, with 10% as a preferred rate for article summarization. In the case of multidocument summarization though, length cannot be determined as a ratio to the original text(s), so the summary always conforms to a pre-determined length. Summary length can also be determined by the physical context where the summary is to be displayed. For example, in the case of delivery of news of summaries to handhelds [20, 25, 35], the size of the screen imposes severe restrictions to the length of the summary. Headline generation is another application where the length of summaries is clearly determined [147, 37]. In very short summaries, coherence is usually sacrificed to informativeness, so lists of words are considered acceptable [71, 149].

2.4 Language coverage

As regards language coverage, systems can be classified as monolingual, multilingual, and crosslingual (a similar classification is commonly used in Information Retrieval systems). Monolingual summarization systems deal with only one language for both the input document and the summary. In the case of multilingual systems, input and output languages are also the same but in this case the system can cover several languages. Crosslingual systems are able to process input document in several languages, producing summaries in different languages.

Multilinguality does not imply additional difficulties. Most of the systems and techniques we will present below can be easily adapted to other languages, assuming, of course, the availability of the knowledge sources needed for the different methods. Roughly speaking, the more amount of linguistic knowledge is needed by a system, the more difficult is to transport it to another language.

A more complex challenge is crosslinguality. There are examples of single document crosslingual summarizers, implying a certain amount of translation, either on the input text or on the summary, but most crosslingual summarizers are multidocument. In this case a lot of problems specific of translinguality arise. Measures of similarity between documents and passages in different languages, for identifying relations or for clustering, have to be envisaged. Similarity between lexical units (words, NEs, multiword terms) belonging to different languages, have to be computed as well. Obviously, the more distant the involved languages are, the harder these problems turn to be, specially if the languages present different lexical units or character sets. Since this is a burning issue, it will be discussed at length in Section 5.

3 Approaches to Text Summarization

There are several ways in which one can characterize 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 [62, 118, 95, 88]. The first classification, following Mani and Maybury (1999) [91], is based in the level of processing that each system performs, the second, proposed in Alonso and Castellón (2001) [4], is based in the kind of information exploited, the third follows Tucker (1999) [142].

3.1 Classification 1: Level of Processing

One useful way to classify summarization systems is to examine the level of processing of the text. Based on this, summarization can be characterized as approaching the problem at the surface, entity, or discourse level [91].

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, following the approach of Edmunson (1969) [42]. 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 directly exploit word distribution in the source [86].

Location relies on the intuition that important sentences are located at positions that are usually genre-dependent, however, some general rules are the *lead method* and the *title-based method*. 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 [81], where they exploit Machine Learning techniques to identify the positions where relevant information is placed within different textual genres. Many of the current systems, specially those applying machine learning techniques, take into account the location of meaning units in a document to assess their relevance.

Bias. The relevance of meaning units is determined by the presence of terms from the title or headings, initial part of text, or user's query. For example, [33, 32, 131] use as features the position in the sentence, the number of tokens and the number of pseudo-query terms.

Cue words and *phrases* are signals of relevance or irrelevance. They are typically meta-linguistic markers (e.g., cues: "in summary", "in conclusion", "our investigation", "the paper describes"; or emphasizers: "significantly", "important", "in particular", "hardly", "impossible"), as well as domain-specific bonus phrases and stigma terms. Although lists of these phrases are usually built manually [72, 139], they can also be detected automatically.

3.1.2 Entity-level

Entity-level approaches build an internal representation of the text by modeling 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 determine saliency. 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. Myaeng and Jang (1999) [106] use 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. Also, in NTT [58, 59], CENTRIFUSER [66], several similarity measures are applied.

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)[13, 99]. Salton et al. (1997), Mitra et al. (1997) [128, 103] 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 [87], 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 [21].
- *Lexical similarity*: words can be related by thesaural relationships (synonymy, hypernymy, meronymy relations). Barzilay (1997) [14] details a system where Lexical Chains are used, based on Morris and Hirst (1991) [104]. This line has also been applied to Spanish, relying on EuroWordNet relations between words, by Fuentes and Rodríguez (2002) [48]. The assumption is that important sentences are those that are crossed by strong chains[1]. 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) [5] 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. Baga and Baldwin (1998), Azzam et al. (1999) [11, 10] use coreference chains for summarization. Baldwin and Morton (1998) [13] exploit co-reference chains specifically for query-sensitive summarization.

  Connectedness method [90] 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.

The system of Baldwin and Morton (1998) [13] uses argument detection in order to resolve co reference between the query and the text for performing summarization.

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. An example of this is SUMMARIST, which applies Topic identification [61, 85]. Topic identification implies previous acquisition of Topic Signatures (that can be automatically learned) and then the identification of a text span as belonging to a topic characterized by its signature. Topic identification, then, includes text segmentation and comparison of text spans with existing Topic Signatures. The topic identified are fused during the interpretation of the process. The fused topics are then expressed in new terms. Other systems are Boros et al. (2001) [22] and MEAD [121, 116, 109]. These systems assign a topic to the sentences in order to create clusters for selecting the sentences to appear in summary.

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 [92]. Ono et al. (1994) [108] and Marcu (1997) [93] 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.

[1] Lexical chains have also been used in other NLP tasks, such as automatic extraction of interdocument links [50].

3.2 Classification 2: Kind of Information

Summarization systems can be classified by the kind of information they deal with [4]. 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. Beginning 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 meaning content.

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 recognize predefined 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) [142]. It considers four main directions in TS: summarizing 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 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 different 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, by using lexicons or thesauri [61].

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 [132] as lexical chains.

3.3.2 Sentence by Sentence

Some summarizing systems decide for each sentence in the source text whether it is important for summarizing, 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 sophisticated summaries, for example, by applying NL generation techniques. However, these methods tend to be highly domain-dependant, because of the huge amount of information they require.

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, 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: [72, 141, 61, 90] 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 [98, 83, 37, 74, 53], among others.

4 Summarization Systems

Table 1 shows how existing summarization systems would be classified according to each of the classifications presented in the previous section. However, it must be taken into account that most current summarization systems are very complex, resorting to very heterogeneous information and applying varied techniques, so a classification will never be clear cut. Moreover, systems tend to evolve with time, which makes their classification still more controversial.

Files with a more extense description of some of these systems (marked with an asterisk) can be found in the Annex (in electronic version only). Additionally, Table 2 lists on-line or downloadable systems.

Multilinguality of the systems is one of the features in each describing file. It is stated whether the system can summarize only a single language, a definite set of languages, or whether its architecture permits unrestricted multilinguality. In this latter case, it is also stated whether experiments with different languages are reported.

As a concrete example of an approach to multilingual summarization, we present the systems developed within project HERMES[2]. The target of project HERMES is to adapt and apply language technologies for Spanish, Catalan, Basque and English to improve access to textual information in digital libraries, Internet, documental Intranets, etc. Therefore, HERMES summarization system should integrate multiple languages in a common architecture. Since the resources available for every language are uneven, this architecture has to be flexible enough to adapt to knowledge-poor representations of text but also to exploit rich representations when available.

[2] http://terral.ieec.uned.es/hermes/

EuroWordNet [144] is a general resource available for these four languages, so a first approach to summarization exploited this resource. A Lexical Chain summarizer was developed for Spanish [48]. As can be seen in Figure 1, the architecture of the summarizer permits easy adaptation to other languages, provided there is at least a morphological analyzer and a version of EuroWordNet available for the language. If other NLP tools are available, like Named Entity Recognizers or co-reference solvers, they can be easily integrated within the system. Once the text has been analyzed and Lexical Chains have been obtained, a summary is built by extracting candidate textual units from the text. Candidate units are chosen applying a certain heuristic, weighting some aspects of Lexical Chains.

A second approach to the task of summarization, seen in Figure 2, [47] tries to overcome this dependency on lexic applying Machine Learning techniques. The system is trained with a corpus of sentences described with a set of features, like position in the text, length, and also being crossed by a Lexical Chain. For each of these sentences, it is previously determined whether it belongs to a summary of the text or not, so that it can be learned which combinations of features characterize summary sentences. In a text to be summarized, each sentence is described with the same set of features, and it is determined whether these describing features characterize the sentence as a summary sentence or not. The summary is composed with sentences qualifying as summary sentences.

This second system does not require any specific feature to produce a summary, not even Lexical Chains. However, the more information available, the more accurate the learning process will be, which will result in better summaries. This approach has been evaluated for English within DUC 2003 contest, but it can be used straightforwardly for any other language, as long as there is a training corpus available.

5 Burning Issues

The field has experienced an exponential growth since its beginnings, but some crucial questions are still open.

5.1 Coherence of Summary texts

Paice (1990) [111] 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.

Much of the work in this area has treated the problem of text summarization from a predominant information-theoretic perspective. Therefore, texts have been modeled 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 on handling textual coherence 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 [72, 8, 89, 80, 58] and transforming a source sentence considered relevant into a summary sentence [64, 69, 53]. 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 *symbolic linguistic 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 [101]. However, recent systems tend to incorporate IE extraction modules that perform a partial understanding of text, either by modeling the typical context of relevant pieces of information [74, 67], or by applying general templates to find, organize and use the typical content of a kind of text or event [53, 37]. This use of IE techniques has produced very good results, as is reflected in the high ranking of Harabagiu and Lacatusu (2002) [53] in DUC 2002. A combination of deeper knowledge with surface clues seems to yield good results, too [83].

5.2 Multidocument summarization

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, mainly based on the work in Carbonell and Goldstein (1998) [27].

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 [27, 121, 54, 100]. Selecting the most relevant fragments from each cluster and assuring coherence of the summaries coming from different documents are other important problems, currently under development in MDS systems.

5.3 Multilingual summarization

As for multilingual summarization, not much work has been done yet, but the roadmap for the DUC contests [12] contemplates this challenge in the near future of the area.

The most well known Multilingual Summarization System is SUMMARIST [61]. The system extracts sentences in a variety of languages (English, Spanish, Japanese, etc.) and translates the resulting summaries. SUMMARIST 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 characterized by its signature. Topic Signatures are tuples of the form <Topic, Signature> where Signature is a list of weighted terms: $\{< t_1, w_1 >, < t_2, w_2 >, ..., < t_n, w_n >\}$. Topic signatures can be automatically learned [79, 85]. Topic identification, then, includes text segmentation (using Marti Hearst's TextTiling) and comparison of text spans with existing Topic Signatures. The identified topics are fused during interpretation, the second step of the process. The fused topics are then reformulated, that is to say, expressed in new terms. The last step is a conventional extractive task.

In order to face multilingual problems the involved knowledge sources have to be as much as possible language independent. In the case of SUMMARIST, sets of Topic Signatures have to be obtained for all the languages involved using the same procedures. Also the segmentation procedure is language independent. So, the accuracy of the resulting summaries depends heavily on the quality of the translators.

As has been said before, a more challenging issue is Crosslingual Multidocument Summarizers. Basically three main problems have to be addressed: 1) clustering of multilingual documents, 2) measuring the distance (or similarity) between multilingual units (documents, paragraphs, sentences, terms), and 3) automatic translation of documents or summaries. Most systems differ on the way they face these problems, the order of performance and the granularity of the units they deal with.

Evans and Klavans (2003) [44] present a platform for multilingual news summarization that extends the Columbia's Newsblaster system [96]. The system adds a new component, translation, to the original six major modules: crawling, extraction, clustering, summarization, classification and web page generation, that have been, in turn, modified for allowing multilinguality (language identification, different character encoding, language idiosyncrasy, etc.).

In this system multilingual documents are translated into English before clustering, so that clustering is performed only on English texts.

Translation is carried out at two levels. Because a low quality translation is usually enough for clustering purposes and assessing the relevance of the sentences, a simple and fast technique is applied for glossing the input documents prior to clustering. Higher (relatively) quality translation (using Altavista's Babelfish interface to Systran) is performed in a second step only over fragments selected to be part of the summary.

The system takes as well into account the possible degradation of the input texts as result of the translation process, since most of the sentences resulting from this process are simply not grammatically correct.

Chen et al. (2003) [30] consider three possibilities for scheduling the basic steps of document translation and clustering:

1. Translation before document clustering (as in Columbia's system), named one-phase strategy. This model clusters the multilingual multidocuments directly resulting in multilingual clusters.
2. Translation after document clustering, named two-phase strategy. This model clusters documents in each language separately and merges the clustering results.
3. Translation deferred to sentence clustering. First, monolingual clustering is performed at document level. All the documents in each cluster refer to the same event in a specific language. Then, for generating the extracted summary of an event all the clusters referring to this event are taken into account. Similar sentences of these multilingual clusters are clustered together, now at sentence level. Finally a representative sentence is chosen from each cluster and translated if needed.

The accuracy of this process depends basically on the form of computing the similarity between different multilingual units. Several forms of such functions are presented and empirially evaluated by the authors.

These measures are multilingual extensions of a baseline monolingual similarity measure. Sentences are represented as bag of words (only nouns and verbs are taken into account). The similarity measure is a function of the number of (approximate) matches between words and of the size of the bags. The matching function in the baseline reduces, except for NE, to the identity. In the multilingual variants of the formula, a bilingual dictionary is used as knowledge source for computing this matching.

Despite of its simplicity the position-free measure (the simplest one) seems to be the most accurate among the studied alternatives. In this approach the translations of all the words of the bag are collected and the similarity is computed as in the baseline. All the other alternatives constraint in some ways the possible mappings between words, using different greedy strategies. The results are, however, worse.

The two-phase strategy outperforms in the experiments the on-phase strategy. The third strategy, deferring the translation to sentence clustering, seems to be the most promising.

A system, covering English and Chinese, following this approach is presented in Chen and Lin (2000) [31]. The main components of the system are a set of monolingual news clusterers, a unique multilingual news clusterer and a news summarizer. A central issue of the system is the definition and identification of meaningful units as base for comparison. For English these units can be reduced to sentences but for Chinese the identification of units and the associated segmentation of the text can be a difficult task. Another important issue of the system (general for systems covering distant languages or different encoding schemata) is the need of a robust transliteration of names (or words not occurring in the bilingual dictionary) for assuring an accurate matching.

5.4 Evaluation

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.

An extensive investigation on the automatic evaluation of automatic summaries was carried out in a six-week workshop at Johns Hopkins University [122], where different evaluation metrics were proposed, including the *relative utility* method. Mani (2001) [88] provides a clear picture of the current state-of-the-art in evaluation, both with human judges and by automated metrics, with a special emphasis on content-based metrics. Hovy and Lin (2003) [84] show that the summaries produced by human judges are not reliable as a gold standard, because they strongly disagree with each other. A consensus summary obtained by applying content-based metrics, like unigram overlap, seems much more reliable as a golden standard against which summaries can be contrasted.

Acknowledgements

This research has been conducted thanks to projects HERMES (TIC2000-0335-C03-02), PETRA (TIC2000-1735-C02-02), and ALIADO (TIC2002-04447-C02-01) 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).

References

[1] Enrique Alfonseca and Pilar Rodríguez. Description of the UAM system for generating very short summaries at DUC-2003. In *HLT/NAACL Workshop on Text Summarization / DUC 2003*, 2003.

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

[3] Laura Alonso, Bernardino Casas, Irene Castellón, Salvador Climent, and Lluís Padró. CARPANTA eats words you don't need from e-mail. In *SEPLN, XIX Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 2003.

[4] Laura Alonso and Irene Castellón. Aproximació al resum automàtic per marcadors discursius. Technical report, CLiC, Universitat de Barcelona, Barcelona, 2001.

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

[6] R. Angheluta, R. De Busser, and M-F. Moens. 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 2002.

[7] Roxana Angheluta, Marie-Francine Moens, and Rik De Busser. K.u. leuven summarization system. In *DUC03*, Edmonton, Alberta, Canada, May 31 - June 1 2003. Association for Computational Linguistics.

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

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

[10] Saliha Azzam, Kevin Humphrey, and Robert Gaizauskas. 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 1999. ACL.

[11] Amit Bagga and Breck Baldwin. 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, 1998.

[12] Breck Baldwin, 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. An evaluation road map for summarization research. TIDES, TIDES 2000.

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

[14] Regina Barzilay. Lexical chains for summarization. Master's thesis, Ben-Gurion University of the Negev, 1997.

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

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

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

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

[19] A. B. Benitez and S.-F. Chang. Multimedia knowledge integration, summarization and evaluation. In *Proceedings of the 2002 International Workshop On Multimedia Data Mining in conjuction with the International Conference on Knowledge Discovery and Data Mining (MDM/KDD-2002)*, Edmonton, Alberta, 2002.

[20] Branimir Boguraev, Rachel Bellamy, and Calvin Swart. Summarisation miniaturisation: Delivery of news to hand-helds. In *NAACL'01*, 2001.

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

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

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

[24] M. Brunn, Y. Chali, and B. Dufou. The University of Lethbridge text summarizer at 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 2002.

[25] Orkut Buyukkokten, Hector Garcia-Molina, and Andreas Paepcke. Text summarization of web pages on handheld devices. In *NAACL'01*, 2001.

[26] N. H. M. Caldwell. An investigation into shallow processing for summarisation. Technical Report Computer science tripos

part II project, University of Cambridge Computer Laboratory, 1994.

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

[28] Y. Chali, M. Kolla, N. Singh, and Z. Zhang. The university of lethbridge text summarizer at DUC 2003. In *HLT/NAACL Workshop on Text Summarization / DUC 2003*, 2003.

[29] Hsin-Hsi Chen. Multilingual summarization and question answering. In *Workshop on Multilingual Summarization and Question Answering (COLING'2002)*, 2002.

[30] Hsin-Hsi Chen, June-Jei Kuo, and Tsei-Chun Su. Clustering and visualization in a multi-lingual multi-document summarization system. In *Proceedings of the 25th European Conference on IR Research*, pages 266–280, 2003.

[31] Hsin-Hsi Chen and Chuan-Jie Lin. A multilingual news summarizer. In *Proceedings of 18th International Conference on Computational Linguistics, COLING 2000*, pages 159–165, 2000.

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

[33] John M. Conroy, Judith D. Schlesinger, Dianne P. O'Leary, and Mary Ellen Okurowski. 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, 2001.

[34] T. Copeck, S. Szpakowicz, and N. Japkowic. 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 2002.

[35] Simon H. Corston-Oliver. Text compaction for display on very small screens. In *NAACL'01*, 2001.

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

[37] H. Daumé III, A. Echihabi, D. Marcu, D.S. Munteanu, and R. Soricut. 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 2002.

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

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

[40] DUC. DUC–document understanding conference. http://duc.nist.gov/.

[41] Daniel M. Dunlavy, John M. Conroy, Judith D. Schlesinger, Sarah A. Goodman, Mary Ellen Okurowski, Dianne P. O'Leary, and Hans van Halteren. Performance of a three-stage system for multi-document summarization. In *DUC03*, Edmonton, Alberta, Canada, May 31 - June 1 2003. Association for Computational Linguistics.

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

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

[44] David Kirk Evans and Judith L. Klavans. A platform for multilingual news summarization. Technical Report CUCS-014-03, Computer Science, University of Columbia, 2003.

[45] A. Farzindar, G. Lapalme, and H. Saggion. 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 2002.

[46] Atefeh Farzindar and Guy Lapalme. Using background information for multi-document summarization and summaries in response to a question. In *DUC03*, Edmonton, Alberta, Canada, May 31 - June 1 2003. Association for Computational Linguistics.

[47] Maria Fuentes, Marc Massot, Horacio Rodríguez, and Laura Alonso. Mixed approach to headline extraction for DUC 2003. In *HLT/NAACL Workshop on Text Summarization / DUC 2003*, Edmonton, Canada, 2003.

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

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

[50] Stephen J. Green. *Automatically generating hypertext by computing semantic similarity.* PhD thesis, University of Toronto, 1997.

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

[52] Udo Hahn and Inderjeet Mani. The cahllenges of automatic summarization. *IEEE Computer*, 33(11):29–36, 2000.

[53] S.M. Harabagiu and F. Lacatusu. 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 2002.

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

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

[56] A. G. Hauptmann and M. J. Witbrock. Informedia: News-on-demand multimedia information acquisition and retrieval. In M. Maybury, editor, *Intelligent Multimedia Information Retrieval*, pages 215–239. AAAI/MIT Press, 1997.

[57] Marti Hearst. Multi-paragraph segmentation of expository text. In *32nd Annual Meeting of Association for Computational Linguistics*, 1994.

[58] T. Hirao, Y. Sasaki, H. Isozaki, and E. Maeda. 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 2002.

[59] T. Hirao, J. Suzuki, H. Isozake, and E. Maeda. NTT's multiple document summarization system for DUC2003. In *HLT/NAACL Workshop on Text Summarization / DUC 2003*, 2003.

[60] Eduard Hovy. *Handbook of Computational Linguistics*, chapter 28: Text Summarization. Oxford University Press, 2001.

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

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

[63] Hongyan Jing. *Cut-and-Paste Text Summarization.* PhD thesis, Graduate School of Arts and Sciences, Columbia University, 2001.

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

[65] Min-Yen Kan. *Automatic text summarization as applied to information retrieval: Using indicative and informative summaries.* PhD thesis, Columbia University, 2003.

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

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

[68] M. Karamuftuoglu. 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 2002.

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

[70] Hideki Kozima. Text segmentation based on similarity between words. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*, pages 286–288, 1993.

[71] W. Kraaij, M. Spitters, and A. Hulth. 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 2002.

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

[73] Finley Lacatusu, Paul Parker, and Sanda Harabagiu. Lite-GISTexter: Generating short summaries with minimal resources. In *DUC03*, Edmonton, Alberta, Canada, May 31 - June 1 2003. Association for Computational Linguistics.

[74] P. Lal and S. Rueger. 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 2002.

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

[76] Abderrafih Lehmam and Philippe Bouvet. Évaluation, rectification et pertinence du résumé automatique de texte pour une utilisation en réseau. In S. Chaudiron and C. Fluhr, editors, *III Colloque d'ISKO-France: Filtrage et résumé automatique de l'information sur les réseaux*, 2001.

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

[78] Anton Leuski, Chin-Yew Lin, and Eduard Hovy. iNeATS: Interactive multi-document summarization. In *ACL'03*, 2003.

[79] C-Y. Lin. *Robust Automated Topic Identification.* PhD thesis, University of Southern California, 1997.

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

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

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

[83] Chin-Yew Lin and Eduard Hovy. 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 2002.

[84] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Marti Hearst and Mari Ostendorf, editors, *HLT-NAACL 2003: Main Proceedings*, pages 150–157, Edmonton, Alberta, Canada, May 27 - June 1 2003. Association for Computational Linguistics.

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

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

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

[88] Inderjeet Mani. *Automatic Summarization.* Nautral Language Processing. John Benjamins Publishing Company, 2001.

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

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

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

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

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

[94] M. Maybury and A. Merlino. Multimedia summaries of broadcast news. In *International Conference on Intelligent Information Systems*, 1997.

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

[96] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. Klavans, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of the Human Language Technology Conference*, 2002.

[97] K. McKeown, 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. Persival, a system for personalized search and summarization over multimedia healthcare information. In *ACM+IEEE Joint Conference on Digital Libraries (JCDL 2001)*, 2001.

[98] K. McKeown, D. Evans, A. Nenkova, R. Barzilay, V. Hatzivassiloglou, B. Schiffman, S. Blair-Goldensohn, J. Klavans, and S. Sigelman. 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 2002.

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

[100] Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Min-Yen Kan, Barry Schiffman, and Simone Teufel. Columbia multi-document summarization: Approach and evaluation. In *Proceedings of the Workshop on Text Summarization, ACM SIGIR Conference*, 2001.

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

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

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

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

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

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

[107] Ani Nenkova, Barry Schiffman, Andrew Schlaiker, Sasha Blair-Goldensohn, Regina Barzilay, Sergey Sigelman, Vasileios Hatzivassiloglou, and Kathleen McKeown. Columbia at the duc 2003. In *DUC03*, Edmonton, Alberta, Canada, May 31 - June 1 2003. Association for Computational Linguistics.

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

[109] J.C. Otterbacher, A.J. Winkel, and D.R. Radev. 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 2002.

[110] Chris D. Paice. 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*, pages 172 – 191. London: Butterworths, 1981.

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

[112] T.A.S. Pardo and L.H.M. Rino. DMSumm: Review and assessment. In E. Ranchhod and N. J. Mamede, editors, *Advances in Natural Language Processing*, pages 263–273. Springer-Verlag, 2002.

[113] T.A.S. Pardo, L.H.M. Rino, and M.G.V. Nunes. GistSumm: A summarization tool based on a new extractive method. In N.J. Mamede, J. Baptista, I. Trancoso, and M.G.V. Nunes, editors, *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken*, number 2721 in Lecture Notes in Artificial Intelligence, pages 210–218. Springer-Verlag, 2003.

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

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

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

[117] Dragomir Radev, Jahna Otterbacher, Hong Qi, and Daniel Tam. MEAD ReDUCs: Michigan at DUC 2003. In *DUC03*, Edmonton, Alberta, Canada, May 31 - June 1 2003. Association for Computational Linguistics.

[118] Dragomir R. Radev. Text Summarization. ACM SIGIR, 2000. tutorial.

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

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

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

[122] Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Arda Çelebi, Hong Qi, Elliott Drabek, and Danyu Liu. Evaluation of Text Summarization in a Cross-lingual Information Retrieval Framework. Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, June 2002.

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

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

[125] J. E. Rush and et al. 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, 1971.

[126] Horacio Saggion and Guy Lapalme. Generating Indicative-Informative Summaries with SumUM. *Computational Linguistics*, 28(4), 2002.

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

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

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

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

[131] J.D. Schlesinger, J.M. Conroy, M.E. Okurowski, H.T. Wilson, D.P. O'Leary, A. Taylor, and J. Hobbs. 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 2002.

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

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

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

[135] H. Sundaram. *Segmentation, Structure Detection and Summarization of Multimedia Sequences*. PhD thesis, Graduate School of Arts and Sciences, Columbia University, 2002.

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

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

[138] S. L. Taylor. *Automatic abstracting by applying graphical techniques to semantic networks*. PhD thesis, Northwestern University, 1975.

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

[140] Simone Teufel and Marc Moens. Sentence extraction and rhetorical classification for flexible abstracts. In *AAAI Spring Symposium on Intelligent Text Summarisation*, pages 16 – 25, 1998.

[141] Simone Teufel and Marc Moens. Summarizing scientific articles – experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4), 2002. Special Issue on Automatic Summarization.

[142] Richard Tucker. *Automatic Summarising and the CLASP system.* PhD thesis, University of Cambridge, 1999.

[143] H. van Halteren. 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 2002.

[144] Piek Vossen, editor. *Euro WordNet: a multilingual database with lexical semantic networks.* Kluwer Academic Publishers, 1998.

[145] H. Wactlar. Multi-document summarization and visualization in the informedia digital video library, 2001.

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

[147] M. Witbrock and V. Mittal. 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)*, 1999.

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

[149] D. Zajic, B. Door, and R. Schwartz. 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 2002.

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

[151] Klaus Zechner. *Automatic Summarisation of Spoken Dialogues in Unrestricted Domains.* PhD thesis, Carnegie Mellon University, 2001.

System	Processing Level	Information Kind	Tucker 1999
Adam [125, 114]	surface	structural	sentencewise
Alfonseca and Rodríguez [1]	surface	structural	sentencewise
* Anes [23]	surface	lexical	att. networks
Barzilay and Elhadad 1997 [15]	entity	lexical	att. networks
Boguraev and Kennedy 1997 [21]	entity	lexical	att. networks
Caldwell 1994 [26]	entity	lexical	att. networks
* CENTRIFUSER [43]	discourse	understanding	info. content
* Chen and Lin (2000) [31]	surface	lexical	info. content
* Columbia MDS [98, 34, 107]	entity/discourse	understanding/structural	info. content
Copeck et al. 2002 [34]	surface	lexical	att. networks
* Cut-and-Paste [63]	surface	structural	info. content
Darsy 1996 [39]	entity	lexical	att. networks
* DiaSumm [151]	surface	lexical	discourse structure
DimSum [9]	surface	lexical	att. networks
* DMSumm [112]	discourse	structural	disc. structure
Edmunson 1969 [42]	surface	structural	sentencewise
FilText [102]	surface	structural	info. content
* FociSum [67]	entity	understanding	att. networks
Frump [38]	entity	understanding	info. content
GISTEXTER [53, 73]	discourse/entity	understanding	info. content
GISTSumm [113]	surface	lexical	att. networks
Gladwin et al. 1991 [49]	entity	lexical	att. networks
* GLEANS [37]	entity/discourse	understanding	info. content
* NTT [58, 59]	surface	structural/lexical	att. networks
* Karamuftuoglu 2002 [68]	surface	structural	att. networks
* Kraaij et al. 2002 [71]	surface	lexical	att. networks
K. U. Leuven [6, 7]	entity	lexical	att. networks
* Lal and Rueger 2002 [74]	entity/discourse	understanding	info. content
Lehnert 1982 [77]	entity	understanding	info. content
* Univ. of Lethbridge [24, 28]	entity	structural/lexical	att. networks
Luhn 1958 [86]	surface	lexical	att. networks
Marcu 1997 [93]	discourse	structural	disc. structure
* MEAD [116, 117]	surface	lexical	att. networks
* MultiGen [99, 17]	entity	structural	info. content
* NeATS [82, 83, 78]	entity	structural	info. content
* Newsblaster [96]	entity/discourse	structural/understanding	info. content
NewsInEssence [119]	surface	lexical	att. networks
Ono et al. 1994 [108]	discourse	structural	disc. structure
NetSumm [115]	surface	lexical	att. networks
Paice 1981 [110]	surface	structural	sentencewise
* PERSIVAL [97]		understanding	info. content
Rafi [75]	surface	structural	att. networks
* RIPTIDES [124, 146]	entity/discourse	understanding	info. content
SAM [129, 36]	entity	understanding	info. content
Dunlavy et al. 2003 [131, 41]	surface	lexical	att. networks
Scisor [123]	entity	understanding	info. content
Scrabble [137]	entity	understanding	info. content
Skorochod'ko 1971 [132]	entity	lexical	att. networks
Smart [127, 103]	entity	lexical	att. networks
* SUMMARIST [61]	surface	lexical	att. networks
SUMMONS [101]	entity	understanding	info. content
SumUM [45, 126, 46]	discourse	structural	discourse structure
* SweSum [136]	surface	lexical	att. networks
Taylor 1975 [138]	entity	understanding	info. content
Tele-Pattan [18]	entity	lexical	att. networks
Tess [148]	entity	understanding	info. content
Teufel and Moens [140, 141]	discourse	structural	disc. structure
TICC [2]	entity	understanding	info. content
TOPIC [51]	discourse	structural	disc. structure
van Halteren 2002 [143]	surface	lexical	att. networks
WebInEssence [120, 149]	surface	lexical	att. networks

Table 1: Classification of summarization systems

On-line or Downloadable 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) http://www.copernic.com/desktop/products/summarizer/download.html
DMSumm downoadable demo	English, Brazilian Portuguese single document http://www.nilc.icmc.usp.br/ thiago/DMSumm.zip
Extractor downloadable demo	English, French, Spanish, German, Japanese, Korean single document (many formats) http://www.dbi-tech.com/dbi_extractor.asp
GISTexter no straightforward access	English Single and Multi-Document form at: http://www.languagecomputer.com/demos/summarization/index.html
GistSumm downloadable demo	multilingual single document http://www.nilc.icmc.usp.br/ thiago/Install_GistSum.zip
Newsblaster on-line demo	Multilingual multi-document http://www1.cs.columbia.edu/nlp/newsblaster/
Island InText no straightforward downloading	English single document form at: http://www.islandsoft.com/orderform.html
Inxight Summarizer / LinguistX / Xerox PARC no straightforward downloading	Chinese, Danish, Dutch, English, Finnish, French, German, Italian, Japanese, Korean, Norwegian, Portuguese, Spanish and Swedish single document form at: http://www.inxight.com/products/oem/summarizer/contact_sales.php
Kmaritime on-line demo	Korean http://nlplab.kmaritime.ac.kr/demo//f_ats.html
Lal and Rüger (2002) on-line demo	English single document http://rowan.doc.ic.ac.uk:8180/summarizer/demo.html
MEAD / NewsInEssence / CLAIR on-line and dowloadable demo	English and Chinese multi-document, multi-lingual http://www.clsp.jhu.edu/ws2001/groups/asmd/ multiple news summ. demo at: 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
Sinope Summarizer Personal Edition 30-day trial downloadable	English, Dutch and 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 30-day trial downloadable demo	probably English only single web pages (Mac OS X.1 only) 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) http://www.nada.kth.se/ xmartin/swesum/index-eng.html
TextWise Content Repurposing Suite no straightforward access	probably English only single document or e-mail form at: http://www.textwise.com/technology/crs/demo.html

Table 2: Some on-line demos of summarization systems, both commercial and academic

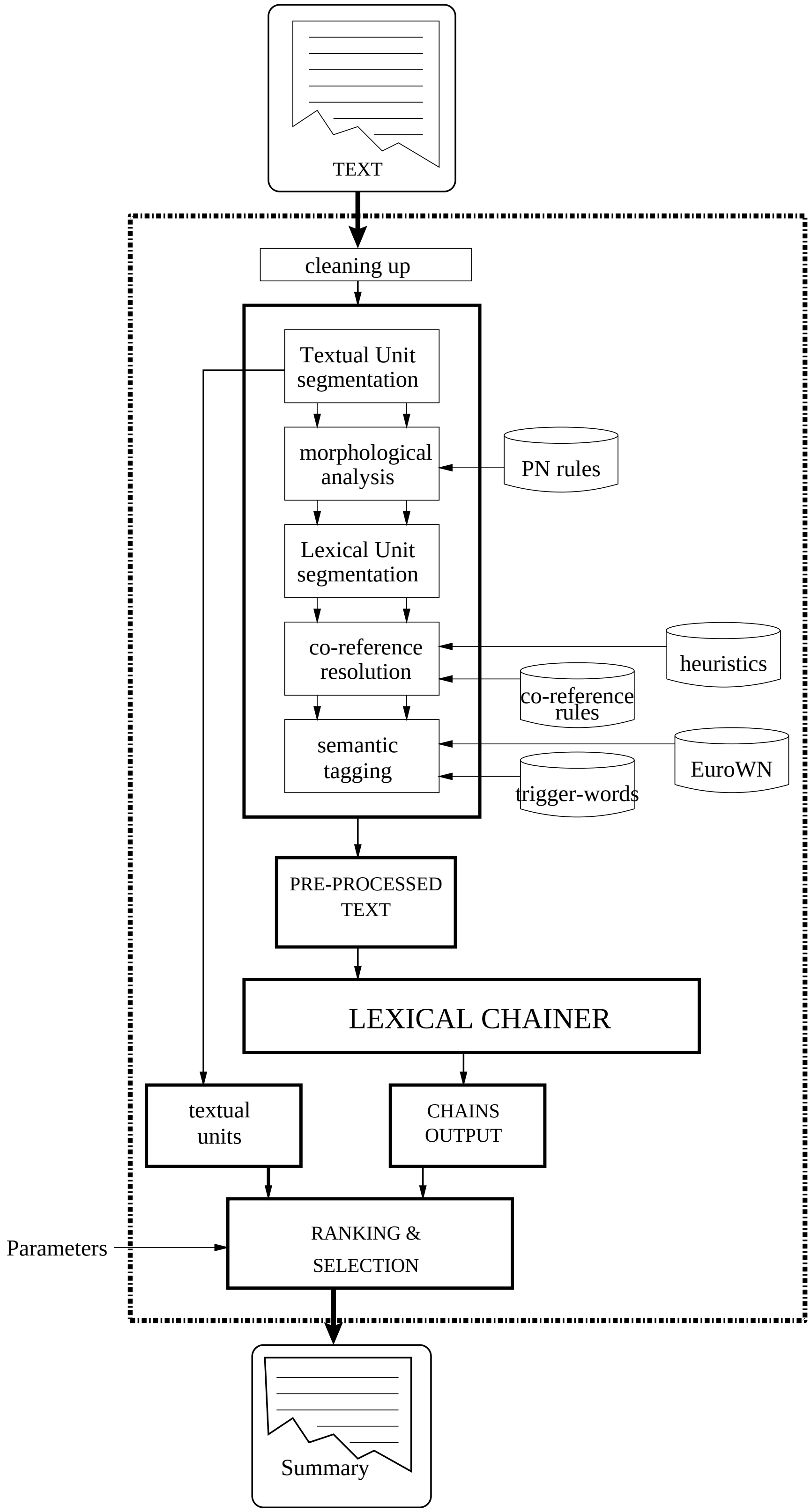

Figure 1: Architecture of HERMES Lexical Chain Summarizer.

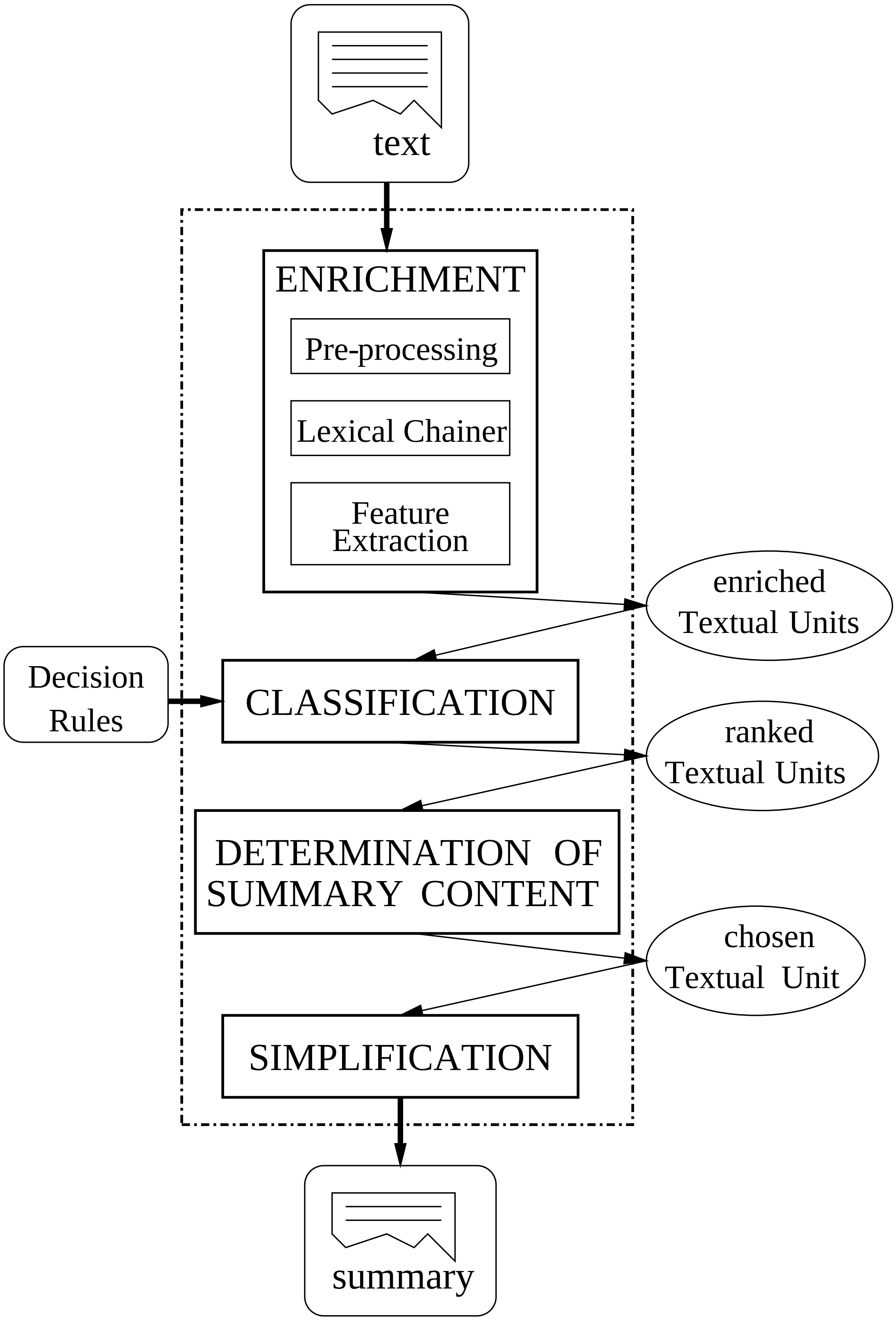

Figure 2: Architecture of HERMES Machine Learning Summarizer.