Computer Standards & Interfaces xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Computer Standards & Interfaces

journal homepage: www.elsevier.com/locate/csi

Tackling redundancy in text summarization through different levels of language analysis

Elena Lloret *, Manuel Palomar

Research Group on Natural Language Processing and Information Systems, Department of Software and Computing Systems, University of Alicante, Apdo. de correos 99, E-03080 Alicante, Spain

ARTICLE INFO

Available online xxxx

Keywords:
Text summarization
Redundancy detection
Natural Language Processing
Information access

ABSTRACT

One of the main challenges to be addressed in text summarization concerns the detection of redundant information. This paper presents a detailed analysis of three methods for achieving such goal. The proposed methods rely on different levels of language analysis: lexical, syntactic and semantic. Moreover, they are also analyzed for detecting relevance in texts. The results show that semantic-based methods are able to detect up to 90% of redundancy, compared to only the 19% of lexical-based ones. This is also reflected in the quality of the generated summaries, obtaining better summaries when employing syntactic- or semantic-based approaches to remove redundancy.

1. Introduction

Nowadays, the vast amount of information available has resulted in a disadvantage rather than an advantage, since users cannot cope with all the information, and the existing tools may be not suitable for their specific needs and purposes. This brings great challenges for different aspects of the current society, such as health [34], business [42], administrative procedures [4], or education [44]. This and other fields need to be improved with the help of automatic tools and procedures which make the analysis, processing and interpreting of the information and data more effective and efficient. The potential application of text summarization (TS) for facilitating information access and helping users to easily manage large amounts of data has reflected in the numerous approaches proposed by the research community in recent years. This task allows users to obtain a brief fragment of text that conveys the essential and most relevant information from a larger one [53]. However, to produce good summaries automatically is still a big challenge.

For instance, redundant information, temporal dimension or coreference resolution are issues which have to be taken into consideration especially when summarizing a set of documents (multi-document summarization), thus presenting this task even greater challenges [22]. In particular, it is worth mentioning that multi-document summarization systems have a crucial role in the management of such information from different perspectives: i) providing users with condensed versions of texts which contain the essence of a range of documents, and serve as substitutes of the original ones (informative summaries); ii) addressing a specific information need expressed in a query (query-focused summaries) or by means of opinions (opinion-oriented summaries); or iii) simply helping people to decide whether it is worth accessing and reading a whole document or not (indicative summaries). These are only a few examples of the most common and well-known kinds of summaries, but more types can be found in [52,26,37].

Recent efforts have been concentrated on multi-document summarization, for instance, [3,62], or [33], as it has turned out an essential task since large amounts of documents have to be dealt with. This task is also much more complex than summarizing a single document, and this difficulty arises from the existing diversity of topics within a large set of documents. A good summarization technology aims to combine the main topics with completeness, readability, and conciseness. According to [22], the main differences between single- and multi-document summarization concern: a) the degree of redundancy, which is much higher than in single-document summarization; b) the temporal dimension of the documents; c) the compression rate, which will typically be much lower for collections of related documents than for single-document summaries; and d) the coreference problem across documents. Since 2003, international evaluation forums, such as the Document Understanding Conferences[1] (DUC) or the Text Analysis Conferences[2] (TAC) have only addressed guidelines for multi-document summarization tasks.

Motivated by the current need toward the generation of high quality summaries, and their potential use and application for facilitating information access in human-computer interaction, it is crucial to investigate and explore different factors that can be widely classified in the following groups: 1) methods that help to detect relevant

* Corresponding author.
E-mail addresses: elloret@dlsi.ua.es (E. Lloret), mpalomar@dlsi.ua.es (M. Palomar).

[1] http://duc.nist.gov/
[2] http://www.nist.gov/tac/

0920-5489/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csi.2012.08.001

fragments of information; 2) techniques that go beyond the simple extraction of sentences; 3) approaches that consider the linguistic quality of summaries; and 4) approaches that deal with redundant information. Redundancy is a well-known problem when dealing with TS, in the sense that a summary should avoid giving a piece of information more than once; otherwise, information will be repeated, affecting the quality of the final summaries and introducing noise. Moreover, there are multiple levels of language processing when humans produce or understand language, which comprise phonology, morphology, lexical, syntactic, semantic, discourse and pragmatics [39]. However, current Natural Language Processing (NLP) systems mainly deal with the lower levels of processing (lexical, syntactic and semantic), partly due to the difficulty associated with the interpretation at the highest levels.

Many different techniques have been proposed to tackle the redundancy problem, and TS systems use a wide range of them to avoid incorporating repeated information in the final summary. To the best of our knowledge, there is no previous study that analyzes the different levels of language analysis (lexical, syntactic and semantic) in the context of TS for removing redundancy. Therefore, the first contribution of this paper is to quantify the influence of the proposed levels of language analysis for detecting redundancy. Specifically, each level is represented by a well-known method: cosine similarity for the lexical level, textual entailment for the syntactic, and sentence alignment for the semantic one. The second contribution is to analyze the usefulness of such methods within the TS task, determining their positive and negative effects, and thus exploring them. On the one hand, redundant information has always been considered as a negative factor for TS, and as a consequence, most approaches remove repeated information at a first stage. On the other hand, redundant information can be also exploited to produce summaries, assuming that the more the information is repeated across documents, the more important it is, following the idea of [7]. This double-perspective for tackling redundancy is analyzed within the scope of this paper, and a comparison with existing TS systems is also provided.

This paper is organized as follows. Section 2 gives an overview on previous work on TS, focusing on multi-document summarization, and how different approaches have been investigated to deal with redundancy in summaries. In Section 3, the methods analyzed in order to detect redundant information across documents are explained in detail. In Section 4, the twofold use of redundant information is dealt with: first, to prevent it from appearing in the summary; and second, to determine potential relevant information within a document. The experiments together with the evaluation are presented in Section 5, and finally some relevant conclusions are drawn and further work is explained in Section 6.

2. Related work

As far as TS is concerned, much effort has been devoted to automatically identify relevant content within a document (or a set of documents) in order to select which sentences should appear in the summary, producing extracts as a consequence.

A wide range of techniques have been widely explored, such as statistical features [10]; linguistic models [23]; graph-based algorithms [29,38,45]; or machine learning techniques [1]. All these methodologies produce extracts by selecting a set of candidate sentences, and very few of the approaches include deeper language analysis, such as language generation or sentence compression techniques in the summarization process, in order to generate abstracts. Examples of abstractive approaches can be also found in the literature, where methods vary from sentence compression [18]; sentence fusion [16]; abstractive operations to convert extracts into abstracts [54]; or introducing language generation for building abstracts [19]. Approaches that even attempt to predict the structure of abstracts using machine learning algorithms are found in [49].

Over the years, redundancy has been an important issue that has been considered by the summarization research community. One of the quality criteria evaluated in the most important workshops and forums, like DUC or TAC is the amount of redundant information a summary has [14,15]. Obviously, the more repeated information a summary includes, the less readable it is, and consequently its quality would be highly affected. When producing a summary from a set of documents, the first issue that many systems consider is redundancy, and this problem is tackled using different methods and approaches. The cosine similarity has been widely employed to measure the level of overlapping between a pair of sentences, with the objective of avoiding redundancy [47,59,55,51]. However, as it is stated in [25], cosine similarity is limited because it performs only a very low level of language analysis (lexical) and as a consequence, it only covers redundancy in its most trivial form (identical words). In contrast to this well-known approach, in [25] redundancy is tackled from a semantic point of view, by detecting semantic similarity, using a word alignment tool and a sentence similarity score based on the inverse document frequency weight. This way the amount of semantic overlap between sentences is measured. However, they were not able to show significant effect of the suggested method, obtaining results lower than when using the cosine similarity method. The explanation given for such results was, on the one hand, the lower performance of the word aligner tool, and on the other hand, the amount of semantic content that most of the sentences of the corpus shared, which was identical to the number of words, being the cosine similarity sufficient in these cases. Another well-known approach to detect redundancy in texts is the *Maximal Marginal Relevance* (MMR) algorithm [9], which attempts to maximize the similarity between a passage and a query, but minimizing at the same time its similarity with already selected passages. This algorithm has been used and adapted in many multi-document summarization systems, such as [24,60,64], and it has been proven to obtain better performance than cosine similarity [61]. However, its main limitation is that it also analyzes the language at a lexical level, employing the cosine similarity to account for the degree of shared content between sentences and the query.

Apart from these techniques, other less employed similarity approaches can be also found. For instance, a basic approach for detecting similar information in texts by means of n-gram overlap between text units is explained in [50], whereas in [56], the whole summary is considered as a baseline first, and an aggregate similarity value based on the cosine similarity is computed against any individual sentence to remove repeated terms. Other approaches not as common as the aforementioned ones, can be found in [13] or [5], where a pivoted QR-matrix decomposition approach similar to the one described in [11] is employed to select a subset of the top scoring sentences, therefore minimizing redundancy in summaries. The pivoted QR approach was also used in conjunction with Latent Semantic Indexing (LSI), as described in [12]. In [27], diversity penalty [63] is considered as a kind of redundancy removal approach that penalizes the sentences to select according to their similarity with the already selected ones.

In more recent approaches, such as [57,31], the task of Textual Entailment (TE) has been successfully combined with summarization, in order to remove redundant information. However, TE has been only applied to single-document summarization and only in [30] a preliminary analysis for multi-document summarization was conducted. Other approaches employ integer linear programming for extracting relevant that do not contain the same information [2].

In contrast to the aforementioned approaches, where redundancy is detected, and consequently redundant information is removed from texts, in [6] it is claimed that redundancy can be exploited to identify relevant information, and this idea is applied to the process of sentence fusion. Other approaches that consider the repetition of information as an indicator of importance can be also found in [48,7] and [40]. Moreover, the identification of repeated information is a central issue in [8], where a summary has to be generated from

Please cite this article as: E. Lloret, M. Palomar, Tackling redundancy in text summarization through different levels of language analysis, Computer Standards & Interfaces (2012), http://dx.doi.org/10.1016/j.csi.2012.08.001

a set of user reviews. In order to determine the most relevant properties of these topics, their multi-document summarization method only selects properties that are stated by a plurality of users, thereby eliminating rare and/or erroneous opinions. Also in [35], redundant information is used to assess the representativeness of an extracted opinion. As can be seen, it seems that with the birth of new textual genres, such as blogs or reviews, the role of redundant information is changing, and while in traditional TS redundant information is often discarded, in opinion-oriented summarization one wants to track and report the degree of redundancy, since in this context the user is typically interested in the number of times a given sentiment is expressed in the corpus [43].

Despite these two uses of redundant information, most research has been focused on applying redundancy techniques for generating summaries without repeated information. However, as it was previously mentioned, different approaches for detecting redundancy across documents have been used, and there is not an in-depth study of the effects they have concerning the level of language analysis they employ, and how they influence the TS task. Only in [41] a comparison between three different redundancy techniques was carried out. These techniques were WordNet distance, cosine similarity and LSI. From this analysis, it was concluded that WordNet distance obtained a poor performance, but Cosine Similarity and LSI performed practically identical. However, the cosine similarity, although a common-used method for detecting redundant information, may not be the most appropriate because it only analyzes language at a lexical level. It would be also very interesting to quantify the improvement that other methods, which use higher levels of language analysis, such as syntactic or semantic, can obtain over it. In the comparison conducted in this paper, the cosine similarity is also taken into account for the lexical level, as it is one of the most popular redundancy detection methods in NLP, but in contrast, two novel approaches under the same experimental setup will be explored from different levels of language analysis: textual entailment and sentence alignment for the syntactic and semantic ones, respectively. Furthermore, these methods are integrated into two TS approaches with the purpose of analyzing and comparing the two roles that redundant information may have in a set of topic-related documents, and the comparison is extended with several settings of the MEAD[3] system [46]. MEAD can be configured to choose among two redundancy detection methods: cosine similarity or MMR. Concerning the summarization process, it employs features that rely on sentence position and similarity with the first sentence or the centroid in order to account for the relevance of each sentence. Then, it extracts the top scored ones, thus producing an extractive summary. Moreover, it also includes a LEADBASED and a RANDOM baseline. In the first one, summaries are produced by selecting the first sentence of each document, then the second sentence of each, etc. until the desired summary size is reached, whereas the latter selects random sentences from documents.

3. Redundancy detection approaches

The objective of this section is to present three methods for detecting redundant information: cosine similarity, textual entailment and sentence alignment.

The reason why these methods have been selected for our research is, on the one hand, the different types of knowledge they employ, and, on the other hand, their popularity among the NLP community. The final goal is to integrate them into a TS approach, and carry out a comparison and an in-depth analysis of their benefits and limitations. In the following subsections, each suggested method is explained in detail.

3.1. Lexical-based redundancy detection

Cosine similarity (CoSim) is a common vector-based similarity measure.

Given two documents d_1, d_2 in a feature vector representation, one way to define how similar they are is by computing the cosine of the angle θ that d_1 and d_2 form in the feature space, as it is shown in Formula 1.

$$sim(d_1, d_2) = cos(\theta) = \frac{\sum_i d_1(i) d_2(i)}{\sqrt{\sum_i d_1(i)^2 \sum_i d_2(i)^2}} \quad (1)$$

As it has been previously said in Section 2, this method has some limitations due to the fact that it only relies on word overlapping. However, since it has been widely used in the literature, it is very interesting to check whether the CoSim is really effective or not. For this reason, it has been analyzed in the scope of this paper, following this approach:

1. Split all documents in a set of sentences;
2. Compute the cosine similarity between one sentence and all of the remaining ones; and
3. Consider as redundant sentences those ones whose cosine similarity is above a specific threshold.

The CoSim value is computed using the publicly available *Text::Similarity* package.[4] A threshold of 0.7 was empirically established, meaning that a sentence whose cosine similarity with respect to another sentence is higher than 0.7 is considered to be redundant. On the one hand, stricter threshold values lead to the detection of hardly redundant sentences, whereas on the other hand, with lower values, too much information is considered as redundant, which is not totally true.

The final set of sentences after processing the CoSim is reduced with regard to the initial one, distinguishing between redundant and non-redundant information. These sets of sentences are the starting point for a TS approach later described.

3.2. Syntactic-based redundancy detection

Textual Entailment (TE) is an NLP task which attempts to detect entailment relations between pairs of sentences. An entailment relation involves determining whether the meaning of one text snippet (the hypothesis) can be inferred from another one (the text) [21]. To successfully achieve this objective, several approaches have been proposed, being the Recognizing Textual Entailment Challenges (RTE)[5] the most referred sources for determining which methodologies are the most relevant.

TS and, in particular, redundancy detection can take advantage of this task, in the sense that TE can be computed between pairs of sentences to determine if one sentence can be deduced from another, thus meaning the information from one sentence is already contained in the other. Here, the approach suggested follows the idea proposed in [31], where entailment relations are computed first, and in case a positive entailment is found, the second sentence is considered to be redundant, and as a consequence, discarded.

A similar version of the TE engine described in [17], achieving a performance of around 63%, is used for carrying out all the experiments and evaluation. This system mainly relies on syntactic information, and then a SVM classifier is trained in order to make the final decision on establishing a true or false entailment.

By computing the entailment relations within a set of documents, repeated information across documents can be identified. However,

[3] www.summarization.com/mead/

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

[5] http://www.nist.gov/tac/2009/RTE/

Please cite this article as: E. Lloret, M. Palomar, Tackling redundancy in text summarization through different levels of language analysis, Computer Standards & Interfaces (2012), http://dx.doi.org/10.1016/j.csi.2012.08.001

due to its computational cost, computing the entailment relations in the same way as for cosine similarity would be very costly computationally. For this reason, the entailment relations between an incremental set of sentences (considered as the text) and a single sentence (considered as the hypothesis) is calculated to build a set of non-redundant sentences. Next, an example to clarify this explanation is provided.

Let's suppose that we have a document D_1 with four sentences $S_1S_2S_3S_4$. The entailment calculation is performed as follows:[6]

$Sent_{nr} = \{S_1\}$

$Sent_{nr} \rightarrow entails \rightarrow S_2 \Rightarrow NO$

$Sent_{nr} = \{S_1, S_2\}$

$Sent_{nr} \rightarrow entails \rightarrow S_3 \Rightarrow NO$

$Sent_{nr} = \{S_1, S_2, S_3\}$

$Sent_{nr} \rightarrow entails \rightarrow S_4 \Rightarrow YES$

$Sent_{nr} = \{S_1, S_2, S_3\}$

In this example, only one redundant sentence (S_4) is detected, and therefore, it is discarded from the set. Once a whole document has been processed, the process continues with the remaining documents in the same way, but starting from the result of the previous one, so in the end the sentences which are not redundant are identified. Another set of redundant sentences is also obtained. This constitutes the starting point for the summarization approaches.

The advantage of using TE over CoSim is that TE takes also into account syntactic information, in contrast to the lexical similarity only used in CoSim. In contrast, the drawback of this method concerns the direction of the entailment computation, since a true entailment in one direction (e.g. $S_i \rightarrow S_j$) does not necessary mean the opposite ($S_j \rightarrow S_i$). In such situations, the entailed sentence may give additional information and we will not take it into account. This shortcoming could be easily solved by computing the entailment relations in both directions, or even extend the same idea using a paraphrase recognition system instead.

3.3. Semantic-based redundancy detection

The idea of using sentence alignment to detect redundancy has been inspired from [25], where a word aligner tool together with a similarity score based on the inverse document frequency and a semantic knowledge resource (i.e., WordNet) was used to measure the amount of semantic overlap between sentences. In this sense, sentence alignment (SentAlign) has the advantage of relying on semantic knowledge, and for this reason, it may be a good approach for detecting redundant information across documents. Therefore, we adopt a similar approach, and we compute sentence alignment at a document level among a set of related documents, using the publicly available *Champollion Tool Kit*[7], which has been proven effective for aligning parallel text in a robust way. Given a set of documents, the alignment between D_i and D_{i+1} is computed, where i refers to the i^{th} document in the set[8]. As a result, we will have a set of equivalent sentences in both documents, in the form of:

$S_{j,i} <==> S_{k,i+1}$

where different cases can be found:

$S_{j,i} <==> S_{k,i+1}$

$S_{j+1,i}, S_{j+2,i} <==> S_{k+1,i+1}$

$omitted <==> S_{k+2,i+1}$

$S_{j+3,i} <==> omitted$

Table 1 shows the generalized cases together with their corresponding actions. The action in the last row of the table tries to maintain the coherence when selecting sentences, taking into account to which document the previous selected sentences pertain (e.g., let's imagine we obtain $5<==>8$ and the last selected sentence from the first document is sentence 3, whereas it is sentence 7 for the second document. In this case, we will select the sentence from the second document, i.e., sentence 8). Moreover, as can be noticed, in each alignment, the lowest number of sentences is kept. This process is repeated for each pair of documents, and the corresponding results are considered as a new document, so the alignment process is repeated between this new resulting document and the next one. Finally, two different set of sentences – one for non-redundant sentences, and another for the redundant ones – are ready to be processed by a TS approach.

Similarly to TE, it is hypothesized that this technique will perform significantly better than cosine similarity, because semantic information is also taken into account.

4. Text summarization approach

In Section 3, three different methods to detect redundant information were described: i) cosine similarity; ii) textual entailment; and iii) sentence alignment.

In order to prove the performance of each proposed method and analyze their benefits and limitations, the objective of this section is to explain how they can be integrated into a TS approach. The core of this TS approach is similar to the one described in [32], which relies on statistical and linguistic features to detect relevant content in documents. On the one hand, redundant information is discarded from documents, as this is the most common approach adopted by TS systems. This way, the insertion of repeated content in the final summary is avoided. On the other hand, redundancy across documents is employed to detect also relevant information.

The suggested TS approach receives different inputs, depending on the redundancy detection method employed and the use of the repeated information. However, in both cases, the purpose is to generate non-redundant summaries.

4.1. Discarding redundant information for text summarization

In this approach, redundancy is considered as a negative factor that affects the quality of summaries. As a consequence, redundant sentences identified by one of the proposed methods (cosine similarity, textual entailment and sentence alignment) are removed from documents, having in the end a single set of non-redundant sentences that is the input for the summarization approach.

The whole approach is depicted in Fig. 1. As can be seen, it has two stages: i) redundancy detection, and ii) summarization. Firstly, the set of documents are passed through the redundancy detection module, which can be set up with one of the proposed methods. Once the redundant sentences have been removed from the set of original documents, the remaining non-redundant ones are inputted to the summarization approach, which focuses on detecting relevant content using statistical (term-frequency -TF-) and linguistic features (the code quantity principle -CQP- [20]). More detail about the

[6] $Sent_{nr}$ is the set of non-redundant sentences.

[7] http://champollion.sourceforge.net/

[8] For clarification on the explanation, we will called D_i and D_{i+1}, first and second document, respectively.

Please cite this article as: E. Lloret, M. Palomar, Tackling redundancy in text summarization through different levels of language analysis, Computer Standards & Interfaces (2012), http://dx.doi.org/10.1016/j.csi.2012.08.001

Table 1
Rules after the sentence alignment process.

Case	Action
A sentence of the first document cannot be aligned with any sentence of the second document (which appears as "omitted")	Keep the sentence of the first document
A sentence of the second document cannot be aligned with any sentence of the first document (which appears as "omitted")	Keep the sentence of the second document
A sentence of the first document is aligned with several sentences of the second document	Keep the sentence of the first document
A sentence of the second document is aligned with several sentences of the first document	Keep the sentence of the second document
A sentence of the first document is aligned with exactly one sentence of the second document	Keep either the sentence of the first or second document, taking into account from which document comes the last kept sentence

summarization approach can be found in [32]. Lastly, the most important sentences are selected and extracted, leading to a final summary, whose length has been already predefined[9].

Fig. 2 shows three examples of generated summaries using the previously mentioned redundancy detection methods. The summaries correspond to cluster *d109h* of the DUC 2002 dataset. This set of documents is about a natural disaster (flooding) occurred in China. As a general overview, it can be seen that the summary which was generated using the cosine similarity to identify repeated information is the poorest one regarding its content. With respect to the other two, it can be seen that they cover different aspects of information about the same topic. The summary generated using textual entailment focuses more in giving a general perspective of the flooding, whereas the one generated using sentence alignment contains statements done by politics.

4.2. Exploiting Redundant Information for text summarization

In contrast to the previous approach where redundancy is considered a negative factor, the underlying assumption for this approach goes in the opposite direction. Taking as a premise that most frequent words can indicate relevance in a document, which has been proven to work successfully for computing the salience of sentences in summarization [36,40,58], the idea behind this approach is in light of this claim. In this paper, it is hypothesized that if a sentence or a piece of information is repeated across different documents, this may mean that such information is relevant and should be worth keeping it in the summary.

Fig. 3 shows graphically the summarization process. Again, the process can be mainly divided in two steps: a redundancy detection step first, and then a summarization step. It is worth mentioning that this approach is very similar to the one presented in Section 1, but as can be seen, the set of potentially redundant sentences is kept for further processing, being the input for the summarization stage. The associated problem with this lies in the fact that this set is very redundant, so when generating the summary, a lot of repeated information can be introduced. In order to prevent this from happening, before inputting the set of redundant sentences to the summarization process, an additional stage is added. It involves detecting redundancy again, in this case using textual entailment, in the same way it was described in Section 3. This will lead to the selection of the most important sentences of a document, but avoiding at the same time incorporating redundant information in the final summary. The reason why TE was chosen as the method for this double-redundancy detection verification is because, as it was reported in [25], on the one hand, the cosine similarity computes the similarity only at a lexical level, and this may be not sufficient for detecting redundancy. On the other hand, their proposed method based on word alignment has not been yet shown to improve cosine similarity. In contrast, TE has been proven to work successfully for detecting redundancy in documents [30]. After this double-redundancy checking process, the most relevant sentences are identified in the same way as for the previous approach, and finally the summary is generated.

Analogous to the examples of the summaries aforementioned, Fig. 4 depicts examples of generated summaries taking into account redundancy to identify relevant information. In order to allow the comparison of this approach with the previous one, the summaries shown correspond to the same cluster of documents of DUC 2002 dataset (i.e. *d109h*). As can be seen, the first summary does not reach the desired length[10], and consequently it is very short. As will be explained in the next section, this happens because the cosine similarity detects very little redundancy, leading to a reduced pool of redundant sentences for the summarization approach. The remaining summaries, although different in content, have both quite good quality. In the next section, an in-depth analysis of the evaluation and the results obtained will be carried out. Therefore, the differences concerning the performance of each method, as well as their performance within the TS approach will be understood.

5. Evaluation and discussion

The objective of this section is to test the proposed methods for detecting redundancy within their applicability in TS and, at the same time, assessing the benefits of the different levels of language analysis for this problem.

In order to provide a detailed evaluation of the different approaches, this section is structured in several subsections. Firstly, the corpora used for developing all the experiments are explained, as well as some details concerning the experimental setup (Section 1). Secondly, an analysis of the amount of redundancy detected by each of the proposed methods is outlined (Section 5.2). Then, two different types of evaluation have been conducted for the two approaches concerning the integration of the redundancy detection methods within TS. A human evaluation is carried out (Section 3) with the purpose of determining the performance of each summary with respect to the amount of redundant information it contains. Moreover, the overall content of the generated summaries is also evaluated automatically using ROUGE [28]. These results are explained in Section 4. For both types of evaluation, a comparison with different configurations of the MEAD system, as well as a baseline, is also reported, together with a detailed analysis and discussion.

5.1. Corpora and experimental setup

Three different data collections from the Document Understanding Conferences (DUC) were used. In particular, the datasets from 2002, 2003, and 2004 were taken into consideration. These data consisted of different clusters of related documents – 59, 60 and 50 – respectively, but each dataset contained at least 500 documents (10 documents per cluster on average). In total, we dealt with 1667 documents. The reason why we used the datasets from these years was due to the fact that since DUC 2005, the multi-document task focused on query-oriented summarization rather than generic one, and this paper deals with generic multi-document summarization.

[9] As it will be explained later, summaries of 100 words length were generated in all the experiments performed in this paper.

[10] All summaries were generated with a 100 word length.

Please cite this article as: E. Lloret, M. Palomar, Tackling redundancy in text summarization through different levels of language analysis, Computer Standards & Interfaces (2012), http://dx.doi.org/10.1016/j.csi.2012.08.001

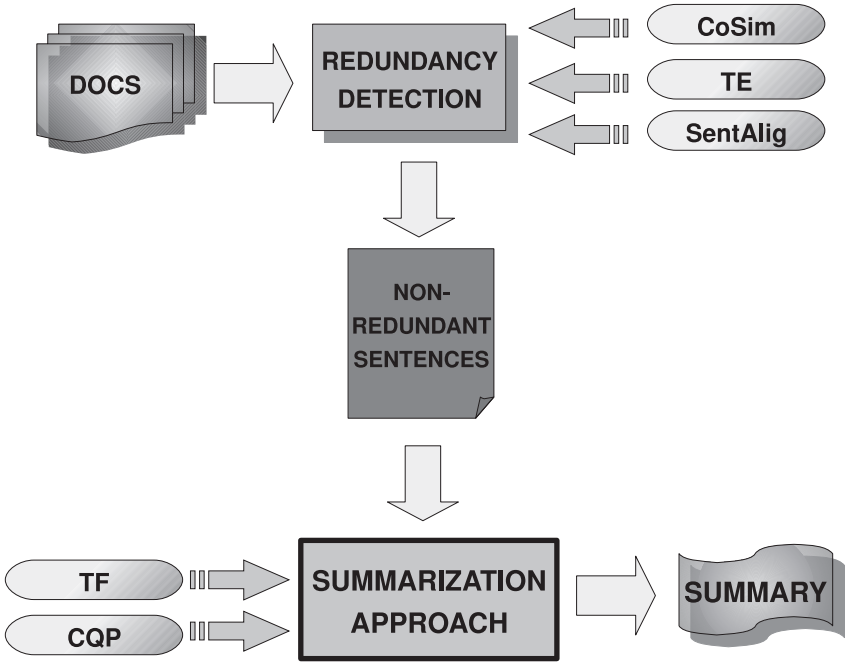

Fig. 1. Text summarization approach where redundancy is discarded.

Concerning the generation of summaries, the DUC guidelines were followed, and as a consequence, summaries of 100 words were generated for each set of documents. Moreover, the summaries were generated with respect to the two proposed uses of redundancy: first removing redundant information from texts with cosine similarity, textual entailment or sentence alignment; and second, taking advantage of the redundant information identified with the same methods. Once the different inputs were ready, they were passed through the summarization process as it was shown in Section 4. In addition, a baseline was also generated without taking into account any redundancy method (baselineCQP + TF), which will be used for comparison purposes. However, when carrying such kind of analysis, to have only one baseline for comparison is not enough. For this reason, we also ran the MEAD system over the same data with different configurations for tackling redundancy: the cosine similarity method (with a similarity threshold of 0.7[11]) and the MMR.

Moreover, we also test the two baselines which MEAD included, i.e., LEADBASED and RANDOM.

In total, 11 approaches were finally evaluated and analyzed (six different approaches for dealing with redundancy coupled with a TS approach, three baselines, and two configurations for the MEAD system).

5.2. Redundancy detection evaluation

Firstly, we analyze and quantify the amount of redundant sentences that each of the proposed methods is able to detect. Table 2 shows the results of this analysis. As can be seen, the amount of redundant sentences detected by the proposed methods is consistent across all the data of the different DUC forums, being the SentAlign method, which performs a semantic-level language analysis, the one which detects the highest number of redundant sentences (90% on average), whereas the CoSim technique, which is only lexical-based, is the one which detects the lowest (19% on average). The method relying on syntactic knowledge, TE, is able to identify that approximately 74% of the sentences are redundant. As far as these results are concerned, we could expect the lowest results for the summaries employing language analysis at a semanic level (SentAlign), because it may be too strict in detecting redundancy, and therefore, some important information could be deleted. However, in the approach that uses redundancy to identify important content in documents, the opposite could happen, since at this stage we deal with the 90% of the content of the documents.

In the next section, the results obtained will show whether these expectations can be confirmed or not.

5.3. Assessing redundancy in summaries

At the present moment, there are no tools able to automatically assess the degree of redundancy a summary contains. Tools such as ROUGE are good to provide a overall idea of how good a generated summary is with respect to a model one. However, they cannot specifically give any information about the amount of repeated information a summary contains. As a consequence, a manual evaluation is necessary to determine to what extent a summary contains redundant information or not. In this section, we propose this type of evaluation to analyze the proposed redundancy detection methods in the context of complete automatic summaries. In this manner, it would be possible to determine the benefits and limitations of the proposed redundancy detection methods when they are combined with a TS approach.

To carry out the human evaluation, a group of seven undergraduate and postgraduate students was given specific guidelines about how summaries should be evaluated. The same guidelines as in the evaluation of DUC conferences were followed, but focusing only on the non-redundancy aspect.

The focus of this evaluation was to determine the amount of repeated information in the generated summaries, so they were asked to provide a score in a five-point scale for each summary, according to the non-redundancy quality criteria: "Unnecessary repetition might take the form of whole sentence that are repeated, or repeated facts, or the repeated use of a noun or noun phrase when a pronoun would suffice." In order to specify more this issue, the participants were told to rate a summary according to this question:

[11] The same threshold was set up for our cosine similarity approach

Please cite this article as: E. Lloret, M. Palomar, Tackling redundancy in text summarization through different levels of language analysis, Computer Standards & Interfaces (2012), http://dx.doi.org/10.1016/j.csi.2012.08.001

a

As soon as they stepped out from the train, they set out for Changshayuan, the front line of the province's flood resistance.
I interviewed Governor Chen Bangzhu here on the spot.
Reporter How are you, Governor Chen?
You have inspected some flood-stricken areas in Changsha City this morning.
Will you please tell our audience something about the province's flood situation?
Governor This year, Hunan has been hit by a flood of a severity unseen for the past 100 years in terms of both water level and quantity.

b

BFN Text Changsha, June 19 XINHUA — This capital of central China's Hunan province has been hit by floods in the wake of a succession of rainstorms on the upper and middle reaches of the Xiangjiang River.
The water level of the river, which flows through Changsha, is still rising sharply, according to local sources.
By four o'clock this afternoon the level recorded at the Changsha monitoring station had risen to 38.93 m, some 0.56 m higher than the highest figure ever previously recorded.
The flood has caused serious losses to people along both banks of the river, especially in Changsha, according to the sources.

c

The same evening, Wen Jiabao and his party earnestly listened to reports by the Jiangxi Provincial Party Committee and government on Jiangxi's flood situation and its flood-fighting and relief work.
Wen Jiabao pointed out: At present, the Ganjiang River's water-level is declining, but the Jiangxi Provincial Party Committee and government are still facing very arduous tasks.
While continually and properly conducting flood-fighting and relief work along the Ganjiang River valleys, we should stress flood-prevention in the Poyang Lake area and should promptly organize the masses in the disaster-stricken areas to resume production so as to be self-reliant and to rebuild their homes.

Fig. 2. Examples of summaries discarding redundant information using: a) cosine similarity, b) textual entailment, and c) sentence alignment.

To what degree does the summary say the same thing over and over again?

1. Quite a lot; most sentences are repetitive
2. More than half of the text is repetitive
3. Some repetition
4. Minor repetitions
5. None; the summary has no repeated information

It is important to note that the summaries they were given did not contain any information about how they were generated, thus making the evaluation process totally blind. Once all the summaries were evaluated, each numeric ranking was mapped into a qualitative scale of three possible values: *very good*, *acceptable*, and *very bad*. The first one, *very good*, groups ratings 4 and 5; *acceptable* corresponds to value 3; and finally *very bad* is bound to values 1 and 2. At the end, each summary pertained to one of these three categories. The reason for doing this was that sometimes the barrier between one category and another was very fuzzy, so the initial fine-grained scale was mapped onto this new one.

A total of 11 approaches for three DUC year datasets (2002, 2003 and 2004) were evaluated, grouped in three categories (redundancy as a negative factor, redundancy as a positive factor, and a comparison with different configurations of MEAD). The results are presented with regard to the previously mentioned categories below.

5.3.1. Assessing redundancy in summaries when discarding redundant information for text summarization

Table 3 shows the results for the approach discarding redundancy from documents. As can be seen, the manual evaluation also consider TE and SentAling as more appropriate methods for detecting redundant information, since the summaries generated by either of these two approaches obtain higher number of *acceptable* and *very good* summaries than the ones produced by cosine similarity or without any redundancy technique. TE and SentAlign performs very similarly whereas cosine similarity differs a little from them. Moreover, the approach without taking into account any redundancy detection method (baselineCQP + TF) differs greatly from the approaches accounting for redundancy.

If no redundancy is detected, the percentage of *very bad* summaries reaches almost 15%. In contrast, for the approaches that employ a redundancy detection method, the percentage of *very bad* summaries

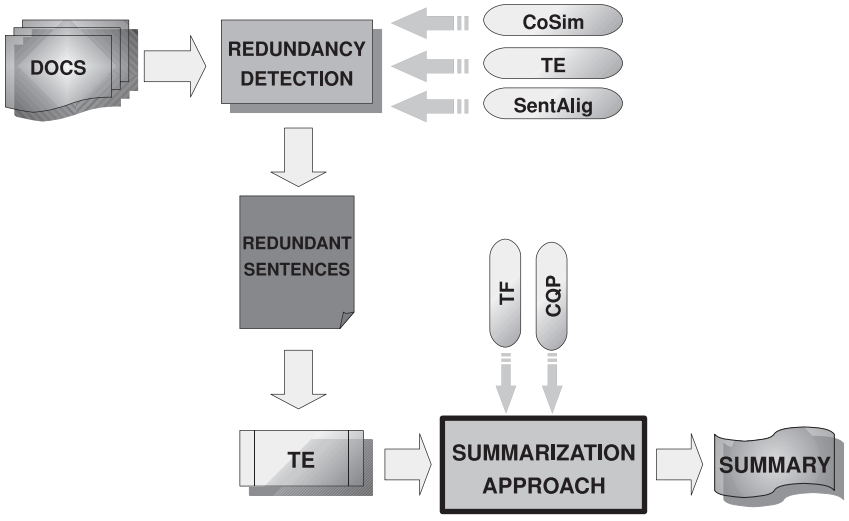

Fig. 3. Text summarization approach where redundancy is exploited.

Please cite this article as: E. Lloret, M. Palomar, Tackling redundancy in text summarization through different levels of language analysis, Computer Standards & Interfaces (2012), http://dx.doi.org/10.1016/j.csi.2012.08.001

a Major flood monitoring stations on the two rivers recorded their highest water levels, all four meters above the danger mark.

b About 118 towns of eight prefectures and cities along the Xiangjiang River were flooded.
According to him, 11 million people in the province's nine cities and 55 counties were affected, and more than 200 people died in the natural disaster, with 189,000 rooms destroyed and 1.2 million hectares of crops ruined.
According to the latest information offered by the Provincial Agricultural Bureau, a total area of 2.465 million mu of crops have been hit by the flood and of which, 446 000 mu of crops will have no yield, 1.957 million mu of grain have been heavily hit and some 300 million kilogram grain are expected to be lost.

c A flood peak occured in Bei Jiang in Guangdong's Sanshui at 0400 on 20 June, raising the water level to 10.39 meters -- the largest flood level ever recorded.
Major flood monitoring stations on the two rivers recorded their highest water levels, all four meters above the danger mark.
Disaster-relief groups have been sent to the flood-stricken areas striving for minimizing losses.
In the past, inadequate consideration has been given to urban flood control and the coordination of urban and rural flood control.
So far, more than 500,000 soldiers and local residents have moved in to help fight the floods.

Fig. 4. Examples of summaries taking advantage of the redundancy using: a) cosine similarity, b) textual entailment, and c) sentence alignment.

is only 3.6% for the cosine similarity method, and 1.1% for TE and SentAlign methods.

5.3.2. Assessing redundancy in summaries when exploiting redundant information for text summarization

Table 4 shows the results when redundancy is exploited to identify important content in texts. As can be seen from this table, results for TE and SentAlign are very good and quite similar as well. This confirms that they are both appropriate methods for detecting repeated information, and they can be used in conjunction with a summarization approach, having a positive influence in its quality regarding the non-redundancy aspect. What it is worth mentioning of this approach is the fact that cosine similarity does not perform better than the baseline, obtaining a percentage of very bad summaries of around 22%. This may be due to the fact that this method identifies too few sentences as redundant[12], according to the similarity threshold we established (please see Section 1 for more details).

5.3.3. Comparison with MEAD system

Table 5 shows the results of the manual evaluation of the summaries generated using different configurations of the MEAD system. On the one hand, concerning the MEAD system, both MEAD-cosine and MEAD-MMR achieve quite good performance (between 90% and 80%, respectively grouping *very good* and *acceptable* summaries). In this case, the cosine similarity performs better than MMR. On the other hand, comparing the MEAD cosine with respect to our proposed approaches, our cosine similarity approach when it is used to discard redundant information from documents obtains similar results, although the number of *very bad* summaries is slightly higher than in MEAD system. In contrast, when cosine similarity is used to exploit redundancy in text, the results are lower than with MEAD, due to the reasons explained above.

With regard to the MEAD baselines, the RANDOM baseline performs very good, and despite selecting sentences randomly, the corresponding summaries does not contain much redundant information.

5.3.4. Assessing redundancy in summaries: discussion

The manual evaluation carried out to determine what degree of redundancy a summary shows that either textual entailment or sentence alignment helps to avoid the incorporation of repeated information in summaries. They can be used instead of cosine similarity for this purpose, obtaining better results. From the results obtained in this type of evaluation, it can be seen how summaries not taking into account redundancy (baseline-CQP + TF) are not as good as the ones that do, being the percentage of *very bad* summaries reported more than ten times with respect to the best redundancy methods (TE and SentAlign).

The results obtained after the human evaluation show that both approaches, TE and SentAlign would be appropriate for detecting redundancy and thus, being integrated into a TS approach. That is reasonable, since after grouping redundant sentences we applied textual entailment as a redundant detection approach to eliminate the potential duplicated information.

Regarding the results for MEAD, using the data of DUC 2002, 2003 and 2004, cosine similarity and MMR performs quite similarly with respect to the percentage of summaries rated as *very good* or *acceptable*, although the MMR method obtained higher number of *very bad* summaries.

5.4. Content evaluation using ROUGE

ROUGE was used for evaluating the resulting summaries generated employing the redundancy detection methods, the proposed baselines, as well as the different configurations for the MEAD system.

This evaluation tool computes the number of overlapping n-grams of different lengths (unigrams, bigrams, the longest common subsequence, skip-bigrams) between an automatic summary and one or more model summaries. In particular, ROUGE-1, ROUGE-2 and ROUGE-L are measured taking into account unigrams, bigrams, and the longest common subsequence, respectively. To account for the significance of the results, a t-test is performed. The convention used for indicating the significance level in the tables with respect to the cosine similarity redundancy detection method is the following: $^{***} = p < .0001$, $^{**} = p < .001$, $^{*} = p < .05$ and no star indicates non-significance. The same significance levels, but with a † symbol are calculated with regard to the baseline approach without taking into consideration the repeated information. Next, the results for the different approaches are presented in different tables and a discussion is provided for each one, together with a general discussion at the end of this section, where the most important findings as well as the potential problems encountered are explained.

Table 2
Percentage of redundancy detected with the different methods.

% of redundancy detected				
Method	DUC-02	DUC-03	DUC-04	Mean
CoSim (lexical-based)	10.0	21.0	26.0	19.0
TE (syntactic-based)	71.0	76.0	74.0	73.6
SentAlign (semantic-based)	91.0	91.0	89.0	90.3

[12] It detects only 19% of the sentences as redundant.

Please cite this article as: E. Lloret, M. Palomar, Tackling redundancy in text summarization through different levels of language analysis, Computer Standards & Interfaces (2012), http://dx.doi.org/10.1016/j.csi.2012.08.001

Table 3
Manual evaluation when redundancy is discarded.

Approach	%	DUC-02	DUC-03	DUC-04	Mean
baselineCQP + TF	Very good	40.7	68.3	68.0	59.0
	Acceptable	32.2	21.7	24.0	26.0
	Very bad	25.4	10.0	8.0	14.5
CoSim + CQP + TF	Very good	78.0	81.7	66.0	75.2
	Acceptable	20.3	13.3	30.0	21.2
	Very bad	1.7	5.0	4.0	3.6
TE + CQP + TF	Very good	86.4	85.0	**98.0**	89.8
	Acceptable	10.2	15.0	2.0	9.1
	Very bad	3.4	0	0	1.1
SentAlign + CQP + TF	Very good	**93.2**	**95.0**	**98.0**	**95.4**
	Acceptable	6.8	1.7	2.0	3.5
	Very bad	0	3.3	0	1.1

Table 5
Manual evaluation for the MEAD system.

Approach	%	DUC-02	DUC-03	DUC-04	Mean
LEADBASED	Very good	17.0	13.3	30.0	20.1
	Acceptable	39.0	30.0	44.0	37.7
	Very bad	44.1	56.6	26.0	42.2
RANDOM	Very good	**76.1**	**76.7**	**86.0**	**79.6**
	Acceptable	22.0	23.3	12.0	19.1
	Very bad	1.7	0	2.0	1.2
MEAD-CoSim	Very good	55.9	73.3	36.0	55.1
	Acceptable	39.0	13.3	50.0	34.1
	Very bad	5.1	13.4	14.0	10.8
MEAD-MMR	Very good	55.9	61.7	38.0	51.9
	Acceptable	37.3	26.7	20.0	28.0
	Very bad	6.8	11.6	42.0	20.1

5.4.1. ROUGE results when discarding redundant information for text summarization

Table 6 shows the F-measure for the ROUGE-1 results for each of the datasets (DUC 2002, 2003 and 2004, respectively) when redundant information is first removed from documents. The last column shows the average results along the three datasets.

Against our expectations from the preliminary redundancy analysis, it can be observed that summaries generated employing the sentence alignment method generally obtain the highest results, whereas the cosine similarity the lowest. Furthermore, most of the results for textual entailment and sentence alignment of DUC 2002 and 2003 are significant with respect to it. Therefore, we can confirm that cosine similarity, which is based only on lexical overlapping, is not enough to detect redundancy to be subsequently applied to a TS process. However, textual entailment and sentence alignment carry out a deeper analysis of language, taking also into consideration syntactic and semantic information, and consequently, they have a positive influence on TS. Both approaches obtain similar results and the differences are not statistically significant, according to the t-test performed.

What we did not expect is that without dealing with redundancy at all, the baselineCQP + TF obtain quite close results to the other approaches. The reason why this may happen is because we use the ROUGE tool for evaluating summaries, which favors the number of common n-grams, no matter the number of times they appear in the summary. Therefore, if a summary contains redundant information, but the vocabulary matches the human summaries, the results would increase. However, as it was found out in the human evaluation, these summaries contain higher degree of redundancy than for instance the ones where TE has been taken into account.

From the results obtained it is worth stressing the fact that although TE and SentAlign discard a high number of sentences (around 74% and 90%, respectively), they obtain good results. This means that both methods do not remove sentences containing relevant content from documents. The last column in Table 6 also shows the average results for the ROUGE scores within the different methods according to the F-measure. Over the three years of DUC we focused on, it can be observed that the cosine similarity performs the poorest, whereas the results for TE and SentAlign are indeed very close. Again, it is confirmed that the results obtained by the method without dealing with redundancy are higher than expected and close to the other approaches where redundancy has been taken into account.

5.4.2. ROUGE results when exploiting redundant information for text summarization

The results concerning our second approach fall short of our expectations. ROUGE results are lower than in the previous approach. As far as Table 7 is concerned, it can be seen that in general terms, the CoSim is the poorest in performance, whilst TE and SentAlign obtain better results compared to it. In contrast, the proposed baseline outperforms all the approaches. Considering redundancy as a sign of relevant content within a set of documents has not been remained true in the experiments performed when evaluating the summaries with the ROUGE tool. In this approach, the number of sentences left after the redundancy detection stage is quite high for the TE and SentAlign methods (74% and 90%, respectively). In contrast, for the CoSim only a 19% of the sentences have been identified as redundant. However, after the second redundancy detection stage, the number of sentences decreases significantly, since TE is used to remove redundant information afterwards, having a small number of sentences to determine which are the most important to include in the summary. On average the number of sentences decreases from 15 sentences on average per document to only 6. This may be the reason of such low ROUGE performance. Regarding the comparison of the redundancy detection methods, it must be noted that TE and SentAlign performs significantly better than CoSim for most of the ROUGE values in DUC 2002 and DUC 2003.

Table 4
Manual evaluation when redundancy is exploited.

Approach	%	DUC-02	DUC-03	DUC-04	Mean
baselineCQP + TF	Very good	40.7	68.3	68.0	59.0
	Acceptable	32.2	21.7	24.0	26.0
	Very bad	25.4	10.0	8.0	14.5
CoSim + TE + CQP + TF	Very good	46.5	44.2	89.3	60.0
	Acceptable	14.0	32.7	8.5	18.4
	Very bad	39.5	23.1	2.2	21.6
TE + TE + CQP + TF	Very good	88.4	**100.0**	**94.0**	**94.1**
	Acceptable	10.2	0	6.0	5.4
	Very bad	3.4	0	0	1.1
SentAlign + TE + CQP + TF	Very good	**86.4**	75.0	80.0	80.5
	Acceptable	10.2	23.3	20.0	17.8
	Very bad	3.4	1.7	0	1.7

Table 6
ROUGE results when redundancy is avoided.

Approach	F-measure	DUC-02	DUC-03	DUC-04	Mean
baselineCQP + TF	Rouge-1	0.29653	0.28753	0.31089	0.29832
	Rouge-2	0.05212	0.05397	0.06298	0.05636
	Rouge-L	0.26051	0.25335	0.27626	0.26337
CoSim + CQP + TF	Rouge-1	0.26872	0.29231	0.29422	0.28508
	Rouge-2	0.03405	0.04840	0.04150	0.04132
	Rouge-L	0.23779	0.26025	0.26252	0.25352
TE + CQP + TF	Rouge-1	0.30137**	0.28977 †	**0.31091***	0.30068
	Rouge-2	0.05327**	0.05481 ††	0.06316**	0.05708
	Rouge-L	0.26373**	0.25399	0.27633	0.26468
SentAlign + CQP + TF	Rouge-1	**0.30621****	**0.30007** †	0.31047*	**0.30558**
	Rouge-2	0.05138**	0.05226	0.05499**	0.05288
	Rouge-L	0.26730**	0.25931	0.27148	0.26603

Please cite this article as: E. Lloret, M. Palomar, Tackling redundancy in text summarization through different levels of language analysis, Computer Standards & Interfaces (2012), http://dx.doi.org/10.1016/j.csi.2012.08.001

5.4.3. Comparison with MEAD system

Table 8 shows the results obtained for the MEAD system. It can be seen that the cosine similarity and the MMR perform similarly, so we could not prove that for the dataset used MMR was more appropriate than cosine similarity. It is worth noting that the best results are obtained with DUC 2004 dataset, and in this case, the cosine similarity performs higher than the MMR method, although the difference between them is not statistically significant, except for ROUGE-1. This may happen due to the nature of the documents, since the DUC 2004 dataset contains less noisy information than the other ones. Furthermore, the MEAD system uses different features from our text summarization approach, relying on positional and similarity features. Regarding the two proposed baselines, it can be seen that a random baseline obtains on average better results than for instance the MEAD system with MMR for tackling redundancy.

Regarding the comparison of MEAD with the proposed approaches previously explained, we are going to take into consideration only the approach with the highest results, that is when redundancy is first detected and removed from documents. It is worth mentioning that the results vary depending on the dataset. For instance, for DUC 2002 dataset, summaries generated using textual entailment and sentence alignment are significantly better than the MEAD methods with a 95% confidence. For DUC 2003, our approach using textual entailment and the one using sentence alignment are significantly better than MEAD-CoSim for ROUGE-2. Finally, for DUC 2004, both suggested methods using TE and SentAlign perform better than the MEAD cosine with a significant value of 0.001 for ROUGE-1 and ROUGE-2 and 0.05 for ROUGE-L. Concerning our approach using cosine similarity compared to the MEAD-CoSim, it performs significantly better than the latter for DUC 2002 dataset. In contrast, the MEAD-CoSim method performs better than ours for DUC 2003 and 2004 datasets.

5.4.4. Content evaluation using ROUGE: discussion

From all the results previously shown, it can be claimed that textual entailment as well as sentence alignment perform better than cosine similarity, since they go beyond the lexical similarity calculus, attempting also to exploit syntactic and semantic information in sentences. Among the two different perspectives that have been proposed in this paper for considering redundant information in documents, the traditional one of removing redundant information from texts seems to be more appropriate than the one taking redundant content as a relevant piece of information. According to ROUGE results, the comparison between cosine similarity and MMR method within the MEAD system shows that cosine similarity works better than MMR for these datasets.

Although focusing on the content, ROUGE may provide insights on how good our generated summaries perform comparing them to different model summaries. In this sense, it can be assumed that model summaries do not contain redundant information, and therefore, the more similarity between them (i.e., higher ROUGE scores), the better. However, it is also interesting to note that lower ROUGE results do not necessarily imply that the automatic summaries contain repeated information. In our case, this leads for instance to the fact that a baseline without taking into account redundancy could obtain higher ROUGE performance than other approaches that include some mechanisms to deal with the repeated information. However, as it was previously shown, when assessing this aspect, the results for these summaries are not as good, thus containing much more redundancy than others.

Table 7
ROUGE results when redundancy is taken into account.

Approach	F-measure	DUC-02	DUC-03	DUC-04	Mean
baselineCQP + TF	Rouge-1	**0.29653**	**0.28753**	**0.31089**	**0.29832**
	Rouge-2	0.05212	0.05397	0.06298	0.05636
	Rouge-L	0.26051	0.25335	0.27626	0.26337
CoSim + TE + CQP + TF	Rouge-1	0.19336	0.15043	0.27269	0.20549
	Rouge-2	0.03033	0.02217	0.05037	0.03429
	Rouge-L	0.17173	0.12933	0.23706	0.17937
TE + TE + CQP + TF	Rouge-1	0.28209**	0.28343**	0.30340	0.28964
	Rouge-2	0.03478	0.04243**	0.04131	0.03951
	Rouge-L	0.24322**	0.24563**	0.26435	0.25107
SentAlign + CQP + TF	Rouge-1	0.27509**	0.27965**	0.29609	0.28361
	Rouge-2	0.03436	0.04024**	0.04075	0.03845
	Rouge-L	0.23977**	0.24546**	0.25941	0.24821

Table 8
ROUGE results for the MEAD system.

Approach	F-measure	DUC-02	DUC-03	DUC-04	Mean
LEADBASED	Rouge-1	0.22369	0.21501	0.31948	0.25273
	Rouge-2	0.03580	0.03794	0.07174	0.04849
	Rouge-L	0.18738	0.18725	0.27636	0.21699
RANDOM	Rouge-1	**0.28000**	0.28194	0.30575	0.28923
	Rouge-2	0.04292	0.04322	0.07174	0.05263
	Rouge-L	0.24468	0.24764	0.27636	0.25623
MEAD-CoSim	Rouge-1	0.23018	**0.29494**	**0.34630*****	**0.29047**
	Rouge-2	0.04165	0.06266	0.08405	0.06279**
	Rouge-L	0.18860	0.25031	0.29454	0.24448**
MEAD-MMR	Rouge-1	0.23235	0.29067	0.33038	0.28447
	Rouge-2	0.04139	0.06075	0.07837	0.06017
	Rouge-L	0.19389	0.25113	0.28212	0.24238

6. Conclusions and future work

In this paper an analysis of three redundancy detection methods that employ different levels of language analysis (lexical, syntactic and semantic) and their influence within the text summarization task was presented. In particular, the proposed methods for tackling the redundancy problem were cosine similarity, textual entailment and sentence alignment, corresponding each of them to one of the different levels of language analysis previously mentioned, respectively. It has been shown that those methods that employ deeper language analysis at a semantic level are able to detect higher number of redundant sentences (90%). Furthermore, the detection of redundant information decreases for approaches using syntactic- (73%) or lexical-knowledge (19%). This means that processing the language at a semantic level may be more beneficial that relying only on the lexical or syntactic levels.

In order to prove the suitability of these methods within NLP tasks, text summarization was chosen, because in this task it is crucial to take into account redundancy in order to produce summaries that do not repeat the same information over and over again. This analysis was carried out by suggesting two different uses of redundancy. On the one hand, an approach where repeated information was discarded first, and then a summary was generated from the non-redundant information was suggested.

On the other hand, the redundant information was used to detect salient sentences.

The human evaluation performed, where a group of humans analyzed the generated summaries with respect to the degree of redundancy they contained, showed that both approaches for text summarization got very good results according to the redundancy aspect, and in all cases, the results reported for textual entailment and sentence alignment increased with respect to cosine similarity, showing again the appropriateness of syntactic- and semantic-based redundancy methods to be integrated within any text summarization approach.

It is worth mentioning that, although the syntactic- and semantic-based redundancy methods detected a higher number of repeated information than the lexical-based method, this did not affect the quality of the summaries. In this sense, when performing the content evaluation using ROUGE, the summaries generated with such methods obtained significantly better ROUGE scores with these methods than with cosine similarity. The comparison of such methods with different implementations of the well-known MEAD system showed that

Please cite this article as: E. Lloret, M. Palomar, Tackling redundancy in text summarization through different levels of language analysis, Computer Standards & Interfaces (2012), http://dx.doi.org/10.1016/j.csi.2012.08.001

syntactic and semantic approaches for detecting redundancy performed better on average (0.30558 vs. 0.29047, according to mean F-measure value of ROUGE-1 for all the datasets). This means that those methods analyzing language at a syntactic or semantic level are really discarding redundant information, and both are appropriate for detecting redundancy in text summarization.

Given that we use three standard datasets (i.e., DUC 2002, DUC 2003, and DUC 2004) and ROUGE as a tool for the content evaluation, it is worth comparing our results not only with the MEAD system, but also with the best participants for such editions. Having access to the results in such evaluations, the results obtained for the top performing system for ROUGE-1 are: 0.3515, 0.3798, 0.3823, for DUC 2002, 2003, and 2004, respectively. As can be seen, these results are better than the ones we reported. This may be happen because we are more focused in detecting redundancy rather than extracting important content, and our TS approach would need to incorporate more features. In addition, DUC participant results are also better than for MEAD system. However, concerning the human evaluation performed, it is worth noting that the same best DUC 2004 participant obtained the following results for the redundancy detection aspect[13]: 84% of the summaries were rated as "very good", 14% as "acceptable", and 2% as "very bad". Despite being good results, such results are very similar to the ones obtained for the cosine similarity and sentence alignment for DUC 2004 data, and what is more important is that our proposed TE method overperforms them, obtaining 94%, 6% and 0%, respectively.

Although it has been showed that the redundancy detection methods are appropriate for detecting redundant information, the main limitation encountered is the selection of relevant information, which is the other key aspect in text summarization. Therefore, in the future we plan to analyze to what extent the suggested redundancy detection approaches are suitable for improving our summarization strategies, for instance, graph-based approaches which have been shown to obtain good results. This manner, we could select and extract relevant information more precisely. It would be also very interesting for a short-term period to study whether the proposed redundancy detection methods can complement each other in some way according to the different levels of language analysis, and replicate the experiments for the text summarization approach which considers redundant information as an indicator of salience, but using sentence alignment to double-check redundancy instead of textual entailment. This would allow to determine whether this text summarization approach could be really effective for generating good summaries or not.

Acknowledgments

This research has been funded by the Spanish Government under the project TEXT-MESS 2.0 (TIN2009-13391-C04-01). Moreover, it has been also supported by Conselleria d'Educació — Generalitat Valenciana (grant no. PROMETEO/2009/119 and ACOMP/2010/286).

The authors would also like to thank Ester Boldrini, Helena Burruezo, Javi Fernández, Óscar Ferrández, José Manuel Gómez and Juanma Martínez for participating in the manual evaluation of summaries.

References

[1] A. Abuobieda, N. Salim, A. Albaham, A. Osman, Y. Kumar, Text summarization features selection method using pseudo genetic-based model, in: Proceedings of the International Conference on Information Retrieval Knowledge Management (CAMP), 2012, pp. 193–197.
[2] R.M. Alguliev, R.M. Aliguliyev, M.S. Hajirahimova, C.A. Mehdiyev, MCMR: maximum coverage and minimum redundant text summarization model, Expert Systems with Applications 38 (2011) 14514–14522.
[3] R.M. Alguliev, R.M. Aliguliyev, C.A. Mehdiyev, pSum-SaDE: a modified p-median problem and self-adaptive differential evolution algorithm for text summarization, Applied Computational Intelligence and Soft Computing 2011 (2011) 1–13.
[4] L. Álvarez Sabucedo, L. Anido Rifón, R. Míguez Pérez, J. Santos Gago, Providing standard-oriented data models and interfaces to egovernment services: a semantic-driven approach, Computer Standards & Interfaces 31 (2009) 1014–1027.
[5] M.R. Amini, N. Usunier, A Contextual Query Expansion Approach by Term Clustering for Robust Text Summarization, in: the Document Understanding Workshop (Presented at the *HLT/NAACL*), Rochester, New York, USA, 2007.
[6] R. Barzilay, K. McKeown, M. Elhadad, Information fusion in the context of multi-document summarization, in: Proceedings of ACL, 1999, pp. 550–557.
[7] R. Barzilay, K.R. McKeown, Sentence fusion for multidocument news summarization, Computational Linguistics 31 (2005) 297–328.
[8] S.R.K. Branavan, H. Chen, J. Eisenstein, R. Barzilay, Learning document-level semantic properties from free-text annotations, Journal of Artificial Intelligence Research 34 (2009) 569–603.
[9] J. Carbonell, J. Goldstein, The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries, in: ACM, New York, NY, USA, 1998, pp. 335–336.
[10] M. Chandra, V. Gupta, S. Paul, A statistical approach for automatic text summarization by extraction, in: Proceedings of the International Conference on Communication Systems and Network Technologies (CSNT), 2011, pp. 268–271.
[11] J. Conroy, D.P. O'leary, Text summarization via hidden Markov models and pivoted QR matrix decomposition, in: Technical Report, SIGIR, 2001.
[12] J.M. Conroy, D.P. O'Leary, J.D. Schlesinger, Classy Arabic and English multi-document summarization, in: Multi-Lingual Summarization Evaluation, 2006.
[13] J.M. Conroy, J.D. Schlesinger, Back to basics: CLASSY 2006, in: The Document Understanding Workshop (Presented at the *HLT/NAACL*), 2006.
[14] H.T. Dang, Overview of DUC 2006, in: The Document Understanding Workshop (Presented at the *HLT/NAACL*), Brooklyn, New York USA, 2006.
[15] H.T. Dang, K. Owczarzak, Overview of the TAC 2008 update summarization task, in: Proceedings of the Text Analysis Conference (TAC), 2008.
[16] M. Elsner, D. Santhanam, Learning to fuse disparate sentences, in: Proceedings of the Workshop on Monolingual Text-To-Text Generation, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 54–63.
[17] O. Ferrández, R. Muñoz, M. Palomar, Alicante University at TAC 2009: experiments in RTE, in: Proceedings of the Text Analysis Conference, 2009.
[18] K. Filippova, Multi-sentence compression: finding shortest paths in word graphs, in: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 322–330.
[19] P.E. Genest, G. Lapalme, Framework for abstractive summarization using text-to-text generation, in: Proceedings of the Workshop on Monolingual Text-To-Text Generation, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 64–73.
[20] T. Givón, Syntax: A Functional–Typological Introduction, II, John Benjamins, 1990.
[21] O. Glickman, 2006. Applied Textual Entailment. Ph.D. thesis. Bar Ilan University.
[22] J. Goldstein, V. Mittal, J. Carbonell, M. Kantrowitz, Multi-document summarization by sentence extraction, in: NAACL-ANLP 2000 Workshop on Automatic Summarization, 2000, pp. 40–48.
[23] P. Gupta, V. Pendluri, I. Vats, Summarizing text by ranking text units according to shallow linguistic features, in: Proceedings of the 13th International Conference on Advanced Communication Technology (ICACT), 2011, pp. 1620–1625.
[24] T. He, J. Chen, Z. Gui, F. Li, Ccnu at tac 2008: Proceeding on using semantic method for automated summarization, in: Proceedings of the Text Analysis Conference (TAC), 2008.
[25] I. Hendrickx, W. Daelemans, E. Marsi, E. Krahmer, Reducing redundancy in multi-document summarization using lexical semantic similarity, in: Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009), 2009, pp. 63–66.
[26] E. Hovy, C.Y. Lin, Automated multilingual text summarization and its evaluation, in: Technical Report. Information Sciences Institute, University of Southern California, 1999.
[27] F. Jin, M. Huang, X. Zhu, A query-specific opinion summarization system, in: 8th IEEE International Conference on Cognitive Informatics, 2009, pp. 428–433.
[28] C.Y. Lin, Rouge: a package for automatic evaluation of summaries, in: Proceedings of ACL Text Summarization Workshop, 2004, pp. 74–81.
[29] H. Lin, J. Bilmes, Multi-document summarization via budgeted maximization of submodular functions, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 912–920.
[30] E. Lloret, O. Ferrández, R. Muñoz, M. Palomar, Integración del reconocimiento de la implicación textual en tareas automáticas de resúmenes de textos, in: Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), 2008, pp. 183–190.
[31] E. Lloret, O. Ferrández, R. Muñoz, M. Palomar, A text summarization approach under the influence of textual entailment, in: Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008) in Conjunction with the 10th International Conference on Enterprise Information Systems (ICEIS 2008), 12–16 June, Barcelona, Spain, 2008, pp. 22–31.
[32] E. Lloret, M. Palomar, A gradual combination of features for building automatic summarisation systems, in: Proceedings of the 12th International Conference on Text, Speech and Dialogue (TSD), 2009, pp. 16–23.
[33] E. Lloret, L. Plaza, A. Aker, Multi-document summarization by capturing the information users are interested in, in: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, RANLP 2011, Organising Committee, Hissar, Bulgaria, 2011, pp. 77–83.
[34] M. López-Nores, Y. Blanco-Fernández, J.J. Pazos-Arias, J. García-Duque, The iCabiNET system: Harnessing Electronic Health Record standards from domestic

[13] We extracted the official results, and then we mapped them into the same rating scale than ours.

Please cite this article as: E. Lloret, M. Palomar, Tackling redundancy in text summarization through different levels of language analysis, Computer Standards & Interfaces (2012), http://dx.doi.org/10.1016/j.csi.2012.08.001

and mobile devices to support better medication adherence, Computer Standards & Interfaces 34 (2012) 109–116.
[35] Y. Lu, C. Zhai, Opinion integration through semi-supervised topic modeling, in: Proceeding of the 17th international conference on World Wide Web, 2008, pp. 121–130.
[36] H.P. Luhn, The automatic creation of literature abstracts, in: Inderjeet Mani, Mark Maybury (Eds.), Advances in Automatic Text Summarization, MIT Press, 1958, pp. 15–22.
[37] I. Mani, Automatic Summarization, John Benjamins Pub Co., 2001.
[38] J.P. Mei, L. Chen, SumCR: a new subtopic-based extractive approach for text summarization, Knowledge and Information Systems 31 (2012) 527–545.
[39] L. Moreno Boronat, M. Palomar Sanz, A. Molina Marco, A. Ferrández Rodríguez, Introducción al procesamiento del lenguaje natural, Universidad de Alicante, 1999.
[40] A. Nenkova, L. Vanderwende, K. McKeown, A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization, in: SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006, pp. 573–580.
[41] E. Newman, W. Doran, N. Stokes, J. Carthy, J. Dunnion, Comparing redundancy removal techniques for multi-document summarisation, in: Proceedings of STAIRS, August, 2004, pp. 223–228.
[42] S. Overbeek, M. Janssen, P. van Bommel, A standard language for service delivery: enabling understanding among stakeholders, Computer Standards & Interfaces 34 (2012) 355–366.
[43] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2 (2008) 1–135.
[44] M. Paule Ruiz, M. Fernández Díaz, F. Ortín Soler, J. Pérez Pérez, Adaptation in current e-learning systems, Computer Standards & Interfaces 30 (2008) 62–70.
[45] L. Plaza, A. Díaz, P. Gervás, A semantic graph-based approach to biomedical summarisation, Artificial Intelligence in Medicine 53 (2011) 1–14.
[46] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, Z. Zhang, MEAD — A Platform for Multidocument Multilingual Text Summarization, Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, 2004.
[47] D.R. Radev, S. Blair-Goldensohn, Z. Zhang, Experiments in single and multi-document summarization using mead, in: First Document Understanding Conference, 2001, pp. 1–7.
[48] D.R. Radev, H. Jing, M. Budzikowska, Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies, in: NAACL-ANLP 2000 Workshop on Automatic Summarization, 2000, pp. 21–30.
[49] H. Saggion, A classification algorithm for predicting the structure of summaries, in: Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009), 2009, pp. 31–38.
[50] H. Saggion, R. Gaizauskas, Multi-document summarization by cluster/profile relevance and redundancy removal, in: The Document Understanding Workshop (Presented at the *HLT/NAACL* Annual Meeting), 2004.
[51] C. Sauper, R. Barzilay, Automatically generating Wikipedia articles: a structure-aware approach, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009, pp. 208–216.
[52] K. Spärck Jones, Automatic summarizing: factors and directions, in: Advances in Automatic Text Summarization, MIT Press, 1999, pp. 1–14.
[53] K. Spärck Jones, Automatic summarising: the state of the art, Information Processing and Management 43 (2007) 1449–1481.
[54] J. Steinberger, M. Turchi, M.A. Kabadjov, R. Steinberger, N. Cristianini, Wrapping up a summary: from representation to generation, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL (Short Papers), 2010, pp. 382–386.
[55] N. Stokes, J. Rong, B. Laugher, Y. Li, L. Cavedon, NICTA's update and question-based summarisation systems at DUC 2007, in: The Document Understanding Workshop (Presented at the *HLT/NAACL*), 2007.
[56] Y. Sun, S.C. Park, Generation of non-redundant summary based on sum of similarity, in: ITCC '05: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05), Volume II, 2005, pp. 782–783.
[57] D. Tatar, E. Tamaianu-Morita, A. Mihis, D. Lupsa, Summarization by logic segmentation and text entailment, in: Proceedings of Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2008), 2008, pp. 15–26.
[58] Z. Teng, Y. Liu, F. Ren, S. Tsuchiya, F. Ren, Single document summarization based on local topic identification and word frequency, in: MICAI '08: Proceedings of the 2008 Seventh Mexican International Conference on Artificial Intelligence, 2008, pp. 37–41.
[59] K. Toutanova, C. Brockett, M. Gamon, J. Jagarlamudi, H. Suzuki, L. Vanderwende, The PYTHY Summarization System: Microsoft Research at DUC 2007, in: The Document Understanding Workshop (presented at the *HLT/NAACL*), 2007.
[60] B. Wang, B. Liu, C. Sun, X. Wang, B. Li, Adaptive maximum marginal relevance based multi-email summarization, in: Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 417–424.
[61] S. Xie, Y. Liu, Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 4985–4988.
[62] O. Yeloglu, E. Milios, N. Zincir-Heywood, Multi-document summarization of scientific corpora, in: Proceedings of the 2011 ACM Symposium on Applied Computing, ACM, New York, NY, USA, 2011, pp. 252–258.
[63] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, W.Y. Ma, Improving web search results using affinity graph, in: SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005, pp. 504–511.
[64] Z. Zhou, Combined features to maximal marginal relevance algorithm for multi-document summarization, Journal of Convergence Information Technology 6 (2011) 298–304.

Dr. Elena Lloret is a post-doctoral researcher at the University of Alicante at the European Project FIRST: A Flexible Interactive Reading Support Toool (grant no. FP7-287607). She is a Computer Science graduate and she received her Ph.D at the University of Alicante. Her main field of interest is text summarization, text simplification and text comprehension. She is the author of over 25 scientific publications in relevant journals and international conferences. She has been collaborating with international researchers and has participated in a number of projects at a national level (TIN2006-15265-C06, TIN2009-13391-C04). She has also been collaborating with international groups in Wolverhampton, Sheffield and Edinburgh.

Prof. Dr. Manuel Palomar is the University President of the University of Alicante and head of the Natural Language Processing and Information Systems Research Group of the same university. He is also a full professor of this University since 1991 and his main teaching area focuses on the analysis, design and management of databases, datawarehouses, and information systems. He received his Master's degree and Ph.D in Computer Science at the Polytechnic University of Valencia, Spain. His research interests are Human Language Technologies (HLT) and Natural Language Processing (NLP), in particular text summarization, semantic roles, textual entailment, information extraction and anaphora resolution. He has supervised more than 12 theses and he is the author of more than 70 scientific publications on international journals and conferences on different topics related to HLT and NLP. Furthermore, he has coordinated and been involved in a number of regional, national and international research projects funded by the Generalitat Valenciana (Valencian Government), the Ministry of Science and Innovation (Spanish Government) and the European Council.

Please cite this article as: E. Lloret, M. Palomar, Tackling redundancy in text summarization through different levels of language analysis, Computer Standards & Interfaces (2012), http://dx.doi.org/10.1016/j.csi.2012.08.001