*Natural Language Engineering* **19** (2): 147–186. © Cambridge University Press 2012
doi:10.1017/S1351324912000198

# *COMPENDIUM: a text summarisation tool for generating summaries of multiple purposes, domains, and genres*

ELENA LLORET and MANUEL PALOMAR

*Department of Software and Computing Systems, University of Alicante, Apdo. de correos, 99, E-03080, Alicante, Spain*
*e-mail*: {elloret,mpalomar}@dlsi.ua.es

(*Received 26 September 2011; revised 14 June 2012; accepted 18 June 2012; first published online 16 July 2012*)

## Abstract

In this paper, we present a Text Summarisation tool, COMPENDIUM, capable of generating the most common types of summaries. Regarding the input, single- and multi-document summaries can be produced; as the output, the summaries can be extractive or abstractive-oriented; and finally, concerning their purpose, the summaries can be generic, query-focused, or sentiment-based. The proposed architecture for COMPENDIUM is divided in various stages, making a distinction between core and additional stages. The former constitute the backbone of the tool and are common for the generation of any type of summary, whereas the latter are used for enhancing the capabilities of the tool. The main contributions of COMPENDIUM with respect to the state-of-the-art summarisation systems are that (i) it specifically deals with the problem of redundancy, by means of textual entailment; (ii) it combines statistical and cognitive-based techniques for determining relevant content; and (iii) it proposes an abstractive-oriented approach for facing the challenge of abstractive summarisation. The evaluation performed in different domains and textual genres, comprising traditional texts, as well as texts extracted from the Web 2.0, shows that COMPENDIUM is very competitive and appropriate to be used as a tool for generating summaries.

## 1 Introduction

In the current society, information plays a crucial role that brings competitive advantages to users, when it is managed correctly. However, due to the vast amount of available information, users cannot cope with it, and therefore new methods and approaches based on Natural Language Processing (NLP) are essential to process all the information in an effective and efficient manner. Applications such as Information Retrieval (IR), Question Answering (QA), or Text Summarisation (TS) can help users to *access the information* more easily, on the one hand, reducing the time they have to spend dealing with the information, and on the other, selecting the information most useful for them. However, issues such as the nature of the different sources, together with the fact that the same information can be repeated across different documents (redundancy), are great challenges for the above-mentioned applications. In particular, the aim of TS is to process, synthesise, and present the information to users, avoiding the arduous task of having to read everything, as well as facilitating the process of guiding the user to what is important in texts.

This paper is focused on TS, and its main goal is to present a TS tool, COMPENDIUM, which is able to produce different types of summaries from different domains and textual genres. For specifying the types of summaries, we follow the taxonomy suggested in Spärck Jones (2009). In this way, concerning the *input*, COMPENDIUM can either take one or several texts, and produce single- or multi-document summaries. Regarding the *purpose* of the resulting summaries, these can be generic, query-focused, or sentiment-based, and their aim is to provide information about the source document(s), thus being informative. As *output*, the final summaries can be extracts or abstractive-oriented summaries (i.e. a combination of extractive and abstractive information). Finally, it is important to mention that COMPENDIUM is a mono-lingual TS tool, working only for one language, i.e. English.

As far as the architecture of COMPENDIUM is concerned, an architecture based on specific stages is proposed. In particular, COMPENDIUM relies on five core stages that constitute the backbone of the TS process (*surface linguistic analysis*; *redundancy detection*; *topic identification*; *relevance detection*; and *summary generation*) together with a series of additional stages that can be integrated into the core ones, thus enhancing the capabilities of COMPENDIUM, generating also query-focused, sentiment-based, or abstractive-oriented summaries. Specifically, the additional stages are *query similarity*; *subjective information detection*; and *information compression and fusion*. With respect to its architecture, the main contributions of COMPENDIUM are (i) the use of textual entailment for avoiding redundant information in the summaries; (ii) the combination of statistical and cognitive-based techniques for detecting relevant information; and (iii) the generation of abstractive-oriented summaries.

In order to assess that COMPENDIUM is appropriate for a wide range of domains and textual genres, we conduct an intrinsic evaluation using different types of documents. The results obtained show that COMPENDIUM performs very well, and therefore it is a suitable TS tool for generating summaries of multiple purposes, domains, and textual genres.

The structure of the paper is as follows. Section 2 puts in context the summarisation task by explaining the types of existing summaries (Section 2.1), as well as the common approaches and techniques for TS (Section 2.2). Also, it remarks the novelties of our TS tool with respect to the existing ones. Section 3 describes our proposed TS tool: COMPENDIUM. In this section, we first provide an overview of its main characteristics and architecture (Section 3.1), and then we focus on the stages it comprises (Sections 3.2 and 3.3, for the core and additional stages, respectively).

The evaluation of COMPENDIUM is provided in Section 4. Section 4.1 contains the description of the different corpora used, whereas Section 4.2 shows the experiments and the results obtained in a wide range of domains and types of texts, comprising newswire, image captions, blogs, and medical research papers. Finally, the main conclusions together with future research lines are explained in Section 5.

## 2 Related work

Although it started in the late 1950s, TS has experienced a great development in recent years, and a wide range of techniques and systems have been proposed in this research field (Lloret and Palomar 2011b). However, to produce a summary automatically is very challenging. Issues such as redundancy, temporal dimension, coreference, or sentence ordering, to name a few, have to be taken into consideration, especially when summarising a set of documents (multi-document summarisation), thus making this field even more difficult (Chali and Hasan n.d.). Moreover, research attempting to overcome the lack of coherence that summaries often present has been fuelled in the last years, resulting in combined approaches that identify relevant content and merge it into new fragments of information (Barzilay and McKeown 2005; Zajic, Dorr and Lin 2008). In order to adapt to new society requirements, the types of summaries have increased in recent years. For instance, the Web 2.0 (social web) has led to the emergence of new types of websites, such as blogs, forums, or social networks, where anybody can express his/her opinions towards a topic, entity, product, or service. This has resulted in a new type of summaries with the purpose of summarising users' opinions (sentiment-based summaries). Therefore, when carrying out research into this area, it is essential to be aware of previous TS approaches and systems so that new or improved methods can be suggested in order to tackle the different types of summaries and their requirements.

In this section, we provide a general overview of the TS task. In this manner, we first explain the most important types of summaries (Section 2.1); and then we focus on the process of TS from a computational point of view, explaining the most common approaches for TS, as well as outlining the differences with respect to our proposed tool (Section 2.2).

### *2.1 Common factors for classifying summaries*

A wide range of summaries can be generated depending on different factors, such as the type of input/output, the purpose of the summary, or the type of reader, which makes the TS especially challenging in NLP. In the literature, three main taxonomies can be found (Hovy and Lin 1999; Mani and Maybury 1999; Sparck Jones 1999) that group summaries from different perspectives.

Taking as a basis the aforementioned taxonomies and the TS approaches, we can depict the most common summary types in Figure 1.

As it can be seen in the figure, although it has traditionally focused on text (Yu *et al.* 2007), the input to the summarisation process can also be multimedia information, such as images (Fan *et al.* 2008); video (Dumont and Mérialdo 2009), or audio (Liu and Liu 2009), as well as on-line information or hypertexts (Steinberger, Jezek and Sloup 2008; Tigelaar, Op Den Akker and Hiemstra 2010). Furthermore, we can talk about summarising only one document (*single-document summarisation*) (Svore, Vanderwende and Burges 2007) or multiple ones (*multi-document summarisation*) (Haghighi and Vanderwende 2009). Regarding the output, a summary may be an *extract* (i.e. when a selection of "significant" sentences of a document is shown) (Zhang and Fung 2009); *abstract*, when new vocabulary

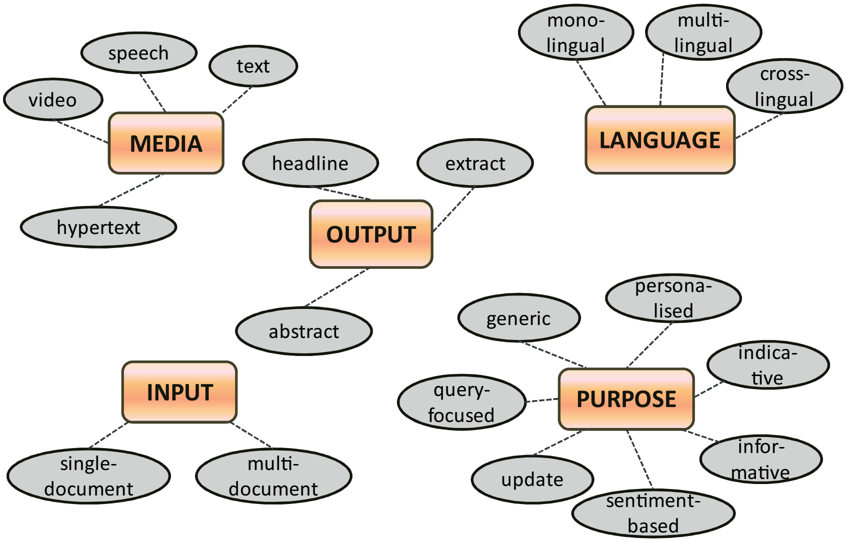

Fig. 1. Summarisation types according to the most common factors.

is added (Ou, Khoo and Goh 2007), or even a *headline* (or title) (Sarkar and Bandyopadhyay 2005; Hennig 2009). It is also possible to distinguish between *generic* summaries and *query-focused* summaries (also known as user-focused or topic-focused). Examples of approaches that generate these types of summaries can be found in Schilder and Kondadadi (2008) or Zhao, Wu and Huang (2009). The first type of summaries can serve as a surrogate of the original text, as these may try to represent all relevant facts of the source text. In the latter, the content of a summary is biased towards a user need, query, or topic. Concerning the style of the output, a broad distinction is normally made between two types of summaries. *Indicative* summaries are used to indicate what topics are addressed in the source text. As a result, these can give a brief idea of what the original text is about. The other type, *informative* summaries, are intended to cover the topics in the source text and provide more detailed information. It is also possible to combine both types, as in the approach suggested in Saggion and Lapalme (2002). In recent years, new types of summaries have appeared. For instance, the birth of the Web 2.0 has encouraged the emergence of new types of textual genres, containing a high degree of subjectivity, thus allowing the generation of *sentiment-based summaries*. A lot of approaches have been proposed for tackling these types of summaries (Balahur *et al.* 2009; Bossard, Généreux and Poibeau 2009; Lerman, Blair-Goldensohn and McDonald 2009). Furthermore, *update summaries* are other examples of new summary types. These assume that users already have a background and they only need the most recent information about a specific topic (Nastase, Filippova and Ponzetto 2008). Finally, concerning the language of the summary, it can be distinguished between *mono-lingual*, *multi-lingual*, and *cross-lingual* summaries, depending on the number of languages dealt with. The cases where the input and the output languages are the same lead to mono-lingual summaries, as in El-haj and Hammo (2008). However, if different languages are involved, the summarisation approach is considered multilingual or cross-lingual (Kabadjov *et al.* 2010; Lehmam 2010; Litvak, Last and Friedman 2010).

### 2.2 The process of summarisation from a computational perspective

The types of summaries mentioned previously have to be created following a summarisation process, thus allowing to transform the source document or documents into a summary. According to the summarization chapter in (Mitkov, 2003), the following three stages have to be taken into account for producing a summary from a computational point of view:

- *Topic identification.* It consists of determining the subject of the document. It is usually approached by assigning each unit (words, sentences, phrases, etc.) a score which is indicative of its importance. In the end, the top score units, up to a desired, length are extracted.
- *Interpretation or topic fusion.* During this stage, the topics identified as important are fused, represented in new terms, and expressed using a new formulation, which includes concepts or words not found in the original text. This stage is what distinguishes extractive from abstractive summarisation, because in the former, the interpretation stage is almost inexistent in most of the approaches, as they only select and extract important sentences for building the final summary. In contrast, for abstractive approaches, this stage is crucial, since the relevant information is generalised and merged.
- *Summary generation.* In this stage, the final summary is built. When dealing with abstractive summarisation, natural language generation techniques are generally included as part of this process. In the case of extractive summarisation, in this stage, it is important to pay attention to the ordering of sentences in the summary, and the coherence between them.

However, most of the systems focus on extractive summarisation, and as a consequence they only take into account the first stage of the process.

Next, we are going to explain the most common techniques that are employed for generating summaries, distinguishing between statistical-based, topic-based, graph-based, discourse-based, and machine-learning-based approaches. Within each of these groups, existing TS systems are going to be described as well.

- *Statistical-based approaches.*

  These approaches determine the relevance of a sentence based on statistical methods, such as term frequency (TF), or *tf*idf*. Whereas term frequency is solely based on how many times a word appears in a document, the idea behind *tf*idf* is that frequent terms in a document are important only if they are not very frequent in the whole collection.

  In McCargar (2005), several statistical approaches were analysed, as well as the potential problems that these types of features may have. It was claimed in Filatova and Hatzivassiloglou (2004) that these methods might not be sufficient for building high-quality summaries; however, a deeper review of statistical techniques conducted in Orăsan, Pekar and Hasler (2004) and Orăsan (2009) showed that these techniques are appropriate for building good summaries. In addition to term frequency and *tf*idf*, mutual information, information gain, and residual inverse document frequency were also analysed. Mutual information is used to measure the dependency or the common information between two words, whereas information gain is a good metric for deciding the relevance of an attribute, and in this case, it could be perfectly applied to the terms or sentences in a document. Residual inverse document frequency is a variant of the inverse document frequency, which computes the term frequency according to a probabilistic distribution. The approach suggested in Mori (2002) also employs information gain for determining the weight of document terms, and then use it for summarising documents successfully. The idea is to first build clusters of documents according to the similarity among them, and then compute the weight of each word in the clusters. The final summary is produced by selecting the highest scored sentences on the basis of the weight of their words, previously computed using information gain.

- *Topic-based approaches.*

  These approaches rely on the identification of key words (topics) for establishing the relevance of the sentence. This is the case of the topic signatures suggested in Lin and Hovy (2000), where it is assumed that the topic of a document can be represented using a set of terms. Following this idea, in Harabagiu and Lacatusu (2005), it was analysed how the structure of a document is characterised in terms of topics' themes, which are representations of events that are reiterated throughout the document collection, and therefore represent repetitive information. Furthermore, in Teng *et al.* (2008), a single-document summarisation approach is suggested, which combines local topic identification with term frequency. The proposed methodology first computes the sentence similarity, and then performs the topic identification by doing sentence clustering. In a second step, sentences from local topics are selected according to the term frequency value. Moreover, not only topic words are used to detect relevant information within a document, in the approach suggested in Kuo and Chen (2008), informativeness and event words are also taken into consideration for producing multi-document summaries. The underlying idea is that these types of words indicate important concepts and relationships, and can be used to detect relevant sentences within a set of documents.

  As it can be seen, each technique establishes a manner of assigning weights to the words included in the document, and then sentences are scored based on these weights in order to determine their relevance.

- *Graph-based approaches.*

  The use of graph-based ranking algorithms has also been shown to be effective in TS. Basically, the nodes of the graph represent text elements (i.e. normally words or sentences), whereas edges are links between those text elements, previously defined (for instance, semantic relations, such as synonymy). On the basis of the text representation as a graph, the idea is that the topology of the graph will reveal interesting things about the salient elements of the text, for example concerning the connectivity of different elements. LexRank (Erkan and Radev 2004) is an algorithm used in the MEAD system (Radev *et al.* 2004), in which all candidate sentences that can be potentially included in the summary are represented in a graph. In this graph representation, two sentences are connected if the similarity between them is above a predefined threshold. Then, once the network is built, the system finds the most central sentences by performing a random walk on the graph. In Mihalcea (2004), an analysis of several graph-based algorithms is carried out, evaluating also their application to automatic sentence extraction in the context of TS.

The SummGraph system (Plaza 2011) uses concepts identified with Wordnet (Fellbaum 1998) and *is-a* relationships for building a graph representation of each sentence in a document. This method has been proved to work successfully in the newswire, biomedical, and tourist domain.

- *Discourse-based approaches.*

  The Rhetorical Structure Theory (RST), proposed in Mann and Thompson (1988), served as a basis for the summarisation approach developed in Marcu (1999). In this approach, the rhetorical relations are extended, and this kind of discourse representation (nucleus and satellite relations, depending on how relevant the information is) is used to determine the most important textual units in a document. In Khan, Khan and Mahmood (2005) the RST is combined with a generic summariser in order to add linguistic knowledge to the summarisation process. Although the results obtained for this mixed approach did not improve the ones obtained by the generic summariser, it was claimed that the drawback of this approach was mainly due to parser, which could not detect all the RST relationships, otherwise linguistic knowledge could have improved the overall summarisation performance. Furthermore, in Cristea, Postolache and Pistol (2005) an approach similar to the RST is described, differing from the previous ones in the lack of relation names and the use of binary trees. This summarisation approach is intended to exploit the coherence and cohesion of a document.

  Cohesion and coherence are two of the main challenging issues for TS. Some approaches rely on the identification of such relations in order to improve the quality of the generated summaries. In Gonçalves, Rino and Vieira (2008), coreference chains are used to deal with referential cohesion problems that are frequent in the extractive summarisation approach. A post-processing system is developed in order to rewrite referential expressions in the most possible coherent way, and it is applied after the summary is generated, obtaining considerable improvements in comparison with the original summaries. In order to guarantee the coherence of a summary, a widespread approach is to use lexical or coreference chains. However, the use of coreference chains is not novel in TS. The first approaches can be found in Baldwin and Morton (1998) and Azzam, Humphreys and Gaizauskas (1999). The main assumption is that the longest coreference chain indicates the main topic of the document, and shorter chains represent subtopics. Therefore, one possible strategy for building summaries is to select only those sentences which related to the longest chain. This strategy helps to maintain the coherence of the text. A similar idea is to use *lexical chains*, which consist in determining sequences of semantic related words (e.g. by concept repetition or synonymy relations). By using lexical chains, the main topics of a document can also be detected. This technique has been widely used in summarisation, and the approaches like the ones described in Barzilay and Elhadad (1999), (Medelyan 2007), or Ercan and Cicekli (2008) exploit them to produce summaries.

- *Machine-learning-based approaches.*
  The high amount of techniques that can be used in TS led to the necessity of combining all of them in an optimal manner in order to come up with the sentence relevance. In regard to this, a wide range of machine-learning techniques can be used for TS. For instance, NetSum (Svore *et al.* 2007) bets on single-document summarisation and produces extracts from newswire documents based on neuronal nets, using RankNet (Burges *et al.* 2005) as a learning algorithm to score the sentences and extract the most important ones. Besides the common features based on keywords and sentence position, a new set of features based on Wikipedia[1] and query logs are also used in a way that, for example, sentences containing query terms or Wikipedia entities are indicative of important content. In Schilder and Kondadadi (2008), a query-focused multi-document summariser is presented, named as FastSum, where sentences are ranked using a machine-learning technique called *Support Vector Regression* (SVR), and *Least Angle Regression* for feature selection. SVR was used in summarisation before, in the approach described in Li *et al.* (2007), where word-, phrase-, or semantic-based features, as well as sentence position or name entities were used to train the classifier automatically. Further on, the extracted features were combined, and then sentences were scored. In Wong, Wu and Li (2008), an extractive summarisation approach is presented, employing supervised and semi-supervised learning methods. The features involved are grouped into different types – surface, content, relevance, and event features – which include sentence position, number of words in a sentence, centroid and high frequent terms, or similarity between sentences, among others. Regarding the supervised approach, a *Support Vector Machine* (SVM) algorithm is used, whereas for the semi-supervised approach, a probabilistic SVM and a *Naïve Bayesian* classifier are co-trained to exploit unlabelled data. SVM technique was also used in Fuentes, Alfonseca and Rodríguez (2007) to detect relevant information to be included in a query-focused summary, where structural, cohesion-based, and query-dependent features were used for training.

COMPENDIUM, our proposed TS tool, is not based on machine-learning techniques. It partially relies on statistical techniques, and it differs from the already existing TS

[1] http://www.wikipedia.org/

sytems in three aspects, mainly as follows: (i) It specifically address the redundancy problem analysing a novel method for such purpose (i.e. textual entailment); (ii) sentence relevance is determined taking into account cognitive-based techniques in addition to statistical ones; (iii) it faces the challenge of abstractive summarisation by proposing an hybrid approach that produces abstractive-oriented summaries combining extractive and abstractive techniques.

In the next section, the architecture of COMPENDIUM is explained.

## 3 Architecture of COMPENDIUM

In this section, COMPENDIUM TS tool is described in detail. The aim of COMPENDIUM is to produce different types of summaries automatically. Therefore, in Section 3.1 we first explain the main characteristics of the summaries generated, as well as an overview of its proposed architecture. Then we go into detail and explain its stages, distinguishing between a set of core ones that are the most important in COMPENDIUM (Section 3.2), and the additional stages (Section 3.3), which are used to generate summaries for multiple purposes (query-focused, sentiment-based, and abstractive-oriented summaries).

### *3.1 COMPENDIUM's overview*

Taking the schema depicted in Figure 1 as a basis, the summaries generated with COMPENDIUM according to the proposed factors (media, input, output, purpose, and language) can be characterised. Next, each of these factors are explained in the context of COMPENDIUM.

- **Media**. Concerning this factor, at the moment, COMPENDIUM deals only with text documents.
- **Input**. Regarding this factor, COMPENDIUM takes one or several texts, thus being able to produce single- and multi-document summaries.
- **Output**. The types of summaries generated can be either extracts (i.e. the most important sentences are selected) or abstractive-oriented summaries (in our case, information is compressed and fused to generate new sentences different from the original ones, and then these sentences are combined with the previously extracted ones).
- **Purpose**. The summaries generated with COMPENDIUM aim at being substitutes for the original documents. Therefore, they must contain the most relevant facts, thus being informative. In addition, specific purposes according to user's interests are taken into account. This leads to generic, query-focused, and sentiment-based summaries. Generic summaries provide a general overview of the document; query-focused summaries are biased towards a user need, question, or topic; and finally, sentiment-based summaries contain a high degree of subjective information, reflecting the opinions of users about a topic.
- **Language**. COMPENDIUM is a mono-lingual TS tool, which work with English, both for the input and the output.

In addition to all these factors, it is worth stressing upon the fact that the summaries generated by COMPENDIUM are not of a fixed length. In contrast, their length can be configured depending on the interest of the user (either in a compression ratio or in number of words), thus being a factor that could be included within the output or purpose characteristics of the TS tool.

Concerning the stages of COMPENDIUM, five core stages that constitute the backbone of the TS process are distinguished. These stages are as follows:

- **Surface linguistic analysis**, which pre-processes the input text.
- **Redundancy detection**, which identifies and removes repeated information.
- **Topic identification**, which determines the main topics of the document/s to be summarised.
- **Relevance detection**, which identifies the most relevant sentences of the document/s.
- **Summary generation**, which extracts the most relevant sentences, and presents them maintaining the same order as in the source document.

It is worth noting that the order of the proposed stages is not arbitrary. We rely on Van Dijk's theory (Van Dijk 1980) concerning the macrostructure of a text. The macrostructure represents the global meaning of discourse, and it is the result of the reading and comprehension process performed by humans. Therefore, this is related to the summarisation process, in the sense that a summary *is the explicit representation of the macrostructure of a text* (i.e. its overall meaning) (Álvarez Angulo 2002). Moreover, the macrostructure of a text can be obtained through the application of a set of macrorules, which can have both a reductive nature (to remove unnecessary information), and a constructive nature (allowing certain elements to be combined in new and more complex units of information). Specifically, the macrorules are *deletion*, *strong deletion*, *generalisation*, and *construction.* The former for deleting redundant information and trivial details, and the latter for generating new information. As it can be seen, the proposed stages in COMPENDIUM are directly related to these macrorules.

Furthermore, we suggest three additional stages in order to increase the capabilities of COMPENDIUM, allowing it to generate query-focused, sentiment-based, or abstractive-oriented summaries. Specific stages for achieving each of these types of summaries are as follows:

- **Query similarity**, needed for generating query-focused summaries.
- **Subjective information detection**, necessary for identifying subjective information and producing sentiment-based summaries.
- **Information compression and fusion**, crucial for generating new information that will appear in the summary, thus resulting in an abstractive-oriented summary, rather than an extract.

Figure 2 depicts the general architecture of COMPENDIUM.[2] In this figure, the core stages are represented within a big rectangle with rounded borders, whereas dotted

[2] COMPENDIUM demo is available at: http://intime.dlsi.ua.es:8080/compendium/

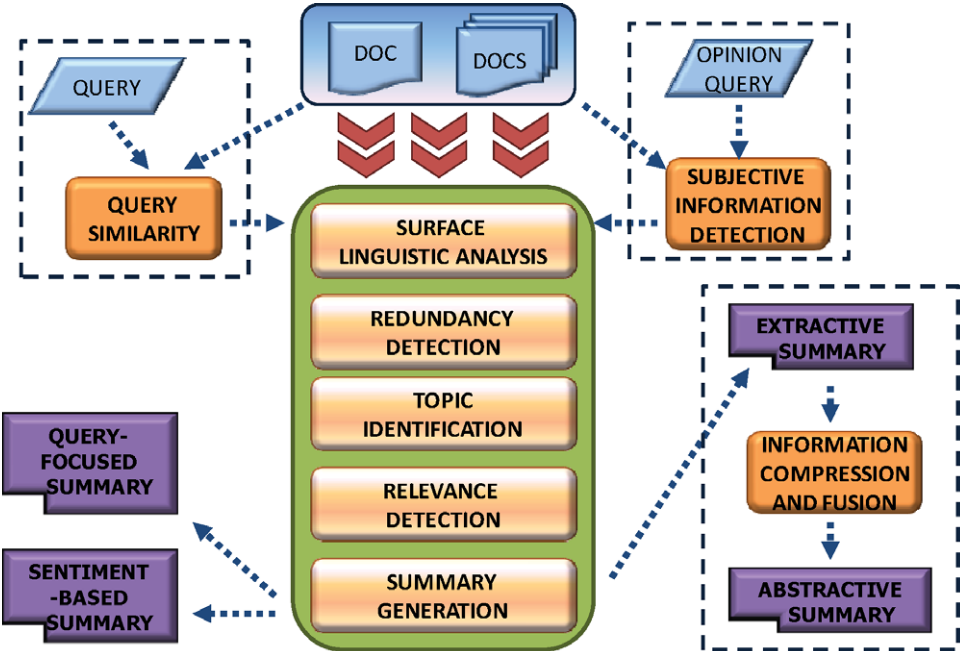

Fig. 2. Overview of COMPENDIUM's architecture.

rectangles correspond to the additional stages. By applying only the core stages, the resulting summaries are single- or multi-document generic informative extracts. In contrast, by taking into consideration the additional stages, query-focused or sentiment-based extracts, as well as abstractive-oriented summaries from a single or several documents can be generated. The remaining of this section focuses on the explanation of the core and additional stages.

## *3.2 Core stages*

As aforementioned, the core stages are (a) *surface linguistic analysis*; (b) *redundancy detection*; (c) *topic identification*; (d) *relevance detection*; and (e) *summary generation*. Next, we explain each of the stages in detail.

### *3.2.1 Surface linguistic analysis*

This stage aims at carrying out a basic linguistic analysis on the input document, thus preparing it for further processing. In order to carry out this analysis, external state-of-the-art tools and resources are used. In particular, for this stage we propose the following:

- **Sentence segmentation.** The text is split into sentences, which are the textual units considered for generating the summary. For this purpose, the sentence segmentation tool provided at DUC evaluation campaigns[3] is used.

[3] http://duc.nist.gov/duc2004/software/duc2003.breakSent.tar.gz

- **Tokenisation.** A tokeniser allows us to identify each word in the document, since we will need to compute, for instance, the frequency of each word, or distinguish between stop words and non-stop words in later stages. In order to be able to identify each word of the text, a tokeniser is used. In particular, we employ *Word Splitter*.[4]
- **Part-of-speech tagging.** A part-of-speech tagger assigns each word with its corresponding morphological category (noun, verb, adjective, preposition, adverb, determiner, pronoun, and conjunction). This process is useful for distinguishing different types of words, since some of them (e.g. nouns or verbs) can be more important than others (e.g. determiners). This tool will be used in the additional stages of *information compression and fusion*. In particular, TreeTagger[5] was used as a part-of-speech tagger, because it is very easy to use, and it can be used for different languages, not only for English.
- **Stemming.** This process consists in reducing words to their stem form. It is very useful in the cases where there is no need to differentiate between two inflected words that belong to the same family (e.g. *running* and *runs*, both come from the verb *run*). The *Porter Stemmer*[6] is employed for performing this task, which will be necessary for considering all terms sharing a common stem as a single one, for later computing their frequency.
- **Stop word identification.** Stop words are words that appear very frequently in documents, but they do not carry any semantic information. They are normally not used for further processing, because they are not relevant (e.g. articles: *the*, *a*; conjunctions: *and*, *or*, etc.). In our case, this process is essential, and consequently we tag them in order not to take them into consideration when performing the remaining stages of the TS process. For instance, when computing the frequency of a word for determining the topic of a document (please see the *topic identification* stage), words such as *"and"*, *"is"*, or *"the"* could be wrongly identified as document's topics if they are not identified as stop words and treated as normal terms. For carrying out this task, a list of stop words is needed. In particular, we use an English stop word list.[7]

We would like to note that this version of COMPENDIUM does not include any coreference resolution module within the *Surface Linguistic Analysis* stage. Although we think it is an essential part of the process and should be incorporated in the short-term future. However, this is not a trivial task. State-of-the-art results in coreference resolution, and more concretely, in anaphora resolution, are around 65% on average (Delmonte *et al.* 2006). Therefore, before including a step focusing on the coreference, we would like to analyse in depth, which anaphora or coreference resolution system would be the most appropriate to be used for our purposes.

[4] http://cogcomp.cs.illinois.edu/page/tools_view/8
[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[6] http://tartarus.org/~martin/PorterStemmer/
[7] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

### *3.2.2 Redundancy detection*

The aim of this stage is to identify redundant information in the source documents in order not to include it in the summary. For this purpose, we propose Textual Entailment (TE) as a technique to detect redundancy. On the one hand, attempts to study the influence of TE on TS have been focused for the evaluation of summaries (Harabagiu, Hickl and Lacatusu 2007) to determine which candidate summary, among a set of these, better represents the content in the original document depending on whether the summary entails it or not. On the other hand, TE has been combined with TS to generate a summary directly from the entailment relations found in a text (Tatar *et al.* 2008), or by extracting the highest scored sentences of a document, where the score of each sentence is computed as the number of sentences of the text that are entailed by it. However, in none of the already existing approaches, TE has been employed for identifying and removing repeated information, thus being novel in this sense.

Textual entailment is a suitable technique for detecting redundant information because its aim is to determine whether the meaning of a text snippet, also known as hypothesis (H), can be inferred from another one, called the text (T) (Glickman 2006). In order to illustrate this objective, we provide the following examples, taken from the RTE corpora.[8] As it can be seen, the first example shows a true entailment relation, whereas the second example shows a false entailment.

**Pair id = 50** (entailment = true)
*T: Edison decided to call "his" invention the Kinetoscope, combining the Greek root words "kineto"(movement), and "scopos" ("to view").*
*H: Edison invented the Kinetoscope.*
**Pair id = 18** (entailment = false)
*T: Gastrointestinal bleeding can happen as an adverse effect of non-steroidal anti-inflammatory drugs such as aspirin or ibuprofen.*
*H: Aspirin prevents gastrointestinal bleeding.*

Taking the rationale behind TE, the manner in which it is employed for detecting redundancy within our summarisation process is next explained. We can compute the entailment relationship between two sentences, discarding the second one, if they both contain a true entailment. This means that the meaning of the second sentence is already embedded in the first one, thus stating the information only once. Therefore, if we repeat the process with all the sentences in a document, those sentences whose meaning is already contained in other sentences can be discarded, as the information has been previously mentioned. As a result, by applying TE we can obtain a set of sentences from the text that do not hold an entailment relation with any other, and then keep this set of sentences for further processing.

It is worth stressing upon the fact that the order in which the entailment relationships are computed is the same order that the sentences have in the original documents. In this manner, we ensure that the coherence of the resulting summary is

[8] http://pascallin.ecs.soton.ac.uk/Challenges/RTE3/

not highly affected after this stage. Moreover, we only perform TE in one direction to avoid a very high computational cost.

Specifically in our research work, we used the TE approach presented in Ferrandez (2009) for computing the entailment relations within a document. This TE system relies on lexical (cosine similarity, Leveshtein distance), syntactic (dependency trees), and semantic measures based on WordNet (Fellbaum 1998), and although its performance is around 60%, it has been shown in previous research (Lloret *et al.* 2008a, 2008b) that this technique is appropriate for detecting redundant information when addressing summarisation.

### *3.2.3 Topic identification*

The objective of this stage is to identify the main topics of a document, so we can later use them to determine which sentences are more important. In Section 2.2 we reviewed the most common techniques for generating summaries, including the topic-based approaches. The term frequency calculation is one of the most employed techniques for achieving this goal, as it has been shown in previous work (Montiel Soto and García-Hernández 2009). Moreover, it has been proven that frequent terms are very likely to appear in human-written summaries (Nenkova, Vanderwende and McKeown 2006).

Following these ideas, in COMPENDIUM summariser, the topics of a document are represented by the frequency of the terms it contains. In this manner, we assume that the most frequent terms of a document will be indicative of the topics included in it, and therefore, as we have previously stated, we will reward the sentences containing the topics (i.e. frequent terms) in the *Relevance Detection* stage. However, it is important to note that stop words, which were previously identified in the *stop word identification* stage, are not taken into consideration for calculating the frequency of a document, thus not forming part of the topics of a document.

### *3.2.4 Relevance detection*

The relevance detection stage assigns a weight to each sentence, depending on how relevant it is within the text. Then we will be able to distinguish between the sentences that carry important information and those that do not.

For tackling the TS task, we wanted to study and analyse a discourse theory that has not been applied before for this purpose. Therefore, in this stage, we take into consideration a feature with a cognitive background, *The Code Quantity Principle* (CQP) (Givón 1990), which is related to how humans understand and retain the information they read, and it has been proven to hold true in written texts (Ji 2007). Specifically, this theory states that (1) a larger chunk of information is given a larger chunk of code; (2) the less predictable information, the more coding material; and (3) the more important information, the more coding material. This means that when an item provides a specific information within a text, it has to be assigned with a coding that it would be more or less stressed according to the relevance degree of such information within the text. In other words, the most important information within a text will contain more lexical elements, and therefore it will be expressed by a high number of units (for instance, syllables, words, or phrases) (Becker 2002). This is also in line with *The Code Quantity, Attention and Memory Principle* (Givón 1990), where it is stated that there is a proportional relation between the relevance of information and the amount of quantity through which it is coded, since the more salient and different coded information in a text, the more easily the reader's attention will be caught. As a result, readers will retain, keep, and retrieve this kind of information more efficiently.

On the basis of this, a coding element can range from characters to phrases. We decided to analyse noun-phrases, because a noun-phrase is the syntactic structure that allows more flexibility in the number of elements it can contain (pronouns, adjectives, or even relative clauses). Moreover, it is able to carry more or less information (words) according to what the writer wants to express (Becker 2002). Furthermore, previous research (Mittal *et al.* 1999) has shown that the average length of complex noun-phrases in summary sentences was more than twice as long than those in non-summary sentences. In addition, Lloret and Palomar (2009) carried out a preliminary study of the percentage of noun-phrases contained in both source documents and model summaries of a corpus of a newswire and another one of fairy tales. This analysis showed that words belonging to noun-phrases were predominant over other types of words in the documents as well as in the summaries, representing on average more than 70% of all content words (i.e. without taking stop words into account), and approximately 30% of the total words of documents. For instance, we take these two sentences as an example:

*$S_1$: The Spanish Academy of Motion Pictures Arts and Sciences presented an honorific award for the best actor.*
*$S_2$: The Academy presented an honorific award.*

In this case, $S_1$ contains more information than $S_2$. Although at first sight, the second sentence might be more appropriate for TS, since it reflects the same facts as that of the first one, but in a shorter manner the first one contains more details, and this would lead to more informative summaries, which is the purpose of our TS process.

Therefore, to consider noun-phrases as coding elements for our approach based on CQP seemed appropriate.

For computing the relevance of a sentence in COMPENDIUM, we assume that when an author writes a document, he or she writes it following the CQP principle, i.e. emphasizing the most important information by means of larger coding material. Therefore, we will analyse to what extent this principle is useful for detecting relevance and generate automatic summaries. As it was shown in Lloret and Palomar (2009, 2010), computing the relevance of a sentence focusing only on the length of the noun-phrases it contained should not be always appropriate since the terms contained may not be relevant. Therefore, we took into consideration the topics identified in the previous stage, and the relevance of each sentence was computed by combining CQP and the weight assigned to each term through the term frequency calculation, as it can be seen in Formula 1.

$$r_{s_i} = \frac{1}{\#NP_i} \sum_{w \in NP} |tf_w| \tag{1}$$

where:
$r_{s_i}$ = the relevance of sentence *i*,
$\#NPi$ = number of noun-phrases contained in sentence *i*,
$tf_w$ = frequency of word *w* that belongs to the sentence's noun-phrase.

In order to identify noun-phrases within a sentence, the *BaseNP Chunker*,[9] is employed. One important aspect to take into consideration is that the use of a chunker (as well as any other NLP-based tool) can introduce some error rate. This tool achieves recall and precision rates of roughly 93% for base noun-phrase chunks, and 88% for more complex chunks (Ramshaw and Marcus 1995), thus being suitable for our purposes.

### *3.2.5 Summary generation*

The objective of this stage is to generate a summary of a specific length. This length can be expressed in words or in the form of the compression rate (i.e. the percentage of information the summary contains with respect to the source document). However, the way this stage is carried out strongly depends on the summary type and purpose we want to produce. Consequently, four main types of summaries can be distinguished: (1) generic; (2) query-focused; (3) sentiment-based; and (4) generic abstractive-oriented. Type 1 is directly produced when the core stages of COMPENDIUM are applied, whereas for types 2, 3, and 4, the additional stages of *query similarity*, *subjective information detection*, and *information compression and fusion* are also required, respectively.

Next, we briefly explain each of these strategies for generating the final summary.

(1) **Generic summaries (COMPENDIUM$_E$).** Once the final score of a sentence is computed by means of the *relevance detection* stage, the most important sentences (i.e. the ones with the highest scores) are selected and extracted to form the final summary up to a desired length or compression rate. In this case, the final summary is a generic extract.
(2) **Query-focused summaries (COMPENDIUM$_{QE}$).** Having computed the two different weights for each sentence (its relevance – $r_{s_i}$–, and its similarity with regard to the query – $qSim_{s_i}$ –, in the *relevance detection* and *query similarity* stages,[10] these two values are combined within the same formula (Formula 2), which is based on the F-measure calculation. As it can be deduced, $\beta$ does not have a fixed value. In contrast, it can be assigned different weights between 0 and 1, so that the TS tool is more flexible, and let the user determine the

[9] This resource is free, available at ftp://ftp.cis.upenn.edu/pub/chunker/
[10] Please see the description of the *query similarity* stage in Section 3.3.

importance given to the generic relevance of sentences or to the query-related sentences,

(2) $$Sc_{s_i} = (1+\beta^2)\frac{r_{s_i} * qSim_{s_i}}{\beta^2 * r_{s_i} + qSim_{s_i}}$$

where:
$Sc_{s_i}$ = the score associated to sentence *i*,
$r_{s_i}$ = the relevance of the sentence *i*,
$qSim_{s_i}$ = the similarity between the query and sentence *i*.

Taking into account the size restrictions, the top-ranked sentences will be selected and extracted, forming the final query-focused extract.

(3) **Sentiment-based summaries (COMPENDIUM$_{SE}$)**. For generating these types of summaries, the same strategy as for the generic ones is followed. The main difference between these is that in this case we only focus on subjective information, which is identified and processed in the *subjective information detection* stage.[11] As a result, extracts containing only subjective information are produced.

(4) **Generic abstractive-oriented summaries**. These summaries (**COMPENDIUM**$_{E-A}$) combine extractive and abstractive TS strategies, and as a consequence, the sentences of the final summary are selected following a strategy that maximises the similarity between each of the new sentences generated in the *information compression and fusion* stage (Section 3.3), and the ones that have been selected as most important in the *relevance detection* stage, given that the similarity[12] between them is above a predefined threshold. In the cases where a sentence in the extract has an equivalent in the set of the new generated sentences, the former will be substituted for the latter; otherwise, the sentence as it appears in the extract will be kept. The resulting summaries are abstractive-oriented.

### *3.3 Additional stages*

Apart from the core stages, there are three additional stages (*query similarity*, *subjective information detection*, and *information compression and fusion*) that, when integrated with the former, enhance the capabilities of COMPENDIUM. These stages allow the generation of other types of summaries (i.e. query-focused, sentiment-based, and a abstractive-oriented summaries). Each of these additional stages are described below.

#### *3.3.1 Query similarity*

When query-focused summaries have to be produced, a query is usually associated to the source documents in order to specify the kind of information the user is

[11] This stage is explained in detail in Section 3.3.
[12] We use the cosine similarity measure to compute the similarity between two sentences.

interested in. From an extractive point of view, the summary should contain the most relevant sentences in the document that also contains the information in the query. Therefore, the goal of this stage is to take into account the information expressed in a given query to tailor the contents of the final summary to such information.

In order to determine which sentences may be potentially related to the given query, the cosine similarity between each sentence and the query is computed, using the Text Similarity package.[13] Formula 3 shows how the query similarity, through cosine similarity, is calculated:

$$qSim_{S_i} = cosine(Q, S_i) = \frac{\sum_{j=1}^{n} w_{j,Q} * w_{j,S_i}}{\sqrt{\sum_{j=1}^{n} {w_{j,Q}}^2} * \sqrt{\sum_{j=1}^{n} {w_{j,S_i}}^2}} \tag{3}$$

where:

$Q$ = the query,
$s_i$ = the sentence $i$ in the document,
$w_{j,Q}$ = the weight of term $j$ in the query, and
$w_{j,S_i}$ = the weight of term $j$ in the sentence $i$.

### *3.3.2 Subjective information detection*

The objective of this stage is to detect and process subjective information, with the purpose of producing sentiment-based summaries. This has to be performed before the relevance of sentences is assigned. In order to be able to carry out such task, a tool capable of analysing and classifying the sentiment associated to a fragment of text (e.g. words, sentences, documents, etc.) is necessary. This manner, it is possible to know whether a fragment of text is subjective or objective, and, in addition, whether it is positive, negative, or netural about a specific topic, entity, product, etc. In particular, COMPENDIUM uses the external opinion mining tool described in Balahur-Dobrescu *et al.* (2009), which relies on opinion lexicons for identifying positive and negative words and expressions. Then it computes the number of positive and negative words in sentences and, depending on the resulting value, it assigns to each sentence in the document one of these three different scores: (i) $score > 0$ if the sentence has a positive nature; (ii) $score < 0$ if it has a negative nature; and (iii) $score = 0$ if the sentence is neutral (i.e. its an objective sentence). Once all the sentences have been analysed according to their associated polarity, we discarded the neutral sentences for further stages so that only those which have been identified as positive or negative according to the opinion mining tool are taken into consideration. In this manner, the resulting summary will contain the different opinions people have towards a topic, product, etc., thus being a sentiment-based (or subjective) summary.

[13] http://www.d.umn.edu/~tpederse/text-similarity.html

### 3.3.3 Information compression and fusion

This stage aims at generating new sentences in one of these forms: either a compressed version of a longer sentence, or a new sentence containing information from two individual ones. The main steps involved in this stage are as follows:

- **Word graph generation**. For generating new sentences, we rely on word graphs adopting a similar approach to the one described in Filippova (2010). Specifically in our approach, we first generate an extractive summary in order to determine the most relevant content for being included in the summary, and then we apply this stage for compressing and merging relevant information, since it has been proven that this approach is the most appropriate one (Lloret and Palomar 2011a).
  Therefore, taking as input the generated extract, it is represented as a directed weighted $DG = (V, E)$, where $V = v_i, v_{i+1}, \ldots, v_{i+n}$ is the set of nodes corresponding to document's words, and $E = e_{i,i+1}, e_{i+1,i+2}, \ldots, e_{i-n,i+n}$ is the set of edges, which consists of adjacency relations between the words. Two words are mapped into the same node only if they have the same part of speech tag. However, it is important to stress that stop words are not mapped together; otherwise, the real meaning of the sentence could be changed when generating the new sentence. The weight of each edge is calculated based on the inverse of the frequency of co-occurrence of two adjacent words, taking also into account the importance of the nodes they link through the Pagerank algorithm (Brin and Page 1998). For implementation we used Python-graph library.[14]
  Once the extract is represented as a word graph, a pool of new sentences is created. In order to produce a new sentence, we employ Dijkstra's algorithm (Dijkstra 1959) to find the shortest paths between an initial node and the remaining ones that are directly or indirectly connected with it. The initial node corresponds to the first word in each sentence, since this manner, we ensure that for each sentence in the extract, we have one derived sentence, so the whole content of the extract is covered. Moreover, we chose the shortest path algorithm because on the one hand, it has been shown to be appropriate for compressing sentences in previous work (Filippova 2010), and on the other hand, the shortest path will also look for minimal-length sentences that contains information from several ones, thus allowing them to include more information.
- **Incorrect paths filtering**. By applying the Dijkstra's algorithm over the graph we obtain all possible shortest paths between one node and the remaining ones. This leads to a high number of resulting sentences, which are not equally good. In fact, some of the sentences might be completely incomprehensible and not correctly formed. Therefore, in order to guarantee the completeness and correctness of a new sentence, this stage is needed to discard the invalid paths (i.e. generated sentences). For instance, some of them may

[14] http://code.google.com/p/python-graph/

suffer from incompleteness. Consequently, in order to reduce the number of incorrectly generated sentences, we define a set of rules, so that sentences not accomplishing all the rules are not taken into account. Three general rules are defined after analysing manually a set of generated sentences derived from a small data set, which are as follows:

— The minimal length for a sentence must be three words.[15]
— Every sentence must contain a verb.
— The sentence should not end in an article (e.g. a, the), a preposition (e.g. of), an interrogative word (e.g. who), or a conjunction (e.g. and).

Once the incorrect sentences have been removed, we can use the new sentences instead of the original ones (see "generic abstractive-oriented summaries" in the *summary generation* stage). It is important to stress upon the fact that, in COMPENDIUM, this stage takes place after the *relevance detection* and before the *summary generation* stages. This is because we need to be aware of the most relevant sentences of the document first.

## 4 Evaluation environment

The aim of this section is to present an intrinsic evaluation of COMPENDIUM. In order to assess its performance, a number of experiments with various corpora from a wide range of domains and different text genres are conducted.

ROUGE[16] was selected as the tool for automatically evaluating our summaries, since it is a widespread TS evaluation tool that has been shown to correlate well with human evaluations (Lin and Hovy 2003). It is able to evaluate how informative an automatic summary is by comparing its content to one or more reference summaries. Such comparison is made in terms of *n-gram* co-occurrence. The best-known ROUGE metrics are ROUGE-1 and ROUGE-2, which compute the number of overlapping unigrams and bigrams, respectively; ROUGE-SU4, which measures the overlap of skip-bigrams and automatic summary, contains with respect to model one, a maximum distance of four words between them; and finally, ROUGE-L, which calculates the longest common subsequence between two summaries.

Conducting a manual evaluation is very costly and time-consuming; however, it is useful for knowing what real users think about the automatic summaries. Therefore, in the cases where we do not have model summaries, we perform a manual evaluation, taking into account different criteria concerning the quality of the generated summaries, such as grammaticality, redundancy, or focus. In other cases, the user satisfaction is evaluated in order to determine the usefulness of an automatic summary from a human point of view. In these cases, we establish a rating scale where different degrees of goodness are analysed. For instance, a 3-level scale may comprise the values "low", "medium", and "high", whereas in a 5-level Likert scale the degrees to measure the agreement with respect to a specific issue

[15] We assume that three words (i.e. subject + verb + object) is the minimum length for a complete sentence.
[16] http://berouge.com/default.aspx

are established ("strongly agree", "agree", "neither agreee nor disagree", "disagree", and "strongly disagree").

In this section, we first describe the corpora used (Section 4.1), and then we focus on the experiments and results obtained (Section 4.2). A discussion for each experiment is also provided together with the results.

### *4.1 Corpora*

As corpora, we use several data sets belonging to different textual genres, depending on the experiment performed and the type of summary we want to evaluate. In this manner, COMPENDIUM can be assessed from a broader perspective by focusing on the newswire, tourist, Web 2.0, and scientific domains. In particular, COMPENDIUM is evaluated using the following corpora:

- **Newswire**. It consists of a collection of English newswire documents provided by DUC.[17] On the one hand, we use the news documents within the DUC 2002 conference. These are 567 documents grouped in fifty-nine clusters, where each cluster represents a set of topic-related documents. Moreover, model summaries for each news are also provided. Specifically, for each document, the number of reference summaries range from 1 to 2 (1,112 model summaries in total). This data are appropriate for carrying out single- and multi-document generic TS. On the other hand, for multi-document summarisation, the data provided in DUC 2003 and DUC 2004 are also suitable. They consist of thirty and fifty clusters of news documents, for DUC 2003 and 2004, respectively, containing approximately ten documents each, and four model summaries for each cluster.
- **Image captions**. This corpus was created by Aker and Gaizauskas (2010), and it contains 308 different images with manually assigned place names. Each image has ten documents in English related to it that have been retrieved using a search engine, where the name of each place has been set as the query. In addition, each image has up to four model summaries (932 in total), which were manually created using the on-line social site, VirtualTourist.[18] This corpus is especially suitable for generating query-focused summaries of multiple input documents.
- **Blogs**. This corpus consists of fifty-one blogs together with their comments extracted from the web. In particular, these were manually selected using Technorati[19] website, which allows us to specifically search for blogs and comments. Since we wanted to cover general topics, we chose among different blogs related to economics, science and technology, cooking, society, and sports. The blogs and their corresponding comments are written in English and all of them have the same structure: The authors create an initial entry containing a piece of news and their opinion on it, and subsequently bloggers

[17] http://www-nlpir.nist.gov/projects/duc/data.html
[18] http://www.virtualtourist.com/
[19] http://technorati.com/

Table 1. *Overview of the corpora used and their properties*

| Corpus | Num. clusters | Num. docs | Avg. length (docs) | Num. model sum | Avg. length (model sum) |
|---|---|---|---|---|---|
| Newswire DUC 2002 | 59 | 567 | 630 | 1,112 | 100 |
| Newswire DUC 2003 | 30 | 298 | 669 | 120 | 100 |
| Newswire DUC 2004 | 50 | 500 | 601 | 200 | 100 |
| Image captions | 308 | 3,080 | 690 | 932 | 200 |
| Blogs | – | 51 | 5,548 | – | – |
| Medical research papers | – | 50 | 2,060 | 50 | 162.7 |

reply expressing their opinions about the topic. Since we wanted to build the corpus for generating summaries, the criteria for selecting blogs were the length and the number of comments they contained. After computing some statistics about our corpus, each blog had on average thirty-four comments associated to the main entry, and they had in total more than 5,500 words on average. We use this corpus for generating sentiment-based summaries.

- **Medical research papers**. This corpus consists of a collection of fifty research papers from specialised journals of medicine that were gathered directly from the web.[20] Each paper contains a human-written abstract, that can be considered as a model summary. This collection of journal papers are appropriate to perform single-document summarisation.

Table 1 shows an overview of all the data sets we have worked with. In addition, some properties are also provided. In particular, for each corpus, the table shows the number of clusters, the number of total documents within each corpus, the average size of the documents (in number of words), the number of model summaries available for each corpus, and finally the average length of the model summaries (in number of words, as well).

### *4.2 Experiments*

In this section, the experiments conducted for evaluating the types of summaries COMPENDIUM is able to generate are described. In particular, we evaluate single- and multi-document summaries; generic, query-focused, and sentiment-based summaries; and extracts and abstractive-oriented summaries.

Our aim is to analyse whether the techniques proposed within COMPENDIUM are appropriate for generating the most common types of summaries, and being applied to different text genres of different domains (newswire, image captions, blogs, and medical research papers). The results obtained will be compared with other TS systems in order to determine the performance of COMPENDIUM with respect to the state-of-the-art.

[20] http://www.elsevier.com/wps/product/cws home/622356

Table 2. COMPENDIUM$_E$ *results for single- and multi-document summarisation for the newswire domain, F-measure* ($\beta = 1$)

| | | Single-document | Multi-document |
|---|---|---|---|
| | ROUGE-1 | 0.45611 | 0.30137 |
| DUC 2002 | ROUGE-2 | 0.20252 | 0.05327 |
| | ROUGE-L | 0.41382 | 0.26373 |
| | ROUGE-1 | – | 0.28977 |
| DUC 2003 | ROUGE-2 | – | 0.05481 |
| | ROUGE-L | – | 0.25399 |
| | ROUGE-1 | – | 0.31091 |
| DUC 2004 | ROUGE-2 | – | 0.06316 |
| | ROUGE-L | – | 0.27633 |

We grouped the different evaluations conducted with respect to the type of textual genre analysed (newswire, image captions, blogs, and medical research papers). We next described in detail the experiments performed and the results obtained. For clarity reasons, each textual genre is in a separate section.

### *4.2.1 Newswire corpus*

Using the different DUC corpora for news, COMPENDIUM$_E$ is evaluated for single- and multi-document summarisation. For both evaluations, we use ROUGE-1, ROUGE-2, and ROUGE-L, and we report the values for F-measure ($\beta = 1$). On the one hand, for single-document summarisation, we use the DUC 2002 corpus, on the other hand, this corpus together with the ones for DUC 2003 and DUC 2004 are used for generating multi-document summaries.

Table 2 shows the results that COMPENDIUM$_E$ obtains for different newswire corpora. As it can be seen, single-document summaries achieve better results (around 45% for ROUGE-1) than multi-document ones (30% on average). The difference in performance may be due to the fact that we do not employ any specific technique for tackling multi-document summarisation. In contrast, we merge all related documents into a single one, and this is used as input for COMPENDIUM$_E$.

An example of a single- and a multi-document summary generated by COMPENDIUM$_E$ is shown in Table 3.[21]

Furthermore, we carry out a comparison of our results with respect to the best scoring system in the respective DUC editions, as well as a Lead baseline that builds the summary taking the first sentence of each document in the case of single-document summarisation, and it takes the first sentence of the first document, then the first sentence of the second document, and so on, for multi-document. In these cases, we report the recall value for ROUGE-1. Table 4 shows such comparison.

[21] The original news for single- and mulit-document summarisation can be found at: http://intime.dlsi.ua.es/downloads/elloret/ORIGINAL_DOCUMENTS/AP880325-0239 and http://intime.dlsi.ua.es/downloads/elloret/ORIGINAL_DOCUMENTS/d078b.zip

Table 3. *Single- and multi-document summaries generated by* COMPENDIUM$_E$ *for document AP880325-0239 (DUC 2002, cluster d078b)*

| COMPENDIUM$_E$ (single-document summary) |
|---|
| What do Charlie Chaplin, Greta Garbo, Cary Grant, Alfred Hitchcock, and Steven Spielberg have in common?<br>They have never won Academy Awards for their individual achievements.<br>Oscar's 60-year history is filled with examples of the film world's highest achievers being overlooked by the Academy of Motion Picture Arts and Sciences.<br>The honorary award has also proved useful to salve the Academy's conscience.<br>Douglas Fairbanks, Judy Garland, Noel Coward, Ernst Lubitsch, Fred Astaire, Gene Kelly, Harold Lloyd, Greta Garbo, Maurice Chevalier, Stan Laurel, Cary Grant, Lillian Gish, Edward G. Robinson, Groucho Marx, Howard Hawks, and Jean Renoir are others who have received honorary awards. |
| COMPENDIUM$_E$ (multi-document summary) |
| Oscar, manufactured by the R. S. Owens Co., Chicago, is made of Britannia metal, copper plate, nickel plate, and gold plate.<br>They have never won Academy Awards for their individual achievements.<br>Oscar's 60-year history is filled with examples of the film world's highest achievers being overlooked by the Academy of Motion Picture Arts and Sciences.<br>The honorary award has also proved useful to salve the Academy's conscience.<br>How long was the longest Oscar ceremony?<br>Free enterprise has collided with the Academy Awards, and everybody's trying to pick up the pieces. |

Table 4. *Comparison of* COMPENDIUM$_E$*'s performance with other text summarisation systems (recall value)*

| | DUC 2002 | | DUC 2003 | DUC 2004 |
|---|---|---|---|---|
| TS system | Single | Multi | Multi | Multi |
| Best DUC participant | 0.42776 | **0.35151** | **0.37980** | **0.38232** |
| COMPENDIUM$_E$ | **0.46008** | 0.30341 | 0.29355 | 0.31362 |
| Lead baseline | 0.41132 | 0.22771 | 0.20967 | 0.31293 |

The single-document summaries achieve very good performance compared with the best system at DUC 2002, and the Lead baseline. In this case, COMPENDIUM$_E$ outperforms the best system by approximately 8% (0.46008 vs. 0.42776 – please see Table 4), and an increase of 12% is obtained over the baseline (0.46008 vs. 0.41132). On the contrary, results for multi-document summarisation are not as good. The Lead baseline is improved by 33% and 40% for the DUC 2002 (0.30341 vs. 0.22771) and DUC 2003 data (0.29355 vs. 0.20967), respectively, but there is a marginal increase for DUC 2004 (0.2%, 0.31362 vs. 0.31293). Despite these improvements, the performance of COMPENDIUM$_E$ for multi-document summarisation does not surpass the best system at DUC. This difference is due to the fact that we approach

Table 5. *Results for generic summarisation in the image caption corpus (recall value)*

| TS system | ROUGE-2 | ROUGE-SU4 |
|---|---|---|
| Wikipedia baseline | 0.09632 | 0.14203 |
| SummGraph | 0.08950 | 0.14290 |
| MEAD | 0.08866 | 0.13769 |
| COMPENDIUM$_E$ | 0.08551 | 0.13371 |
| SUMMA | 0.06423 | 0.10919 |

Table 6. *Results for query-focused summarisation in the image caption corpus (recall value)*

| TS system | ROUGE-2 | ROUGE-SU4 |
|---|---|---|
| Wikipedia baseline | 0.09632 | 0.14203 |
| SummGraph | 0.10075 | 0.15430 |
| MEAD | 0.10192 | 0.15353 |
| COMPENDIUM$_{QE}$ | 0.08864 | 0.13892 |
| SUMMA | 0.06532 | 0.10946 |

multi-document summarisation as a single-document summarisation where all related documents are considered as a single one. This is an indication that this type of summarisation may require a more elaborate processing instead.

### *4.2.2 Image caption corpus*

Generic and query-focused summaries were produced using COMPENDIUM$_E$ and COMPENDIUM$_{QE}$, respectively. For the query-focused summaries generated by COMPENDIUM$_{QE}$, it was experimentally established that, for this research and this corpus, the optimal value for the parameter $\beta$ was 0, thus meaning that the sentences related to the query were considered more important to build the summary. As far as the generic summaries (COMPENDIUM$_E$) are concerned, the query similarity stage was not taken into consideration.

In total, 308 summaries of 200 words each were generated, and to evaluate both approaches, we use ROUGE-2 and ROUGE-SU4. These metrics allowed us to compare the automatic summaries against the model summaries also provided in this corpus.

Furthermore, as baseline the first 200 words of the Wikipedia article describing each image were selected. Also, for comparison purposes, we generated summaries employing different state-of-the-art TS systems. In particular, these were SummGraph (Plaza, Díaz and Gervás 2008), MEAD (Radev *et al.* 2004), and SUMMA (Saggion 2008). In this way, the performance for generic and query-based summarisation for such systems was also compared. The results are given in Tables 5 and 6, respectively.

From the results obtained, we can conclude that query-focused summarisation is more appropriate for these types of data. If we observed the results for the summarisers, all ROUGE scores in Table 6 (query-focused summaries) are higher than the scores in Table 5 (generic summaries). Moreover, the Wilcoxon signed-rank test is computed for assessing the significance of the results. For all summarisers, except SUMMA, the query-focused summaries are significantly better than the generic ones.

Regarding our TS approach, COMPENDIUM, the results for the query-focused approach (COMPENDIUM$_{QE}$) increase by approximately 4% on average with respect to the generic one, COMPENDIUM$_{E}$ (0.08864 vs. 0.08551 for ROUGE-2 and 0.13892 vs. 0.13371 for ROUGE-SU4), thus proving that query-focused summaries are the most appropriate ones for these type of documents.

On the other hand, Wikipedia summaries are a difficult goal to achieve, as it can be seen from the results where only two summarisers, when producing query-focused summaries (SummGraph and MEAD), obtain higher results. The reasons why it is so difficult to perform better than the Wikipedia baseline are as follows: (1) These articles have been created by humans; (2) the first paragraph in a Wikipedia article is usually a summary of the entire document content; and (3) Wikipedia articles almost exclusively contain salient information to the subject matter, and do not present other information somehow related to the topic but not important (e.g. nearby hotels, or transport services).

Although the results achieved by COMPENDIUM are comparable to the best performing systems, we try to elucidate the reasons why COMPENDIUM performs a bit lower than SummGraph and MEAD.

Analysing the types of documents and the resulting summaries, we realised that one of the main problems in these specific types of documents resides in the nature of the corpus, which contained some noisy information, in the sense that they included a lot of noun-phrases within the documents as categories for the objects (e.g *"Mahogany, Maple, crown mouldings, multiple Viking ovens, Sub-Zero refrigerators, antique..."*). In these cases, the method proposed in COMPENDIUM for detecting relevant information did not perform as it would be expected. The reason is that, according to the CQP feature, our method gives more importance to those sentences which contains longer noun-phrases, and consequently, these types of sentences are scored higher. Therefore, they are wrongly considered as relevant, and they are final incorporated in the summary. In the end, this leads to the fact that the quality of the final summaries is directly affected by these sentences, and as a consequence, the content of the generated summary will show greater differences in comparison to the sentences extracted by other summarisers. In order to improve our results, we could have classified which of the noun-phrases may be related to the topic and which may not be.

Nevertheless, the resulting summaries are good enough to provide an idea of the most important facts about a location, as it is illustrated through the two fragments of summaries related to "Nou Camp" shown in Table 7.[22] The summary at the top

[22] The original documents about *Camp Nou* can be found at: http://intime.dlsi.ua.es/downloads/elloret/ORIGINAL_DOCUMENTS/CampNou.zip

Table 7. *Generic and query-focused summaries generated by* COMPENDIUM *for the place* Camp Nou

| COMPENDIUM$_E$ (generic summary) |
|---|
| Nou Camp in both Spanish and English is a football stadium in Barcelona, Spain.<br>The stadium has been the home of FC Barcelona since The stadium's construction in 1957.<br>The stadium is a UEFA 5-star rated stadium, and has hosted numerous international matches at senior level, and UEFA Champions League finals, the most recent being in 1999.<br>The stadium has a capacity of 98,772, making The stadium the largest stadium in Europe, and the eleventh largest in the world.<br>The stadium's official name was Estadi del FC Barcelona FC Barcelona Stadium until 2000, when the club membership voted to change the official name to the popular nickname, Camp Nou.<br>Across Camp Nou is the Palau Blaugrana, the stadium for indoor sports and adjacent is the Ice Rink, the stadium for ice-based sports. [. . . ] |
| COMPENDIUM$_{QE}$ (query-focused summary) |
| The Camp Nou "new field", Catalan pronunciation: ['kam 'now], often erroneously called the "Nou Camp" in both Spanish and English is a football stadium in Barcelona, Spain.<br>The stadium's official name was Estadi del FC Barcelona FC Barcelona Stadium until 2000, when the club membership voted to change the official name to the popular nickname, Camp Nou.<br>Across Camp Nou is the Palau Blaugrana, the stadium for indoor sports and adjacent is the Ice Rink, the stadium for ice-based sports.<br>By the early 1950s, Barcelona had outgrown Barcelona's old stadium, Camp de Les Corts which had held 60,000 supporters.<br>The Camp Nou, built between 1954 and 1957, was designed by architects Francesc Mitjans-Mir, Lorenzo Garc a Barbon, and Josep Soteras Mauri.<br>FC Barcelona won Eulogio Martínez first game at Camp Nou in impressive fashion, a 4-2 victory against Legia Warsaw, with Eulogio Martínez scoring the first goal at the new stadium. [. . . ] |

corresponds to a generic summary produced using COMPENDIUM$_E$, whereas the one at the bottom is a query-focused one (COMPENDIUM$_{QE}$).

### *4.2.3 Blog corpus*

The comments associated to each blog in the corpus are used to produce sentiment-based summaries (COMPENDIUM$_{SE}$). Different from the previous experiments, instead of generating summaries of a fixed length, we experimented with three different compression rates (10%, 15%, and 20% of the document).

The evaluation conducted also varies from the previous ones in the sense that ROUGE is not employed. Instead, a qualitative evaluation is conducted, where summaries are evaluated manually. It is worth stressing upon the fact that we focus more on the quality of the summaries rather than on their content, since the content would depend on the specific need a user has at a particular moment. Moreover, for

Table 8. *Results for* COMPENDIUM$_{SE}$ *within the blog corpus*

| | | Compression Rate | | |
|---|---|---|---|---|
| Criterion | % | 10% | 15% | 20% |
| | Non-acceptable | 26 | 0 | 4 |
| Redundancy | Understandable | 45 | 6 | 10 |
| | Acceptable | 29 | 94 | 86 |
| | Non-acceptable | 4 | 2 | 0 |
| Grammaticality | Understandable | 22 | 27 | 55 |
| | Acceptable | 74 | 71 | 45 |
| | Non-acceptable | 33 | 26 | 14 |
| Focus | Understandable | 43 | 29 | 47 |
| | Acceptable | 24 | 45 | 39 |
| | High | 35 | 18 | 8 |
| Difficulty | Medium | 28 | 35 | 51 |
| | Low | 37 | 51 | 41 |

this corpus we do not have model summaries, and to produce them manually is a very difficult and time-consuming task.

In particular, the criteria proposed for evaluating the opinion summaries are the following: *redundancy*, *grammaticality*, *focus*, and *difficulty*. *Redundancy* measures the presence of repeated information in a summary. *Grammaticality* accounts for the number of spelling or grammatical errors that a summary presents. *Focus* evaluates whether it is possible or not to understand the topic of the summary, that is, the main subject of the text; and finally, *difficulty* refers to the extent to which a human can understand a summary as a whole or not.

If we have a look at the criteria proposed in DUC and TAC conferences, we will realise that we adopt more or less the same, except the difficulty criteria which is non-conventional. The reason why this criterion is included is that it provides an idea of the summary as a whole, regarding its readability.

Furthermore, three different degrees of goodness are established for each evaluated criteria. These were *non-acceptable*, *understandable*, and *acceptable*. In this classification, *acceptable* means that the summary meets the specific criterion and therefore is good, whereas *non-acceptable* means that the summary would not be good enough with respect to a criterion. When assessing the *difficulty*, the summaries were classified with regard to *high*, *medium*, and *low*, being low, the better.

Table 8 shows the results for this evaluation.

Analysing the results obtained, a set of interesting conclusions can be drawn. As far as the grammaticality criterion is concerned, the results show a decrease of grammaticality errors as the size of the summary lowers. We can see that the number of acceptable summaries varies from 74% to 45%, for a compression rate of 20% and 10%, respectively. This is obvious because the longer the summary, the more chances there are for containing orthographic or grammatical errors. Due to the informal language used in blogs, we thought *a priori* that summaries would contain many spelling mistakes. Contrary to our expectations, generated summaries are quite well written, only 4% of them, at most, being non-acceptable. Another important fact that can be inferred from the results is related to how the summaries deal with the topic. According to the percentages shown in the tables previously presented, the number of summaries that have correctly identified the topic and have therefore been evaluated as acceptable, change considerably with respect to the different summary sizes, increasing when we changes from 10% to 15%, but decreasing when changing from 15% to 20%. However, as a general trend, we can see that when taking into account the number of summaries that have not performed correctly in the focus parameter, there is a decreasing trend, reducing the incorrect summaries from 33% to 14%. This means that for longer summaries, the topic may be stated along the summary, although not necessarily at the beginning of it, whereas for shorter summaries, there is no such flexibility, and as a consequence if the topic does not appear at the beginning, the most probable thing is that it does not appear in the summary at all. Finally, regarding redundancy, results indicate that summaries of 15% and 20% contain less repeated information than shorter ones. What can be seen from the results is that the summaries of 20% compression rate obtain the best results on average over the rest of the size experimented with. This is due to the fact that this compression rate achieves higher percentage (for the *understandable* and *acceptable* degrees of goodness) in two (grammaticality and focus) out of the three criteria proposed. Only the 15% compression rate summaries obtained better results in the redundancy criterion. On the other hand, as far as the difficulty criteria is concerned, results are also encouraging. According to the evaluation performed, the longer the summaries, the easier they are to be understood in general. Grouping the percentages of summaries, we obtained that 65%, 86%, and 92% of the summaries of size 10%, 15%, and 20%, respectively, have either medium or low level of difficulty, which means that they could be understood as a whole without serious difficulties. Again, for this criterion, 20% summaries achieve the best results. This has also been proved by previous research work, which demonstrated that this compression rate is more suitable for an acceptable quality of summaries (Morris, Kasper and Adams 1992). It is worth mentioning that this criterion is rather subjective and depends to a large extent on different factors, such as the knowledge the person who reads the summaries, the number of grammatical errors the text contains, or the connectedness of the sentences. Moreover, it is reasonable to think that long summaries can be more difficult to understand, but our experiments show that it is actually the other way around, because longer summaries may contain more information than shorter ones, which allows the user to have more awareness of the content and what the summary is about.

In general terms, while evaluating the summaries obtained, we noted some recurrent mistakes. The first one is the punctuation; in some cases we noted some commas missing or instead of having a comma, contain a full stop (e.g. “So. One option. . . ”). Also, in some cases, apostrophes are missing, in examples such as *dont*. The second error is that in some cases the summaries start with a sentence containing a coreference element that we cannot resolve, because the antecedent has been deleted or sentences that imply some concept previously mentioned in the original text have not been selected. It is also worth mentioning that some of the

Table 9. *Sentiment-based summary generated by* COMPENDIUM$_{SE}$ *for blog 29*

| COMPENDIUM$_{SE}$ (sentiment-based summary) |
|---|
| Clothilde, I love the wallpapers!<br>They keep everything tasty and fresh!<br>Thanks a lot for the gorgeous calender desktop background.<br>What a great idea and beautiful photo.<br>I've just started recreating some of the easier and more attainable recipes.<br>Another lovely calendar! Clotilde, have you discontinued your "Bonjour mois" newsletter?<br>I'm terribly late this month but was enjoying the cheese so much that I just forgot! The peas are another winner of course.<br>My only quibble would be about the name. |

grammatical errors are due to users' misspellings, for example "I thikn". The third error concerns the spelling mistakes found in the summaries, which are directly transferred from the initial blog posts, which also contains such kind of errors (e.g. calender). Finally, we also found some void sentences, that do not contribute to the general meaning of the summary as for example, "I m an idiot", "Just an occasional visitor", or "welcome back!!!".

Table 9 shows an example of an automatic summary for blog 29 with a compression rate of 10%.[23] In this case, the summary contains mostly positive opinions, having only the last sentence a negative charge (*"My only quibble would be about the name"*).

As it can be seen, only opinions have been considered and these are presented grouped into positives, on the one hand, and negatives, on the other. We considered it as good due to the fact that there are no objectives or useless sentences. The system presents subjective sentences with an emotional charge, and as a consequence this summary meets our purposes.

### *4.2.4 Medical research papers corpus*

With this set of documents, we want to analyse the capabilities of COMPENDIUM for generating abstractive-oriented summaries. In this particular evaluation, we generate summaries of 162 words because this was the average length of the abstracts included in the medical papers (please see Table 1), and we wanted to use them as model summaries. Therefore, since our goal is to analyse to what extent the resulting automatic summaries are valid, we set up a comparison between COMPENDIUM$_{E}$ and COMPENDIUM$_{E-A}$, considering as model summaries the ones included in the medical papers.

Table 10 shows an example of two summaries generated with COMPENDIUM$_{E}$ and COMPENDIUM$_{E-A}$, respectively.[24] It is worth mentioning that these approaches

[23] The original blog and its comments can be found at: http://intime.dlsi.ua.es/downloads/elloret/ORIGINAL_DOCUMENTS/blog29.txt

[24] For both types of summaries, the original medical paper can be found at: http://intime.dlsi.ua.es/downloads/elloret/ORIGINAL_DOCUMENTS/medicalDoc

Table 10. *Example of summaries generated with* COMPENDIUM$_E$ *and* COMPENDIUM$_{E-A}$ *for the medical corpus*

| COMPENDIUM$_E$ (extractive summary) |
|---|
| Histologic examination of lesions plays a key role in the diagnostics of cutaneous lupus erythematosus LE. |
| LE has a broad spectrum of histological signs which are related to the stages of the lesions, but some signs apply to all stages e.g. mucin deposition. |
| Histologic findings of skin lesions are essentially identical for systemic lupus erythmatosus SLE and cutaneous LE. |
| From the histological standpoint, LE can be classified only into early, fully developed, late LE, and special manifestations of LE. |
| The early histologic findings of LE lesions are sparse superficial perivascular lymphocytic infiltrates, neutrophils, and sometimes nuclear dust immediately beneath the dermoepidermal junction. |
| Few individual necrotic keratinocytes and focal vacuolar alteration of basal cells may occur. |
| Fully developed lesions are characterized by moderately dense to dense perivascular and periappendageal lymphocytic infiltrates in the papillary and reticular dermis with abundant mucin deposition in the reticular dermis. |
| According to Kuhn *et al.* the presence of even slight epidermal or junctional involvement should exclude LE tumidus. |
| COMPENDIUM$_{E-A}$ (abstractive-oriented summary) |
| LE lesions plays a key role in the diagnostics. |
| LE has a broad spectrum of histological signs which are related to the stages of the lesions, but some signs apply to all stages e.g. mucin deposition. |
| LE lesions are essentially identical for systemic lupus erythmatosus SLE. |
| LE can be classified only into early histologic. |
| LE lesions are sparse superficial perivascular lymphocytic infiltrates neutrophils and sometimes nuclear dust immediately beneath the dermoepidermal junction. |
| Few individual necrotic keratinocytes and focal vacuolar alteration of basal cells may occur. |
| Fully developed lesions are characterized by moderately dense to dense perivascular and periappendageal lymphocytic infiltrates in the papillary and reticular dermis with abundant mucin deposition in the reticular dermis. |
| According to Kuhn *et al.* the presence of even slight epidermal or junctional involvement should exclude. |

produce generic summaries, and for generating them neither the keywords of the original paper nor the information in the titles or the abstracts were taken into consideration. As it can be seen, both summaries share some of the sentences, whereas others have been shorten in the latter.

Specifically, we set up two types of evaluation. In the first one, we use the model abstracts of the papers and we compare these with the ones generated by COMPENDIUM$_E$ and COMPENDIUM$_{E-A}$, for our extractive and abstractive-oriented approach, respectively. The second evaluation aims at assessing the summaries with regard to user satisfaction with respect to a 5-level Likert scale. Next, each of these types of evaluation is explained in more detail.

Table 11. *ROUGE-1 results for* COMPENDIUM$_{E-A}$ *and its comparison with other TS approaches*

| TS Approach | Recall | Precision | $F_{\beta=1}$ |
|---|---|---|---|
| COMPENDIUM$_E$ | **0.44022*** | 0.40525 | **0.42201*** |
| COMPENDIUM$_{E-A}$ | 0.38658 | **0.41809*** | 0.39533 |
| MS-Word 2007 | 0.43610 | 0.40456 | 0.41974 |

- **Comparison with human abstracts.** In this evaluation, the summaries generated by COMPENDIUM$_E$ and COMPENDIUM$_{E-A}$ are assessed with respect to the human abstracts provided in the original papers. We use ROUGE-1 and report the values of recall, precision, and F-measure ($\beta = 1$). We also compare the results obtained with the state-of-the-art summariser: MS-Word 2007 Summarizer,[25] and we run a t-test to account for statistical significance of the results at a 95% confidence level (statistical significant results are marked with a star).

  Table 11 shows the results for the first evaluation. In this evaluation, COMPENDIUM$_{E-A}$ summaries are evaluated with respect to the human abstracts, and compared to COMPENDIUM$_E$ and MS-Word 2007.

  As it can be seen, our both TS approaches are comparable with respect to the state-of-the-art TS tool (i.e. MS-Word 2007 Summarizer). Regarding COMPENDIUM$_{E-A}$, it is worth mentioning that the precision obtained is higher and statistically significant compared with the remaining approaches, thus meaning that the information contained is the right one. However, its recall is lower, so in the end the final value of F-measure is negatively affected, being COMPENDIUM$_E$ statistical significant with respect to COMPENDIUM$_{E-A}$ for these specific measures. This is due to the fact that for this TS approach we rely on the sentences detected as important in the relevance detection stage, and we compress or merge some information within them. Therefore, the resulting summaries are shorter than the extracts, and since no extra information is added, the recall value will be never higher than it is for COMPENDIUM$_E$. One possible solution to address this issue would be to rely on the source document and generate new sentences from it instead of the most relevant sentences. Another strategy would be to include in the COMPENDIUM$_{E-A}$ summary the next highest ranked sentence in the document according to the relevance detection stage that was not included in the extract because of summary length restrictions. Regarding COMPENDIUM$_E$, it achieves slightly better results than MS-Word 2007; however, there are no statistical differences between them. Only statistically significant results are obtained with respect to COMPENDIUM$_{E-A}$.
- **User satisfaction study.** In the last evaluation, we aim at assessing the user satisfaction with respect to the generated summaries. For this purpose, we

[25] http://www.microsoft.com/education/autosummarize.aspx

Table 12. *Qualitative questions to evaluate the summaries*

**Q1:** The summary reflects the most important issues of the document.
**Q2:** The summary allows the reader to know what the article is about.
**Q3:** After reading the original abstract provided with the article, the alternative summary is also valid.

Table 13. *User satisfaction results for the different text summarisation approaches*

| % | TS approach | Q1 | Q2 | Q3 |
|---|---|---|---|---|
| 1. Strongly disagree | COMPENDIUM$_E$ | 9.76 | 19.51 | 19.51 |
| | COMPENDIUM$_{E-A}$ | 2.44 | 0 | 2.44 |
| 2. Disagree | COMPENDIUM$_E$ | 41.46 | 19.51 | 34.15 |
| | COMPENDIUM$_{E-A}$ | 31.37 | 21.95 | 31.71 |
| 3. Neither agree nor disagree | COMPENDIUM$_E$ | 24.39 | 29.27 | 26.83 |
| | COMPENDIUM$_{E-A}$ | 21.95 | 29.27 | 26.83 |
| 4. Agree | COMPENDIUM$_E$ | 21.95 | 21.95 | 7.32 |
| | COMPENDIUM$_{E-A}$ | 41.46 | 39.02 | 34.15 |
| 5. Strongly agree | COMPENDIUM$_E$ | 2.44 | 9.76 | 12.20 |
| | COMPENDIUM$_{E-A}$ | 2.44 | 9.76 | 4.88 |

perform a qualitative evaluation and asked ten humans to evaluate our summaries[26] according to a 5-level Likert scale (1= strongly disagree . . . 5 = strongly agree). For each summary, humans were asked to respond to three questions concerning the appropriateness of the summaries. The questions asked, as well as the percentage of summaries for each question in scales 1 to 5 are shown in Tables 12 and 13, respectively.

As it can be seen from the results shown in Table 13, our abstractive-oriented approach (COMPENDIUM$_{E-A}$) obtains better results than the extractive one (COMPENDIUM$_E$). Although the evaluation concerning the information contained in the summaries generated with COMPENDIUM$_{E-A}$ was not as good as for the extractive approach, taking into consideration their quality from a human point of view, the abstractive-oriented summaries are much better than the extractive ones. When we have a look at the different percentages of summaries that have been rated in one of each categories, we observe that there is a higher percentage of abstractive-oriented summaries that humans agree with, compared with the extractive summaries for the same rating. Moreover, it is worth stressing upon the fact that, analogously, the percentage of summaries with lower ratings (strongly disagree and disagree) also decreases when COMPENDIUM$_{E-A}$ is employed.

Furthermore, concerning the average individual scoring results (between 1 and 5), COMPENDIUM$_{E-A}$ achieves at most 3.37 for Q2 and 3.1 for Q1 and Q3, whereas the maximum average value for COMPENDIUM$_E$ is 2.83 for Q2,

[26] The humans were also provided with the original papers and their abstracts.

the remaining questions obtaining values lower than 2.60. In light of the results obtained, it has been proved that the combination of extractive and abstractive techniques is more appropriate and leads to better summaries than extracts.

## 5 Conclusion and future work

This paper presented COMPENDIUM, a TS tool that is able to generate informative summaries for different purposes (generic, query-focused, sentiment-based, extractive, and abstractive-oriented), and it can deal with documents pertaining to a wide range of domains and textual genres (newswire, image captions, blogs, and medical research papers).

The architecture of COMPENDIUM is divided into two kinds of stages (core and additional stages), depending on the types of summaries one would like to generate. In light of the techniques employed within this architecture, three main contributions can be remarked that allow COMPENDIUM to be distinguished with respect to the state-of-the-art TS systems:

- **The use of textual entailment as a redundancy detection method**. This method combines lexical, syntactic, and semantic information in order to detect entailment relationships between sentences. Although textual entailment and TS had been used before to generate and evaluate summaries, this technique was not employed for dealing with the problem of redundancy in TS. Therefore, in this paper, we showed that textual entailment can be used for discarding sentences whose content has been already stated in previous sentences, thus avoiding repeated information in the summaries.
- **The combination of statistical and cognitive-based techniques, the latter directly related to how humans remember the information.** We analyse to what extent the *Code Quantity Principle* combined with the frequency of terms is appropriate for detecting relevant information in texts, and we showed that this combination led to good summaries in most of the cases.
- **Generate abstractive-oriented summaries** through the combination of extractive and abstractive techniques. A pool of new sentences (either compressed or fused) is created from an extract, and then these sentences are used to substitute those sentences in the extract that express the same information. In the cases where no equivalence is found, the sentences are not replaced. From the quantitative evaluation performed, we showed that this approach is appropriate, since it is able to keep the most relevant information. Moreover, from the user evaluation carried out, we confirmed that this approach generated better summaries than the extractive TS approach.

The proposed features and contributions of COMPENDIUM have lead to acceptable results, as it is proven from the extensive evaluation conducted. Specifically, an intrinsic evaluation was performed, whose goal was to assess the capabilities of COMPENDIUM for generating different types of summaries belonging to different domains and textual genres, as well as to compare it to the state-of-the-art TS systems. Therefore, the suitability of COMPENDIUM as a TS tool is shown.

However, from this research we have identified a number of issues that we have to work on in the future. As a first issue, we would like to analyse additional techniques and methods for the topic identification stage, such as the ones proposed and studied in Lin (1997) and Coursey and Mihalcea (2009). These are different from the word frequency counting, and therefore the problems and limitations of the word frequency technique would be avoided. In addition, coreference resolution as well as the use of concepts instead of words should also be investigated and included within our TS process.

Furthermore, we plan to carry out research into other issues. On the one hand, we want to enrich COMPENDIUM with semantic knowledge. This line of research broadens the research carried out into abstractive-oriented summarisation. The objective of this research is to incorporate semantic knowledge to COMPENDIUM by means of concept graphs or other semantic techniques. Semantic knowledge will allow us to have a higher level of abstraction. In the long term, we would also like to face the problem of evaluation by analysing the automation of several quality metrics with the final goal of developing a qualitative evaluation framework.

## Acknowledgments

This research has been supported by the project "Desarrollo de Técnicas Inteligentes e Interactivas de Minería de Textos" (PROMETEO/2009/119) and the project reference ACOMP/2011/001 from the Valencian Government, as well as by the Spanish Government (grant no. TIN2009-13391-C04-01).

## References

Aker, A., and Gaizauskas, R. 2010. Model summaries for location-related images. In *Proceedings of the 7th Language Resources and Evaluation Conference*, Valletta, Malta, pp. 3119–24.

Álvarez Angulo, T. 2002. *El Resumen Como Estrategia de Composición Textual y su Aplicación Didáctica*. PhD thesis, Universidad Complutense de Madrid, Madrid, Spain.

Azzam, S., Humphreys, K., and Gaizauskas, R. 1999. Using coreference chains for text summarization. *In Proceedings of the ACL'99 Workshop on Coreference and its Applications*, Baltimore, MD, USA.

Balahur, A., Lloret, E., Boldrini, E., Montoyo, A., Palomar, M., and Martinez-Barco, P. 2009. Summarizing threads in blogs using opinion polarity. In *Proceedings of the International Workshop on Events in Emerging Text Types*, Borovets, Bulgaria, pp. 5–13.

Balahur-Dobrescu, A., Kabadjov, M., Steinberger, J., Steinberger, R., and Montoyo, A. 2009. Summarizing opinions in blog threads. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation Conference*, Hong Kong, China, pp. 606–13.

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

Barzilay, R., and Elhadad, M. 1999. Using lexical chains for text summarization. In I. Mani and M. Maybury (eds.), *Advances in Automatic Text Summarization*, pp. 111–22. Cambridge, MA, USA: MIT Press.

Barzilay, R., and McKeown, K. R. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics* **31**(3): 297–328.

Becker, A. 2002. Análisis de la Estructura pragmática de la cláusula en el español de Mérida (Venezuela). *Estudios de Lingüística del Español* **17**, 18–32.

Bossard, A., Généreux, M., and Poibeau, T. 2009. CBSEAS, A summarization system integration of opinion mining techniques to summarize blogs. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, pp. 5–8.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks ISDN Systems* **30**, 107–17.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. 2005. Learning to rank using gradient descent. In *ICML '05: Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, pp. 89–96.

Chali, Y., and Hasan, Sadid A. n.d. Query-focused multi-document summarization: automatic data annotations and supervised learning approaches. *Natural Language Engineering* **18**(1): 109–45.

Coursey, K., and Mihalcea, R. 2009. Topic identification using wikipedia graph centrality. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pp. 117–20. Stroudsburg, PA, USA: Association for Computational Linguistics.

Cristea, D., Postolache, O., and Pistol, I. 2005. Summarisation through discourse structure. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing,* Mexico City, Mexico, pp. 632–44.

Delmonte, R., Bristot, A., Boniforti, M. A. P., and Tonelli, S. 2006. Another evaluation of Anaphora resolution algorithms and a comparison with GETARUNS' knowledge rich approach. In *Proceedings of the 4th International Workshop on Robust Methods in Analysis of Natural Language Data (Romand 06)*, Trento, Italy, pp. 3–10.

Dijkstra, E. W. 1959. A note on two problems in connexion with graphs. *Numerische Mathematik* **1**: 269–71.

Dumont, E., and Mérialdo, B. 2009. Automatic evaluation method for rushes summary content. In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo*, Cancun, Mexico, pp. 666–9.

El-haj, M. O., and Hammo, B. H. 2008. Evaluation of query-based Arabic text summarization system. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, pp. 1–7, Beijing, China

Ercan, G., and Cicekli, I. 2008. Lexical cohesion-based topic modeling for summarization. In *Proceedings of the 9th International Conference in Computational Linguistics and Intelligent Text Processing*, Haifa, Israel, pp. 582–92.

Erkan, G., and Radev, D. R. 2004. LexRank: graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* **22**: 457–79.

Fan, J., Gao, Y., Luo, H., Keim, D. A., and Li, Z. 2008. A novel approach to enable semantic and visual image summarization for exploratory image search. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, Vancouver, Canada, pp. 358–65.

Fellbaum, C. 1998. *WordNet: An Electronical Lexical Database*. Cambridge, MA, USA: The MIT Press.

Ferrández, Ó. 2009. *Textual Entailment Recognition and its Applicability in NLP Tasks.* PhD thesis, University of Alicante.

Filatova, E., and Hatzivassiloglou, V. 2004. Event-based extractive summarization. In M.-F. Moens and S. Szpakowicz (eds.), *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pp. 104–11. Stroudsburg, PA, USA: ACL.

Filippova, K. 2010. Multi-sentence compression: finding shortest paths in word graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, pp. 322–30.

Fuentes, M., Alfonseca, E., and Rodríguez, H. 2007. Support vector machines for query-focused summarization trained and evaluated on pyramid data. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, pp. 57–60.

Givón, T. 1990. *Syntax: A Functional-Typological Introduction, II.* Amsterdam, Netherlands: John Benjamins.

Glickman, O. 2006. *Applied Textual Entailment.* PhD thesis, Bar-Ilan University.

Gonçalves, P. N., Rino, L., and Vieira, R. 2008. Summarizing and referring: towards cohesive extracts. In *DocEng '08: Proceedings of the 8th ACM Symposium on Document Engineering*, Sao Paulo, Brazil, pp. 253–6.

Haghighi, A., and Vanderwende, L. 2009. Exploring content models for multi-document summarization. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York, USA, pp. 362–70.

Harabagiu, S., Hickl, A., and Lacatusu, F. 2007. Satisfying information needs with multi-document summaries. *Information Processing & Management* **43**(6): 1619–42.

Harabagiu, S., and Lacatusu, F. 2005. Topic themes for multi-document summarization. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, pp. 202–9.

Hennig, L. 2009. Topic-based multi-document summarization with probabilistic latent semantic analysis. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria, pp. 144–9.

Hovy, E. H., and Lin, C-Y. 1999. Automated multilingual text summarization and its evaluation. Technical Report, Information Sciences Institute, University of Southern California, Los Angeles, SC, USA.

Ji, S. 2007. A textual perspective on Givon's quantity principle. *Journal of Pragmatics* **39**(2): 292–304.

Kabadjov, M., Atkinson, M., Steinberger, J., Steinberger, R., and Van Der Goot, E. 2010. NewsGist: a multilingual statistical news summarizer. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, Athens, Greece, pp. 591–4.

Khan, A. U., Khan, S., and Mahmood, W. 2005. MRST: a new technique for information summarization. *The Second World Informatika Conference*, Prague, Czech Republic, pp. 249–52.

Kuo, J.-J., and Chen, H.-H. 2008. Multidocument summary generation: using informative and event words. *ACM Transactions on Asian Language Information Processing* **7**(1): 1–23.

Lehmam, A. 2010. Essential summarizer: innovative automatic text summarization software in twenty languages. In *Proceedings of the Adaptivity, Personalization and Fusion of Heterogeneous Information Conference*, Paris, France, pp. 216–7.

Lerman, K., Blair-Goldensohn, S., and McDonald, R. 2009. Sentiment summarization: evaluating and learning user preferences. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, pp. 514–22.

Li, S., Ouyang, Y., Wang, W., and Sun, B. 2007. Multi-document summarization using support vector regression. *In Proceedings of the Document Understanding Workshop*, Rochester, New York USA.

Lin, C. Y. 1997. *Robust Automated Topic Identification.* PhD thesis, University of Southern California, Los Angeles, SC, USA.

Lin, C.-Y., and Hovy, E. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Conference on Computational Linguistics*, Saarbrcken, Luxembourg, Germany, pp. 495–501.

Lin, C.-Y., and Hovy, E. 2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology Conference*, Edmonton, Canada, pp. 71–8.

Litvak, M., Last, M., and Friedman, M. 2010. A new approach to improving multilingual summarization using a genetic algorithm. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp. 927–36.

Liu, F., and Liu, Y. 2009. From extractive to abstractive meeting summaries: can it be done by sentence compression? In *Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Singapore, pp. 261–4.

Lloret, E., and Palomar, M. 2009. A gradual combination of features for building automatic summarisation systems. In *Proceedings of the 12th International Conference on Text, Speech and Dialogue*, Pilsen, Czech Republic, pp. 16–23.

Lloret, E., and Palomar, M. 2010. Challenging issues of automatic summarization: relevance detection and quality-based evaluation. *Informatica. Special Issue on Computational Linguistics* **34**(1): 29–35.

Lloret, E., and Palomar, M. 2011a. Analyzing the use of word graphs for abstractive text summarization. In *Proceedings of the First International Conference on Advances in Information Mining and Management*, Barcelona, Spain, pp. 61–6.

Lloret, E., Ferrández, Ó., Muñoz, R., and Palomar, M. 2008a. Integración del reconocimiento de la implicación textual en tareas automáticas de resúmenes de textos. *Procesamiento del Lenguaje Natural* 183–90.

Lloret, E., Ferrández, Ó., Muñoz, R., and Palomar, M. 2008b. A text summarization approach under the influence of textual entailment. In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science in conjunction with the 10th International Conference on Enterprise Information Systems*, Barcelona, Spain, pp. 22–31.

Lloret, E., and Palomar, M. 2011b. Text summarisation in progress: a literature review. *Artificial Intelligence Review* **37**(1): 1–41.

Mani, I., and Maybury, M. T. 1999. *Advances in Automatic Text Summarization.* Cambridge, MA, USA: The MIT Press.

Mann, W. C., and Thompson, S. A. 1988. Rhetorical structure theory: toward a functional theory of text organization. *Text* **8**(3): 243–81.

Marcu, D. 1999. Discourse trees are good indicators of importance in text. In I. Mani and M. Maybury (eds.), *Advances in Automatic Text Summarization*, pp. 123–36. Cambridge, MA, USA: MIT Press.

McCargar, V. 2005. Statistical approaches to automatic text summarization. *Bulletin of the American Society for Information Science and Technology* **30**(4): 21–5.

Medelyan, O. 2007. Computing lexical chains with graph clustering. In *Proceedings of the Association of Computational Linguistics Student Research Workshop*, Prague, Czech Republic, pp. 85–90.

Mihalcea, R. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the Association of Computational Linguistics on Interactive Poster and Demonstration Sessions*, University of Michigan, Michigan, USA, pp. 170–3.

Mitkov, R. 2003. *The Oxford Handbook of Computational Linguistics (Oxford Handbooks in Linguistics S.).* Oxford, UK: Oxford University Press.

Mittal, V., Kantrowitz, M., Goldstein, J., and Carbonell, J. 1999. Selecting text spans for document summaries: heuristics and metrics. In *Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference*, Orlando, FL, USA, pp. 467–73.

Montiel Soto, R., and García-Hernández, R. A. 2009. Comparación de tres modelos de texto para la generación automática de resúmenes. *Procesamiento del Lenguaje Natural* **43**: 303–11.

Mori, T. 2002. Information gain ratio as term weight: the case of summarization of IR results. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan, pp. 1–7.

Morris, A. H., Kasper, G. M., and Adams, D. A. 1992. The effect and limitations of automatic text condensing on reading comprehension performance. *Information Systems Research* **3**(1): 17–35.

Nastase, V., Filippova, K., and Ponzetto, S. P. 2008. Generating update summaries with spreading activation. In *Proceedings of the Text Analysis Conference*, Gaithersburg, MD, USA, pp. 189–97.

Nenkova, A., Vanderwende, L., and McKeown, K. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, USA pp. 573–80.

Orăsan, C. 2009. Comparative evaluation of term-weighting methods for automatic summarization. *Journal of Quantitative Linguistics* **16**(1): 67–95.

Orăsan, C., Pekar, V., and Hasler, L. 2004. A comparison of summarisation methods based on term specificity estimation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal, pp. 1037–41.

Ou, S., Khoo, C. S. G., and Goh, D. H. 2007. Automatic multidocument summarization of research abstracts: design and user evaluation. *Journal of American Society for Information Science and Technology* **58**(10): 1419–35.

Plaza, L. 2011. *Uso de Grafos Semánticos en la Generación Automática de Resúmenes y Estudio de su Aplicación en Distintos Dominios: Biomedicina, Periodismo y Turismo*. PhD thesis, Universidad Complutense de Madrid.

Plaza, L., Díaz, A., and Gervás, P. 2008. Concept-graph based biomedical automatic summarization using ontologies. In *Proceedings of the 3rd Textgraphs Workshop on Graph-based Algorithms for Natural Language Processing*, Rochester, NY, USA, pp. 53–6.

Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Drabek, E., Lam, W., Liu, D., Otterbacher, J., Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A., and Zhang, Z. 2004. MEAD – a platform for multidocument multilingual text summarization. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, pp. 699–702.

Ramshaw, L. A., and Marcus, M. P. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*, Boston, MA, USA, pp. 82–94.

Saggion, H. 2008. SUMMA: a robust and adaptable summarization tool. *Traitement Automatique des Languages* **49**: 103–25.

Saggion, H., and Lapalme, G. 2002. Generating indicative-informative summaries with SumUM. *Computational Linguistics* **28**(4): 497–526.

Sarkar, K., and Bandyopadhyay, S. 2005. Generating headline summary from a document set. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, Mexico City, Mexico, pp. 649–52.

Schilder, F., and Kondadadi, R. 2008. FastSum: fast and accurate query-based multi-document summarization. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, Columbus, OH, USA, pp. 205–8.

Spärck Jones, K. 1999. Automatic summarizing: factors and directions. In I. Mani, and M. T. Maybury (eds), *Advances in Automatic Text Summarization*, pp. 1–14. Cambridge, MA, USA: MIT Press.

Steinberger, J., Jezek, K., and Sloup, M. 2008. Web topic summarization. In *Proceedings of the 12th International Conference on Electronic Publishing*, Toronto, Canada, pp. 322–34.

Svore, K. M., Vanderwende, L., and Burges, C. J. C. 2007. Enhancing single-document summarization by combining RankNet and third-party sources. In *Proceedings of the*

*Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech, pp. 448–57.

Tatar, D., Tamaianu-Morita, E., Mihis, A., and Lupsa, D. 2008. Summarization by logic segmentation and text entailment. In *Proceedings of Conference on Intelligent Text Processing and Computational Linguistics*, Haifa, Israel, pp. 15–26.

Teng, Z., Liu, Y., Ren, F., Tsuchiya, S., and Ren, F. 2008. Single document summarization based on local topic identification and word frequency. In *Proceedings of the Seventh Mexican International Conference on Artificial Intelligence*, Atizapán de Zaragoza, Mexico, pp. 37–41.

Tigelaar, A. S., Op Den Akker, R., and Hiemstra, D. 2010. Automatic summarisation of discussion fora. *Natural Language Engineering* **16**(02): 161–92.

Van Dijk, T. 1980. *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*. Hillsdale, NJ, USA: Lawrence Erlbaum.

Wan, X. 2008. Using only cross-document relationships for both generic and topic-focused multi-document summarization. *Information Retrieval* **11**(1): 25–49.

Wong, K.-F., Wu, M., and Li, W. 2008. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester, UK, pp. 985–92.

Yu, J., Reiter, E., Hunter, J., and Mellish, C. 2007. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering* **13**(1): 25–49.

Zajic, D. M., Dorr, B. J., and Lin, J. 2008. Single-document and multi-document summarization techniques for email threads using sentence compression. *Information Processing & Management* **44**: 1600–10.

Zhang, J., and Fung, P. 2009. Active learning of extractive reference summaries for lecture speech summarization. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: From Parallel to Non-parallel Corpora*, Singapore, pp. 23–6.

Zhao, L., Wu, L., and Huang, X. 2009. Using query expansion in graph-based approach for query-focused multi-document summarization. *Information Processing and Management* **45**(1): 35–41.