

## SEMPCA-SUMMARIZER: EXPLOITING SEMANTIC PRINCIPAL COMPONENT ANALYSIS FOR AUTOMATIC SUMMARY GENERATION

Óscar ALCÓN, Elena LLORET

*Department of Software and Computing Systems*

*University of Alicante*

*Apdo. de correos 99, E-03080, Alicante, Spain*

*e-mail: {oalcon, elloret}@dlsi.ua.es*

**Abstract.** Text summarization is the task of condensing a document keeping the relevant information. This task integrated in wider information systems can help users to access key information without having to read everything, allowing for a higher efficiency. In this research work, we have developed and evaluated a single-document extractive summarization approach, named SemPCA-Summarizer, which reduces the dimension of a document using Principal Component Analysis technique enriched with semantic information. A concept-sentence matrix is built from the textual input document, and then, PCA is used to identify and rank the relevant concepts, which are used for selecting the most important sentences through different heuristics, thus leading to various types of summaries. The results obtained show that the generated summaries are very competitive, both from a quantitative and a qualitative viewpoint, thus indicating that our proposed approach is appropriate for briefly providing key information, and thus helping to cope with a huge amount of information available in a quicker and efficient manner.

**Keywords:** Natural language processing, human language technologies, intelligent information processing, automatic text summarization, principal component analysis

**Mathematics Subject Classification 2010:** 68-T50

## 1 INTRODUCTION

In the current digital society, the increase of available data and the impossibility to cope with it in an efficient manner requires more and more the development of techniques capable of reducing and analyzing such data without losing the key and relevant ideas contained. One way of reducing the dimension of a document keeping, and at the same time having a relevant information is thinkable through automatic text summarization [1], which can be applied to a wide range of domains to help digest and manage information in an easier and less-time consuming manner. In this research area, summaries are created with the purpose of identifying the gist of the original text, while discarding irrelevant information, so advanced techniques to perform data analysis and knowledge interpretation are required. As a result, only the key information is selected and extracted. The type of summarization generated under these premises is extractive, in the sense that a summary would be produced just by copying the relevant sentences from the original document, without rewriting them or making any change to the produced text. Although many different approaches have been developed since this research area emerged in the late fifties [2], it is still a field under research, where issues that may seem successfully addressed to some extent, such as the correct identification of key information for building extractive text summarization approaches, need to be further studied and improved [3]. A correct identification of the key information would greatly benefit more advanced summarization processes, where the information is then compressed, fused or re-generated, as in the case of abstractive summarization. Furthermore, companies could take advantage of the integration of text summarization systems as management tools in their daily processes. This integration would allow to optimize information harvesting and digestion. For instance, an automatic process could decide whether to distribute or not some information that may be of interest, thus avoiding employees and CEOs having to read and manually filter tons of information.

Therefore, the aim of this research work is to describe and evaluate a novel approach for single-document extractive summarization using Principal Component Analysis (PCA) technique enhanced with semantic analysis (SemPCA-Summarizer). PCA is a statistical technique to compress data by reducing the number of dimensions, without much loss of information [4]. In the proposed approach, a concept-sentence matrix is built from the input document, and then, PCA is used to identify and rank the concepts, under the hypothesis that semantic analysis may have a positive impact on the summarization process, and thus, in the generated summaries, in contrast to the traditional process of taking only into consideration the individual words contained in the text for computing frequencies. This is where the semantic analysis plays a key role within our text summarization approach. Finally, to create the summary, we analyze four different heuristics to order the sentences based on the presence and importance of each concept, so that the ones leading to the best summaries can be determined and chosen. The results obtained show that our method produces good summaries, both from a quantitative

and qualitative point of view, as well as it overperforms the results of other existing summarization approaches. Our main contributions to the state of the art are:

- Semantic analysis is performed to detect synonymous words within the text, that constitute our working elements, i.e., the concepts to be analyzed.
- A novel approach for computing and determining relevant concepts is proposed.
- Four different heuristics are analyzed to select and order the most relevant sentences, leading to the creation of different types of summaries within the same method, as well as determining the approach that generates the best summaries.
- The performance of the proposed approach against a set of state-of-the-art summarization systems is compared, obtaining very good results for SemPCA-summarizer.

The remainder of the paper is structured as follows. In Section 2, the information about previous research related to extractive text summarization with data mining techniques is described. In Section 3, we explain PCA technique and how we have used it for text summarization, i.e., our Semantic PCA-based extractive summarization approach, SemPCA-Summarizer. Section 4 explains the automatic and manual evaluation methodology carried out to test our suggested approach. The results obtained together with their discussion are provided in Section 5 and finally, the conclusions as well as several future research works are outlined in Section 6.

## 2 RELATED WORK

A wide range of techniques, including statistical [5], graph-based [6, 34], machine learning [7], or linguistically motivated theories [8, 33] have been used by the research community to address extractive text summarization [3]. For this type of summaries, such techniques are mainly used for determining the key information in documents, and scoring sentences based on their relevance using different techniques and algorithms, such as spreading activation [9], genetic algorithms [35], word and phrase embeddings [10], integer linear programming [11], or even neural networks, which have gained a lot of interest in the recent years [12, 13]. Other summarization approaches integrate several techniques (statistical, linguistic, etc.) in order to analyze the benefits of combining them. This is the case of COMPENDIUM summarizer [36], which employs textual entailment together with statistical and linguistic-based features for scoring sentences, and determines which ones are more relevant to take part in the summary. Also, the approach proposed in [37] analyzed different combinations of statistical and linguistic settings, such as anaphora resolution together with Word Sense Disambiguation (WSD) methods for extracting the most important sentences based on the number of concepts they contained, instead of terms.

When focusing on similar or related statistical techniques as the ones used in the current research work (i.e., the use of PCA for summarization), the first ap-

proach using PCA for text summarization was presented in [14]. In this approach, PCA was employed for quantifying both word frequency and co-occurrence in the document to extract thematic words. The significant sentences of the text were selected using those thematic words under the hypothesis that those sentences would represent the main topic of a document. An improved approach derived from this research work was later developed in [15], in which PCA was used to identify relevant terms and then, Singular Value Decomposition (SVD) was employed for extracting the significant sentences associated to such terms (the higher number of terms a sentence had, the more important it was considered) and build the final summary. The experiments performed over a set of Korean newspaper articles showed that the method that used PCA and SVD jointly achieved the best performance for F-measure (0.436), compared to the method in which only PCA was employed (0.416).

A different approach was developed in [16], where single- and multi-document summaries were generated using a modified version of the vector space model, called Semantic Vector Space Model (SVSM) to model the set of documents including what the authors defined as action words. According to them, action words are verbs that are used to strengthen the way experiences are presented regardless expressing positive or negative experiences (e.g. break, destroy, arrive, etc.). In this approach, PCA was used to extract topic features. In the context of single-document summarization, it was evaluated in a controlled environment, using three individual articles belonging to several topics. It obtained an average accuracy of 61.85% when the extracted sentences of the proposed summarization method were compared to the sentences that a human would have extracted. Despite the high accuracy obtained, the authors did not test their approach with a bigger corpus, so this approach could not be validated with a wider set of documents.

Moreover, at some period of time, there seemed to be an increasing trend towards the use of data mining techniques for text summarization, including SVD, Latent Semantic Analysis (LSA), or PCA. Indeed, all these techniques (SVD, LSA, PCA) are tightly related, differing in the way they process and decompose the term-document matrix. In [17], a comparison between different approaches using SVD for generating summaries was carried out. From this comparison, [18] got a F-measure of 0.60 with their feature-based approach created by single LSA-based sentence scores. Their system included a set of textual features belonging to prosodic and lexical features. LSA was also used for computing the score of the sentences. Their experiments were carried out using both human transcriptions and the output of an automatic speech recognizer.

Although PCA has been already employed for text summarization [14], to the best of our knowledge, the influence of semantic analysis for computing relevant concepts with PCA has not been exploited so far. Traditionally, the PCA matrix is built from word co-occurrence, and consequently, the potentials of the knowledge obtained through a semantic analysis process are not integrated.

In the following section, our PCA-based approach for single-document extractive text summarization is described in detail.

### 3 SEMPCA-SUMMARIZER: A SEMANTIC PCA-BASED EXTRACTIVE SUMMARIZATION SYSTEM

Text summarization approaches follow a generic flow distinguishing three phases [19]:

1. *interpretation*;
2. *transformation*; and
3. *summary generation*.

In the interpretation stage, the text has to be understood. From a computational perspective, this is usually done using language processing tools, capable of analyzing the text at different language-levels (e.g., lexical, syntactic, or semantic). Next, in the transformation stage, the text is represented using the information obtained from the previously mentioned tools. The text could be represented in different ways using, for instance, a vector-space model, graphs, etc. Finally, once this representation is obtained, the last stage is devoted to apply some algorithm to determine the important information and generate the final summary based on it. We followed the same convention for defining the stages of our text summarization approach. Figure 1 shows an overview of the process for the proposed *SemPCA-Summarizer* approach.

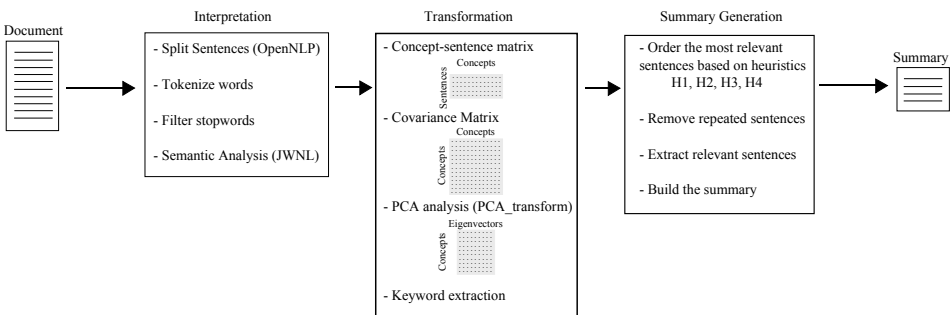


Figure 1. SemPCA-Summarizer approach for extractive single-document summarization

Firstly, the document needs to be interpreted to create the concept-sentence matrix, which will be analyzed using PCA. This is carried out by performing a linguistic analysis of the text, explained in Section 3.1. Next, the concept-sentence matrix reflects the co-occurrence of the concepts (instead of words) in each sentence. That is the way how we obtain a higher level of abstraction by considering concepts instead of just words, since a concept groups a set of synonymous words (for instance, *hurricane*, *tornado*, and *cyclone*). The process of creating such a matrix, applying PCA, and subsequently, extracting the concepts with the highest weights, determined by PCA, corresponds to the transformation stage, explained in Section 3.2. Lastly, in the summary generation phase, using these con-

cepts as a sign of relevance, the most important sentences will be selected and ordered using different heuristics, so the final summary can be generated (Section 3.3).

### 3.1 Interpretation: Linguistic Analysis and Concept Identification

A basic linguistic text preprocessing is necessary to proceed with the creation of the concept-sentence matrix. Once the input text is split in sentences, using the OpenNLP<sup>1</sup> library for Java, each of them is tokenized to subsequently filter stopwords (i.e., those words lacking semantic meaning that are not useful for calculating the frequency of occurrence, such as “the”). Afterwards, semantic analysis is applied to each word in order to identify concepts. For this, the knowledge base WordNet [20] was chosen to perform the semantic analysis using its Java library JWNL<sup>2</sup>. WordNet is a lexico-semantic English resource that groups words into sets of synonyms called synsets, providing short and general definitions. It also provides information about the semantic relationships between the synsets. In our proposed approach, it is used to infer existing sets of synonyms in the documents, thus working with concepts instead of terms.

For identifying concepts, the process searches for the first synset of each word in the document. The first synset WordNet returns corresponds to the most frequent sense of that word, and therefore its most probable meaning. This approach has been proven to obtain very competitive results compared to more sophisticated word sense disambiguation techniques [21, 22, 23, 24]. If two words have the same first synset, we will consider them as synonyms. Therefore, they will be grouped under the same concept, and their occurrences will be added together. For example, *detonation* and *explosion* are different words but their WordNet’s first synsets are the same (07323181), so we keep them as a single concept in the concept-sentence matrix.

At the end of this stage, the text is prepared to compute the concept-sentence matrix and apply PCA technique, as it is next explained.

### 3.2 Transformation: PCA for Key Information Detection

In this stage, the concept-sentence matrix is created, from which the PCA technique will be applied.

PCA is a statistical technique focused on the extraction of information to compress and interpret the data [4]. For large volumes of data, the aim of this algorithm is to find a set of patterns or trends to reduce the dimensionality in the input data set.

PCA creates projections of the input samples in a subspace of a smaller dimension by finding a linear combination of the original data. The linear combina-

---

<sup>1</sup> <https://opennlp.apache.org/>

<sup>2</sup> <http://sourceforge.net/projects/jwordnet>

tion is constructed with respect to the importance in terms (in decreasing order) of the total variability of the sample population. The covariance matrix is computed to obtain the principal components (eigenvectors,  $e$ ) and their corresponding weights (eigenvalues,  $\lambda$ ). In this respect, the covariance matrix can be decomposed into a set of eigenvalue-eigenvector pairs  $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$ , where  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ .

PCA returns a matrix in which the eigenvectors are the columns and the rows are the variables of the covariance matrix. The eigenvectors are composed by the contribution of each variable, which determines the importance of the variable in the eigenvector. Moreover, the eigenvectors are derived in decreasing order of importance determined by the eigenvalues.

In this manner, an eigenvector with high eigenvalue carries a great amount of information. Therefore, the first eigenvectors collect the major part of the information extracted from the covariance matrix.

In our approach, PCA is applied using the `PCA_transform`<sup>3</sup> Java library to process the covariance matrix from the concept-sentence matrix. In our concept-sentence matrix, the concepts (nouns, verbs, adverbs, and adjectives) are considered as variables (columns), whereas the sentences are the observations of the matrix (rows). This is also a novelty with respect to [14], where only nouns were used as variables.

Once PCA is applied to the covariance matrix, for each eigenvector, ordered by importance, the concept(s) with the highest value is/are extracted, as they are considered being more relevant than others. These concepts will be used for selecting the most important sentences, as shown next.

### 3.3 Summary Generation

From the previous stages, the matrix with the eigenvectors from PCA is obtained; however, an important stage in any extractive summarization process is to finally determine and select the specific sentences that will constitute the summary to be used by users or other Natural Language Processing (NLP) processes. Therefore, in this stage, the strategy for choosing the key sentences to form the final summary, based on the values obtained from PCA is defined.

Once we know the relevant keyword, it is frequent that the same keyword appears in more than one sentence of the documents, so we need to come up with a method to determine which sentences should be part of the final summary that contain such keyword. Therefore, four heuristics are defined for selecting and ordering the most relevant sentences from the document. The heuristics were chosen following the definition and purposes of different types of summaries (e.g., generic, informative, topic-focused, etc.) with the aim to allow the approach to be more flexible concerning the generation of various kind of summaries that could be appropriate depending on the users' information needs. Taking into account the concept

---

<sup>3</sup> [https://github.com/mkobos/pca\\_transform](https://github.com/mkobos/pca_transform)

with the highest value for each eigenvector from the PCA matrix, the approach selects and extracts:

**H1: one sentence** (searched in order of appearance in the original text) in which such concept appears. During this process, if a sentence had been already selected by a previous concept to take part in the summary, we would select and extract the next sentence in which the concept appears to avoid including redundant information.

**H2: only the first sentence** in which such concept appears. If a sentence had been already selected by a previous concept to take part in the summary, we would skip that sentence, and go ahead to the next concept. This heuristic is similar to H1, but since H2 focuses only on the first sentence, it does not allow to select and extract other sentences associated to each specific concept.

**H3: all the sentences** in which such concept appears.

**H4: all the sentences** in which at least two relevant concepts appear, giving priority to the concept with higher weight according to the PCA matrix. This heuristic is similar to the one proposed in [14] which selects the sentences based on the number of thematic words determined by PCA the sentence contain, trying to maximize the number of thematic words a sentence has. However, H4 differs from it in the fact that we prioritize the importance of the concepts in the sentence, rather than the number of concepts that a sentence may contain.

If we found different concepts with the same highest value for the same eigenvector, we would extract the corresponding sentences for all these concepts. In the same manner, if a concept was represented by several terms that are synonyms, we would extract the corresponding sentences for each of these synonyms. For instance, let us suppose that the concept id *07323181* was detected as an important concept by the approach. This concept corresponds to the terms *detonation* and *explosion*. In case both terms appear in the original document, the sentences in which those terms appear would be potentially selected to be part of the resulting summary.

Moreover, these strategies provide us with the relevance of the sentences in decreasing order, that are chosen for building the summary until the desired length is reached. In all cases, redundancy in the final summary is avoided by not allowing the inclusion of repeated sentences, if these have already been selected.

## 4 EVALUATION ENVIRONMENT

The conducted evaluation verifies the performance of the proposed automatic text summarization approach, i.e., determines the accuracy of the generated summaries and assesses their quality with respect to the information contained, and their readability.



In this section, the corpus and the evaluation methods designed to test our summaries are described. Both the quantitative and qualitative evaluation was performed. Quantitative evaluation conducts an automatic evaluation comparing the generated automatic summaries with respect to manual summaries, which are taken into account as model summaries. Qualitative evaluation completes the automatic evaluation, by means of a manual evaluation, which gathers and analyzes the users' opinion about the different generated summaries.

#### 4.1 Corpus

Our approach was evaluated with news documents using the DUC 2002 corpus<sup>4</sup>. In particular, this corpus consists of 533 different English news articles, and it contains up to three human summaries for each document. These human summaries, that have a length of approximately 100 words, will be considered as model summaries for our quantitative evaluation.

Despite the fact that more recent datasets are available for testing novel summarization approaches, the reason for selecting this dataset is twofold. Firstly, it is a corpus suitable for developing and testing generic single-document summarization systems, since further DUC editions proposed specific summarization tasks (e.g. headline generation, update summarization), and in most of the cases the provided corpora focused on multi-document summarization. Secondly, this dataset is one of the most widespread corpus for single-document summarization, thus allowing to compare our proposed method with state-of-the-art approaches under the same conditions.

We therefore used the DUC 2002 news documents as input for SemPCA-Summarizer (one document at a time) and generated the corresponding summaries with each of the heuristics previously described. To provide a fair comparison with respect to the model summaries available, automatic summaries were produced taking into account a 100 word length limit.

#### 4.2 Quantitative Evaluation

The quantitative evaluation was performed using ROUGE [25]. This tool allows the automatic evaluation of text summaries by comparing their content to a model one, created by a human. The idea under ROUGE is to measure the number of overlapping n-gram units between both types of summaries (e.g., unigrams, bigrams, or word sequences). Based on these, ROUGE implements different metrics, such as ROUGE-1, ROUGE-2, ROUGE-L, or ROUGE-SU4. Following [26], we will take into account: unigram similarity (ROUGE-1); bigram similarity (ROUGE-2); longest common subsequence (ROUGE-L) and bigram similarity skipping unigrams (ROUGE-SU4).

---

<sup>4</sup> [http://www-nlpir.nist.gov/projects/duc/data/2002\\_data.html](http://www-nlpir.nist.gov/projects/duc/data/2002_data.html)

In addition, for each of the aforementioned metrics, ROUGE returns the commonly used NLP measures: Precision, Recall and F-measure.

$$\textit{Precision} = \frac{\# \textit{CorrectPhrasesExtracted}}{\# \textit{TotalPhrasesExtracted}}, \quad (1)$$

$$\textit{Recall} = \frac{\# \textit{CorrectPhrasesExtracted}}{\# \textit{CorrectPhrasesTest}}, \quad (2)$$

$$\textit{F-measure} = \frac{2 * \textit{Precision} * \textit{Recall}}{\textit{Precision} + \textit{Recall}} \quad (3)$$

where  $\# \textit{CorrectPhrasesExtracted}$  is the number of correct sentences that the evaluated approach extracts,  $\# \textit{TotalPhrasesExtracted}$  the total number of sentences that the evaluated approach extracts and  $\# \textit{CorrectPhrasesTest}$  the total number of sentences included in the human summaries. Human summaries were manually generated by experts, so they are considered as gold-standard summaries.

### 4.3 Qualitative Evaluation

To verify if the generated summaries show an acceptable degree of coherence, they will be qualitatively evaluated through a manual evaluation. With this type of evaluation, our aim is to know the users' opinion about our summaries regarding different quality criteria and their readability.

In order to define an appropriate manual evaluation environment, a revision of previous works also conducting this type of evaluation was carried out first. In [27], seven questions were formulated regarding the following aspects: coherence, non-redundancy (2 questions), clarity, cohesion, grammaticality and readability. Each of them was evaluated giving a scoring from 1 to 5 (1 = very poor, and 5 = very good). This way of proceeding is known as the Likert rating scale [28].

Similar to the previous one, other examples can be found ([27, 29]) which employ a Likert scale with seven questions to evaluate the summaries too, but slightly varying the aspects that are evaluated in a summary: five questions about linguistic features (*grammaticality*, *non-redundancy*, *referential clarity*, *focus*, *structure and coherence*), one question based on amount of information in the summary that contributes to meeting the information need expressed in the topic (*content responsiveness*), and another one about the overall response of the summary (i.e., *overall responsiveness*).

More recently, the criteria to qualitatively evaluate summaries by the National Institute of Standards and Technology [30] focused on *content*, *readability/fluency* and *general responsiveness*. In this case, summaries were truncated to 100 words, and each summary was assessed by four people. Again, each of these criteria was evaluated according the Likert scale. In [31], the Likert-scale assessment was proposed using the following five criteria of quality: *grammaticality*, *non-redundancy*, *clarity*, *accuracy (focus)*, and *coherence*.

For qualitative evaluation, the Likert-scale evaluation seems to be one of the most accepted methodologies, so we also opted for this type of scale for our evaluation. Regarding the types of questions for assessing the summaries, seven questions were defined, some of which were adapted from already existing ones [32]:

1. Reading only the first sentence of the summary, does it allow to have a clear idea of what the original document is about?
2. Do you think the summary reflects the relevant information from the original document?
3. Does the summary contains the information in a non-redundant way?
4. Do you think that the sentences in the summary are correctly ordered?
5. Do you think that the content of the summary can be well understood?
6. Would you say that the summary is easy to read?
7. In general, do you consider it a good summary?

These questions were grouped with respect to:

- Accuracy of the information (questions 1 and 2)
- Redundancy (question 3)
- Readability/understanding (questions 4, 5, and 6)
- Overall assessment (question 7)

The responses associated to each question were measured in the range of 1 to 5, under the following scale:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

## **5 RESULTS AND DISCUSSION**

In this section, the performed experiments and the results are shown and discussed.

### **5.1 Quantitative Results**

As mentioned in Section 4.2, ROUGE was used for the quantitative evaluation. ROUGE was computed for the summaries obtained from the heuristics proposed in Section 3.3. Since we used 533 input texts from DUC 2002 Corpus, we obtained 533 summaries for each heuristic (we have four heuristics), thus having a total of

2 132 summaries for evaluation, and 1 085 DUC 2002 model summaries<sup>5</sup> as reference. Figure 2 provides example summaries generated from document id *WSJ880912-0064* of the DUC 2002 corpus, described in Section 4.1.

To check the appropriateness of performing semantic analysis, our approach was run using semantic analysis (i.e., concept-sentence matrix) and without it (i.e. using word-sentence matrix). The results can be seen in Table 1 and Table 2.

	ROUGE-1			ROUGE-2			ROUGE-L			ROUGE-SU4		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
H1	0.46454	0.46688	<b>0.46508</b>	0.21011	0.21127	<b>0.21041</b>	0.42233	0.42459	<b>0.42290</b>	0.22804	0.22928	<b>0.22834</b>
H2	0.46540	0.46048	0.46148	0.20814	0.20491	0.20571	0.42183	0.41742	0.41831	0.22699	0.22379	0.22453
H3	0.43619	0.44018	0.43799	0.18561	0.18732	0.18638	0.39602	0.39966	0.39767	0.20724	0.20923	0.20814
H4	0.44037	0.43446	0.43535	0.18605	0.18296	0.18359	0.39955	0.39441	0.39521	0.20886	0.20555	0.20614

Table 1. Results for the SemPCA-Summarizer summarization system

	ROUGE-1			ROUGE-2			ROUGE-L			ROUGE-SU4		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
H1	0.46265	0.46691	0.46452	0.20885	0.21052	0.20952	0.41917	0.42299	0.42083	0.22661	0.22850	0.22738
H2	0.46479	0.46594	0.46490	0.20915	0.20904	0.20881	0.42121	0.42217	0.42125	0.22711	0.22713	0.22682
H3	0.43310	0.43840	0.43563	0.18267	0.18503	0.18380	0.39267	0.39750	0.39498	0.20515	0.20781	0.20642
H4	0.43898	0.43882	0.43758	0.18630	0.18541	0.18506	0.39706	0.39711	0.39599	0.20840	0.20889	0.20787

Table 2. Basic PCA-based extractive summarization approach without semantic analysis

The best overall F-measure values are emphasized in boldface. These correspond to H1 with semantic analysis (Table 1). A t-test was conducted to account for statistical significance. We compared the use of semantic knowledge with respect to the basic method without using this type of knowledge for each F-measure result of H1,H2, H3, and H4, but no statistical significance was found. Nevertheless, analyzing and comparing the same heuristic with and without semantic analysis, we find that in two of them (H1 and H3), the semantic analysis contributes to slight improvement of the results, obtaining an average ROUGE improvement of 0.36 % and 0.86 %, respectively. The reason of this small improvement could be probably caused by either the absence of synonyms in texts due to the finer granularity of the resource used (i.e. WordNet). However, for H2 and H4, there was not any improvement at all when employing semantic knowledge, and what is more, the results with semantic analysis even suffer a dropped of 1 % and 0.57 %, respectively. This may be due to the fact that these two heuristics prioritize the amount of concepts in contrast to H1 and H3, that gave priority to the importance of the concepts in sentences. Therefore, this issue could negatively affect the results, since it would be difficult to find synonyms in the same sentences, and as a result, the sentences extracted would be less relevant.

Considering the nature of the four suggested heuristics, and when having a look at the summaries generated, we believe that each of the heuristic could be particularly suitable for generating a specific type of summary. In this manner, H1

<sup>5</sup> The number of model summaries is lower because the range of human summaries existing for each document varies from 1 to 3.

**Human summary:** Hurricane Gilbert swept toward Jamaica yesterday with 100 mph winds, and officials issued warnings to residents on the southern coasts of the Dominican Republic, Haiti and Cuba. Hal Gerrish, with the National Hurricane Center in Coral Gables, Fla., said, "It's still gaining strength. It's certainly one of the larger systems we've seen in the Caribbean for a long time." The governments of Haiti, Cuba and Jamaica issued warnings. In Puerto Rico, the storm injured three people, tore off several roofs, caused coastal flooding and brought down trees and power lines in the west and southwest regions.

**Best DUC 2002:** Hurricane Gilbert swept toward Jamaica yesterday with 100-mile-an-hour winds, and officials issued warnings to residents on the southern coasts of the Dominican Republic, Haiti and Cuba. The storm ripped the roofs off houses and caused coastal flooding in Puerto Rico. Forecasters said the hurricane was gaining strength as it passed over the ocean and would dump heavy rain on the Dominican Republic and Haiti as it moved south of Hispaniola, the Caribbean island they share, and headed west. At 3 p.m. EDT, the center of the hurricane was about 100 miles south of the Dominican Republic and 425 miles east of Kingston, Jamaica.

**SemPCA- H1:** The storm ripped the roofs off houses and caused coastal flooding in Puerto Rico. Forecasters said the hurricane was gaining strength as it passed over the ocean and would dump heavy rain on the Dominican Republic and Haiti as it moved south of Hispaniola, the Caribbean island they share, and headed west. Hurricane Gilbert swept toward Jamaica yesterday with 100-mile-an-hour winds, and officials issued warnings to residents on the southern coasts of the Dominican Republic, Haiti and Cuba. The hurricane was moving west at about 15 mph and was expected to continue this motion for the next 24 hours. In the Dominican Republic, all domestic flights and flights to and from Puerto Rico and Miami were canceled.

**SemPCA- H2:** The storm ripped the roofs off houses and caused coastal flooding in Puerto Rico. Forecasters said the hurricane was gaining strength as it passed over the ocean and would dump heavy rain on the Dominican Republic and Haiti as it moved south of Hispaniola, the Caribbean island they share, and headed west. Hurricane Gilbert swept toward Jamaica yesterday with 100-mile-an-hour winds, and officials issued warnings to residents on the southern coasts of the Dominican Republic, Haiti and Cuba. In the Dominican Republic, all domestic flights and flights to and from Puerto Rico and Miami were canceled. Forecasters said the hurricane's track would take it about 50 miles south of southwestern Haiti.

**SemPCA- H3:** The storm ripped the roofs off houses and caused coastal flooding in Puerto Rico. In the Dominican Republic, all domestic flights and flights to and from Puerto Rico and Miami were canceled. The hurricane center said small craft in the Virgin Islands and Puerto Rico should remain in port until conditions improve. In Puerto Rico, besides tearing off several roofs, the storm caused coastal flooding and brought down power lines and trees along roads and highways in the west and southwestern regions. Three people were injured in Guayanilla, Puerto Rico, when a tree fell on their vehicle as they traveled along Route 97, police reported.

**SemPCA- H4:** The storm ripped the roofs off houses and caused coastal flooding in Puerto Rico. In the Dominican Republic, all domestic flights and flights to and from Puerto Rico and Miami were canceled. The hurricane center said small craft in the Virgin Islands and Puerto Rico should remain in port until conditions improve. In Puerto Rico, besides tearing off several roofs, the storm caused coastal flooding and brought down power lines and trees along roads and highways in the west and southwestern regions. Four policemen stationed on Mona Island, between Puerto Rico and the Dominican Republic, were stranded as a result of the weather.

Figure 2. Example of automatic and human summaries generated from document *WSJ880912-0064* of the DUC 2002 corpus

would be appropriate for generic informative summaries, H2 for generic indicative summaries, H3 for topic-oriented informative summaries, and H4 for topic-oriented indicative summaries. In this case, we performed a one-way ANOVA statistical test with a Tukey HSD Post Hoc analysis for each of the heuristics in order to account for statistical differences between them in each of the methods analyzed (SemPCA-summarizer and Basic PCA-based summarizer). We found that, in both approaches, the results were statistically significant (at least with  $(p\text{-value} < 0.05)$ ) in all the pairwise comparison, except for H1–H2 and H3–H4. This is expected since H1–H2 focus on the production of generic summaries, while H3–H4 are aimed to produce topic-oriented summaries, and the differences between them is the amount of detail about a concept that is provided in the summary, being both types appropriate. Moreover, since the human summaries we used for comparing our automatic summaries were built from a generic and informative point of view, this could explain why for this corpus, H1 leads to the best ROUGE results for all the metrics studied.

Furthermore, our best approach (PCA enhanced with semantic analysis and using H1 heuristic) was compared to other state-of-the-art summarization approaches under the same conditions (same corpus and same evaluation metrics). The comparison can be seen in Table 3, in which ROUGE-1 recall value of our best approach (H1) is compared with:

- the lead baseline of DUC 2002, which takes approximately the first 100 words in the documents;
- the best DUC 2002 approach [33], which was based on a supervised sentence extraction using a Hidden Markov Model and a Logistic Regression Model;
- wMVC summarizer [34]. This approach, called the Weighted Minimum Vertex Cover summarization approach, considered the text summarization task as an optimization problem. A graph-based algorithm was used, where vertices of the graph represented the sentences and graph edges represented the connections between sentences;
- MUSE [35], an approach based on the linear optimization of several sentence ranking measures using a genetic algorithm;
- COMPENDIUM [36], which employed textual entailment, statistical and linguistic-based features for scoring sentences, and determines which ones had to take part in the summary;
- TS + MFS-approach [37];
- TS + UKB-approach [37];

The last two approaches, [37], used anaphora resolution and a Word Sense Disambiguation (WSD) method for enriching the document with semantic knowledge, and extracting the most important sentences based on the number of concepts they contained, instead of terms. The difference between them was the WSD method employed: MFS for the most frequent sense, and UKB for a PageRank-based WSD method [38].

The reason for using ROUGE recall metric for the comparison, instead of precision or F-measure is that we did not reimplement the summarization methods. Instead, we relied on published results for all the compared approaches, so the only common metric across all the approaches was the recall metric for ROUGE-1. From the comparison shown in Table 3, it can be seen that our PCA-based approach (H1) obtains the best results in terms of Rouge-1 recall metric over the DUC 2002 corpus.

	ROUGE -1 (Recall value)
SemPCA-Summarizer (H1)	<b>0.46688</b>
Best DUC 2002 approach	0.42776
Lead baseline DUC 2002	0.41132
wMVC summarizer [34]	0.38800
MUSE [35]	0.45490
COMPENDIUM [36]	0.46008
TS + MFS approach [37]	0.42339
TS + UKB approach [37]	0.42556

Table 3. Comparison with other approaches that used the DUC 2002 corpus

The quantitative results obtained show the appropriateness of the proposed approach for summarization, which is competitive with respect to the state of the art, having as the additional advantage that it is flexible enough to create different types of summaries (generic vs. topic-oriented; informative vs. indicative). This would definitely benefit the adaptation of the generated summary depending on the information needs and user profile. For instance, indicative summaries or headlines may be useful for CEOs to have a daily update of the most important facts without spending too much time reading more detailed information.

## 5.2 Qualitative Results

Given the importance of assessing the readability of a summary besides its content, a manual evaluation was conducted, as explained in Section 4.3. However, due to the difficulty and time-consuming task involving manual evaluation, we did not conduct this type of evaluation to all the generated summaries with our approach. This would have implied the manual evaluation of more than 2000 summaries, meaning a good amount of human resources and effort that was unfortunately not available. Therefore, we finally opted for selecting a representative sample of summaries of each type for the manual evaluation.

To find out the representative number of summaries, a statistical formula called *Representative sample (M)* [39] was employed:

$$M = \frac{N * K^2 * P * Q}{E^2 * (N - 1) + K^2 * P * Q} \quad (4)$$

where  $N$  is the current population,  $K$  is confidence level,  $P$  is the assumed probability of success,  $Q$  is the probability of failure, i.e.,  $1-P$ , and  $E$  is the error rate. The values for each parameter were set according to the suggestions in [40]:

$$K = 0.95; \quad E = 0.05; \quad P = 0.5; \quad Q = 0.5. \quad (5)$$

Since we also wanted to compare our approach with the best DUC 2002 approach, previously mentioned in Section 5.1, and the model summaries, the population was calculated taking into account 3 198 summaries (533 best DUC 2002 approach summaries, 533 human summaries and 2 132 summaries generated with our approach). The other approaches compared (e.g., MUSE, COMPENDIUM, etc.) were not manually evaluated since their generated summaries were not available.

Using Equation 4 we obtained  $M = 87.79 \approx 88$  summaries. On account of this, it was decided to take a total sample of 90 abstracts to be manually evaluated. The total sample of 90 summaries was made up of 60 summaries of PCA text summarization types (15 summaries for each type), 15 summaries of the best DUC 2002 approach, and 15 model summaries created by humans in DUC 2002. All the summaries were truncated to approximately 100 words. The 90 abstracts were evaluated by 18 people (comprising PhD students and senior researchers with advanced knowledge in English), so each one had to evaluate 5 summaries. Summarization evaluation is a subjective task, so people evaluating a summary may have different opinions about its quality. In order to minimize the bias that may be obtained by allowing only one user to evaluate a summary, we enrich the qualitative evaluation, asking an additional external user to evaluate also the 90 summaries. In this manner, a summary would be evaluated by two people, and therefore the assessment would not only depend on one user. The final manual evaluation results were computed by averaging the scores assigned by the two assessors for each summary. In this manner, if the discrepancy between the two assessors was high, for instance, if one person assigned a good score, i.e. 5, and the other one a low score, i.e. 2, we would consider that the summary would not be either very bad or very good, and in this case the average would be a fair value for its final score.

In Table 4, the results of the manual evaluation are shown. As it can be seen, the results prove that, in a general way, the summaries are acceptable, obtaining values closer to the positive range of the Likert scale.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7
DUC 2002 human	4.3	4.1	4.5	4.5	4.2	4.2	4.2
Best DUC 2002 approach	4.1	3.3	4.6	4.6	3.5	4.1	3.4
SemPCA-Summarizer with H1	3.7	3.6	4.4	3.6	3.6	4.0	3.6
SemPCA-Summarizer with H2	3.4	3.3	4.7	3.6	3.3	4.0	3.1
SemPCA-Summarizer with H3	3.1	3.3	3.9	3.8	3.6	3.5	3.2
SemPCA-Summarizer with H4	3.0	3.4	3.8	3.6	3.4	3.5	3.4

Table 4. Qualitative evaluation results obtained through a manual inspection of the generated summaries



As we expected, the human summaries exhibit the best results. This could be because abstractive summaries always seem to be better considered by humans than extractive ones. However, for several questions (e.g., Q3, Q6), the results do not greatly differ from the rest of the approaches (Best DUC 2002 approach, SemPCA-summarizer with H1, and SemPCA-summarizer with H2).

When focusing on the extractive summarization approaches, and firstly comparing only our heuristics among them, H1 is the one which obtains the best results, reaching the highest values in five out of the seven questions outlined. On the other hand, H3 and H4 show the lowest results, although the average obtained between all the questions is above a 3-rating score.

When inspecting the results for each type of question, the results of Q1 and Q2 show that the accuracy of the information in the summaries is coherent with the heuristics tested, being H1 the best heuristic. Here, it is important to mention that the first sentence of the Best DUC 2002 approach shows a clear idea of the information of the original text. This is logical, since this approach ensured that first sentence of the original document was always included in the automatic summary. However, when having a look at Q2, it seems that the rest of the summary contains less important information compared with the generated summaries using the other heuristics. Concerning the redundancy of the summaries (Q3), the results of all the evaluated approaches indicate that they are very good with respect to this issue.

Switching now to the readability/understanding questions, the results of Q4 reflect that the ordering of the sentences in the summary is better in the Best DUC 2002 approach than in any of our suggested heuristics. Again, this is expected, since the sentences in the Best DUC 2002 approach are shown in the same order as in the original document. In contrast, in the case of our heuristics, the sentences extracted are ordered with respect to the importance of the concepts and not necessarily by the order the sentences appear in the original texts. To improve this issue in our heuristics, we could have taken into account the position of the sentence in the text, and order the extracted sentences in the same order as in the original document. However, it seems that in the case of H1 this fact minimally affects the legibility of the summaries, since the results of Q5 and Q6 show good values. Compared to our heuristics, the summaries produced by the Best DUC 2002 approach does not exhibit a very high understanding score (Q5), which is lower than the score obtained for H1 and H3.

Finally, according to overall assessment results (Q7), some of our heuristics (H1 and H4) obtain equal or higher results than the Best DUC 2002 approach. This question (Q7) gives an idea about to what extent the generated summary would be appropriate. In this respect, we computed the number of summaries scored under each rating for this question for the two manual evaluations conducted. Figure 3 graphically depicts the results obtained comparing both manual evaluations.

As it can be seen, although the individual scoring may vary from different assessors, partly due to the subjectivity involved in the evaluation process and the fuzzy differences between the scores assigned (an assessor may have doubts between lower or upper boundaries, e.g., 1–2 or 4–5), we can see that better summaries are clearly

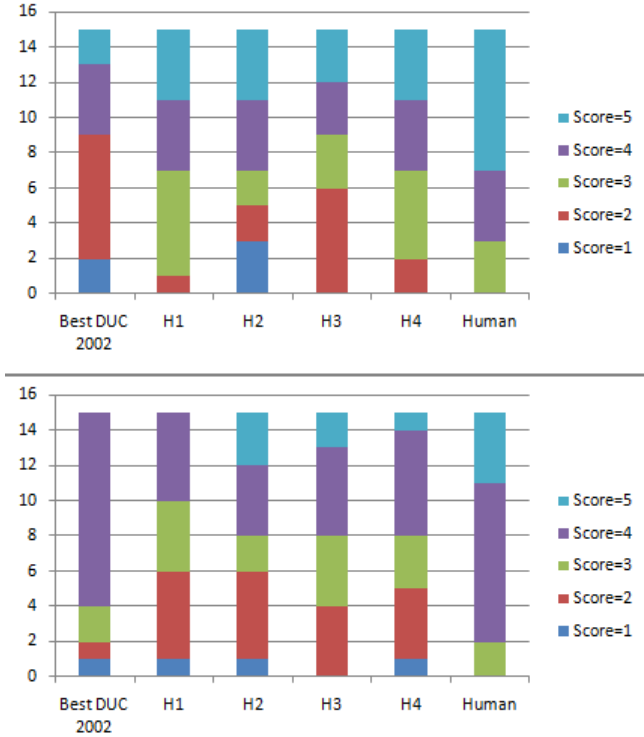


Figure 3. Results of manual evaluation for Q7. The first graphic corresponds to the original manual evaluation, and the second one to the enriched manual evaluation, where an additional external expert reevaluated all the summaries again.

distinguished from poor ones, as in the case of human summaries. In contrast, there are more discrepancies when it comes to automatic summaries, where there may be a greater variability in as far as users' opinions is concerned.

### 6 CONCLUSIONS AND FUTURE WORK

In this paper, a single-document extractive text summarization approach using Principal Component Analysis (PCA) technique enhanced with semantic information was proposed and analyzed. Our approach has been called SemPCA-Summarizer. Specifically, PCA was used as a detector and ranker of the relevant concepts within the text, to further extract the sentences associated to them according to different heuristics, so the final summary could be generated.

The results obtained have confirmed that SemPCA-Summarizer produces good summaries from the quantitative and qualitative point of view, and it is competitive with respect to the state of the art, so it can be very useful for detecting and

providing key information. Having analyzed different heuristics for selecting sentences based on the relevant concepts identified, the best performing heuristic is H1, which selects one sentence (searching in order of appearance in the original text) per relevant concept. In particular, it obtained the best results for all the ROUGE metrics tested, when semantic knowledge was included in the base system. Concerning the qualitative evaluation, H1 also achieved very good results compared with the remaining heuristics. It is also worth noting that the performed experiments and the obtained results allow us to gain some insights concerning the potential that our approach may have for generating other types of summaries, e.g., generic vs. topic-oriented, or indicative vs. informative summaries.

Although there is still some room for improvement, this finding could be very encouraging since it may constitute the starting point for further research.

In the short- and medium-term, we plan to analyze whether the approach could be improved if fragments of sentences instead of complete sentences were taken into account as the linguistic units to process. This could help to raise the accuracy of the relevant information. Moreover, it would be very interesting to test the approach for other types of documents, integrating more advanced word sense disambiguation methods that would take into account the context of the whole document in which a word appears. For this, word embeddings could be used. In the long-term, it would be very useful for the research community to test the appropriateness of our approach for multilingual text summarization, so that the approach could deal with texts of different languages.

## Acknowledgements

This research work has been partially funded by the Generalitat Valenciana and the Spanish Government through the projects PROMETEOII/2014/001, TIN2015-65100-R, and TIN2015-65136-C2-2-R. Additionally, the authors would like to thank all the users who contributed to the summaries evaluation.

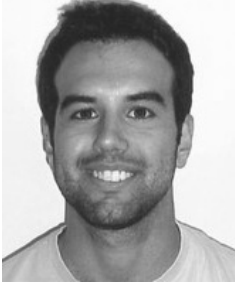
## REFERENCES

- [1] GAMBHIR, M.—GUPTA, V.: Recent Automatic Text Summarization Techniques: A Survey. *Artificial Intelligence Review*, Vol. 47, 2017, No. 1, pp. 1–66, doi: 10.1007/s10462-016-9475-9.
- [2] LUHN, H. P.: The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, Vol. 2, 1958, No. 2, pp. 159–165.
- [3] NENKOVA, A.—K. MCKEOWN: Automatic Summarization. *Foundations and Trends in Information Retrieval*, Now Publishers Inc., Vol. 5, 2011, No. 2-3, pp. 103–233, doi: 10.1561/15000000015.
- [4] JOLLIFFE, I. T.: *Principal Component Analysis*. Springer Verlag, New York, 2002.

- [5] MCCARGAR, V.: Statistical Approaches to Automatic Text Summarization. *Bulletin of the American Society for Information Science and Technology*, Vol. 30, 2004, No. 4, pp. 21–25, doi: 10.1002/bult.319.
- [6] CANHASI, E.—KONONENKO, I.: Multi-Document Summarization via Archetypal Analysis of the Content-Graph Joint Model. *Knowledge and Information Systems*, Vol. 41, 2014, No. 3, pp. 821–842, doi: 10.1007/s10115-013-0689-8.
- [7] MEI, J.-P.—CHEN, L.: SumCR: A New Subtopic-Based Extractive Approach for Text Summarization. *Knowledge and Information Systems*, Vol. 31, 2012, No. 3, pp. 527–545, doi: 10.1007/s10115-011-0437-x.
- [8] UZÊDA, V. R.—PARDO, T. A. S.—NUNES, G. V.: A Comprehensive Comparative Evaluation of RST-Based Summarization Methods. *ACM Transactions on Speech and Language Processing*, Vol. 6, 2010, No. 4, Art. No. 4, doi: 10.1145/1767756.1767757.
- [9] NASTASE, V.—FILIPPOVA, K.—PONZETTO, S. P.: Generating Update Summaries with Spreading Activation. *Proceedings of Text Analysis Conferences (TAC)*, NIST, 2008.
- [10] KÅGEBÄCK, M.—MOGREN, O.—TAHMASEBI, N.—DUBHASHI, D.: Extractive Summarization Using Continuous Vector Space Models. *Proceedings of the 2<sup>nd</sup> Workshop on Continuous Vector Space Models and Their Compositionality (CVSC)*, Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 31–39, doi: 10.3115/v1/W14-1504.
- [11] BOUDIN, F.—MOUGARD, H.—FAVRE, B.: Concept-Based Summarization Using Integer Linear Programming: From Concept Pruning to Multiple Optimal Solutions. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 1914–1918, doi: 10.18653/v1/D15-1220.
- [12] CAO, Z.—WEI, F.—DONG, L.—LI, S.—ZHOU, M.: Ranking with Recursive Neural Networks and its Application to Multi-Document Summarization. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, January 25–30, 2015, Austin, Texas, USA, 2015, pp. 2153–2159.
- [13] RUSH, A. M.—CHOPRA, S.—WESTON, J.: A Neural Attention Model for Sentence Summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 379–389, doi: 10.18653/v1/D15-1044.
- [14] LEE, C. B.—KIM, M. S.—PARK, H. R.: Automatic Summarization Based on Principal Component Analysis. In: Pires, F. M., Abreu, S. (Eds.): *Progress in Artificial Intelligence (EPIA 2003)*. Springer, Berlin, Heidelberg, *Lecture Notes in Computer Science*, Vol. 2902, 2003, pp. 409–413.
- [15] LEE, C.—PARK, H.—OCK, C.: Significant Sentence Extraction by Euclidean Distance Based on Singular Value Decomposition. In: Dale, R., Wong, K. F., Su, J., Kwong, O. Y. (Eds.): *Natural Language Processing – IJCNLP 2005*. Springer, Berlin, Heidelberg, *Lecture Notes in Computer Science*, Vol. 3651, 2005, pp. 636–645.

- [16] VIKAS, O.—MESHARAM, A. K.—MEENA, G.—GUPTA, A.: Multiple Document Summarization Using Principal Component Analysis Incorporating Semantic Vector Space Model. *Computational Linguistics and Chinese Language Processing*, Vol. 13, 2008, No. 2, pp. 141–156.
- [17] BADRY, R. M.—ELDIN, A. S.—ELZANFALLY, D. S.: Text Summarization within the Latent Semantic Analysis Framework: Comparative Study. *International Journal of Computer Applications*, Vol. 81, 2013, No. 11, pp. 40–45.
- [18] MURRAY, G.—RENALS, S.—CARLETTA, J.: Extractive Summarization of Meeting Recordings. *Proceedings of the 9<sup>th</sup> European Conference on Speech Communication and Technology (Interspeech-2005 – Eurospeech)*, Lisbon, Portugal, September 4–8, Association for Computational Linguistics, 2005.
- [19] SPARCK JONES, K.: *Automatic Summarising: Factors and Directions*. *Advances in Automatic Text Summarization*, 1999, MIT Press.
- [20] MILLER, G. A.: WordNet: A Lexical Database for English. *Communications of the ACM*, Vol. 38, 1995, No. 11, pp. 39–41, doi: 10.1145/219717.219748.
- [21] PREISS, J.—DEHDARI, J.—KING, J.—MEHAY, D.: Refining the Most Frequent Sense Baseline. *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (DEW'09)*, 2009, pp. 10–18, doi: 10.3115/1621969.1621973.
- [22] LLORET, E.—MOREDA, P.—MORENO, I.—CANALES, L.: *Meaning Disambiguator: v2.1*. Technical report, University of Alicante, 2014, <http://first-asd.eu/?q=D4.1>.
- [23] BHINGARDIVE, S.—SINGH, D.—RUDRA MURTHY, V.—REDKAR, H.—BHATTACHARYYA, P.: Unsupervised Most Frequent Sense Detection Using Word Embeddings. *Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 1238–1243, <http://www.aclweb.org/anthology/N15-1132>, doi: 10.3115/v1/N15-1132.
- [24] BHINGARDIVE, S.—SHUKLA, R.—SARASWATI, J.—KASHYAP, L.—SINGH, D.—BHATTACHARYYA, P.: Synset Ranking of Hindi WordNet. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May 23–28, 2016.
- [25] LIN, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of the ACL-04 Workshop on Text Summarization Branches Out*, Association for Computational Linguistics, 2004, pp. 74–81.
- [26] STEINBERGER, J.—JEZEK, K.: Sentence Compression for the LSA-Based Summarizer. *Proceedings of the 7<sup>th</sup> International Conference on Information Systems Implementation and Modelling*, 2006, pp. 141–148.
- [27] YEN, J.—OVER, P.: An Introduction to DUC-2004 Intrinsic Evaluation of Generic News Text. *Proceedings of the Document Understanding Conference (DUC 2004)*, 2004.
- [28] LIKERT, R.: A Method of Constructing an Attitude Scale. In: Maranell, G. M. (Ed.): *Scaling: A Sourcebook for Behavioral Scientists*. Chapter 19. Aldine Publishing Company, Chicago, 1974, pp. 233–243.

- [29] CONROY, J.—MCKEOWN, K.—SPARCK-JONES, K.—VANDERWENDE, L.: Overview of DUC 2006. Proceedings of the Document Understanding Conference (DUC 2006), 2006.
- [30] National Institute of Standards and Technology NIST: TAC 2011 Guided Summarization Task Guidelines. 2011.
- [31] LLORET, E.—PALOMAR, M.: Towards Automatic Tweet Generation: A Comparative Study from the Text Summarization Perspective in the Journalism Genre. *Expert Systems with Applications*, Vol. 40, 2013, No. 16, pp. 6624–6630, doi: 10.1016/j.eswa.2013.06.021.
- [32] CARLSSON, B.—JÖNSSON, A.: Using the Pyramid Method to Create Gold Standards for Evaluation of Extraction Based Text Summarization Techniques. Proceedings of the Swedish Language Technology Conference (SLTC 2010), 2010.
- [33] CONROY, J. M.—SCHLESINGER, J. D.—O’LEARY, D. P.—OKUROWSKI, M. E.: Using HMM and Logistic Regression to Generate Extract Summaries for DUC. Proceedings of the DUC 01 Conference, 2001, pp. 13–14.
- [34] GUPTA, A.—KAUR, M.—SINGH, A.—GOEL, A.—MIRKIN, S.: Text Summarization Through Entailment-Based Minimum Vertex Cover. Proceedings of the Third Joint Conference on Lexical and Computational Semantics (SEM 2014), 2014, pp. 75–80, doi: 10.3115/v1/S14-1010.
- [35] LITVAK, M.—LAST, M.—FRIEDMAN, M.: A New Approach to Improving Multilingual Summarization Using a Genetic Algorithm. Proceedings of the 48<sup>th</sup> Annual Meeting of the Association for Computational Linguistics (ACL ’10), Association for Computational Linguistics, 2010, pp. 927–936.
- [36] LLORET, E.—PALOMAR, M.: COMPENDIUM: A Text Summarisation Tool for Generating Summaries of Multiple Purposes, Domains, and Genres. *Natural Language Engineering*, Vol. 19, 2013, No. 2, pp. 147–186, doi: 10.1017/S1351324912000198.
- [37] VODOLAZOVA, T.—LLORET, E.—MUÑOZ, R.—PALOMAR, M.: The Role of Statistical and Semantic Features in Single-Document Extractive Summarization. *Artificial Intelligence Research*, Vol. 2, 2013, No. 3, pp. 35–44, doi: 10.5430/air.v2n3p35.
- [38] AGIRRE, E.—SOROA, A.: Personalizing PageRank for Word Sense Disambiguation. Proceedings of the 12<sup>th</sup> Conference of the European Chapter of the Association for Computational Linguistics (EACL ’09), Association for Computational Linguistics, 2009, pp. 33–41, doi: 10.3115/1609067.1609070.
- [39] PITA FERNÁNDEZ, S.: Sample Size Calculation (Determinación del Tamaño Muestral). *Cad Atención Primaria*, Vol. 3, 1996, pp. 114–138.
- [40] GUTIÉRREZ VÁZQUEZ, Y.—FERNÁNDEZ ORQUÍN, A.—MONTOYO GUIJARRO, A.—VÁZQUEZ PÉREZ, S.: Integration of Semantic Resources Based on WordNet. *Procesamiento del Lenguaje Natural*, Vol. 47, 2011, pp. 161–168.



**Óscar ALCÓN** currently works at CYPE Ingenieros S.A. as Product Development Responsible Manager. His tasks include the specification and supervision of software development for the design of common telecommunications infrastructures in buildings for Spain and Portugal. It includes activities such as market research, determination of deadlines, contacts with customers and collaborators, definition of lines of work, product presentations, commercial activities and relations with marketing and sales departments. In addition, he also conducts research into natural language processing and text summarization as a collaborator with the GPLSI research group at the University of Alicante, and more specifically for the project “DIIM2.0: Desarrollo de técnicas inteligente e interactivas de minería y generación de información sobre la web 2.0”.



**Elena LLORET** is a full-time Ph.D. Assistant Lecturer at the University of Alicante in Spain. There she obtained her Ph.D. focused on text summarisation in 2011. Her main interests are natural language processing and more specifically text summarisation, and natural language generation. She is the author of over 60 scientific publications in international peer-reviewed conferences and refereed journals. She has served in the program committee on international conferences, such as ACL, EACL, RANLP, or COLING. She is a member of the Spanish Society for Natural Language Processing (SEPLN) and she has participated in a number of national and EU-funded projects. She has also been collaborating with international groups at the University of Wolverhampton (UK), the University of Sheffield (UK), the University of Edinburgh (UK), and the Lorraine Research Laboratory in Computer Science and Its Applications in France. Since 2009 she has been involved in teaching activities at the University of Alicante. Specifically, for the degrees in computer science engineering and multimedia engineering and for the master’s programme in information technologies and English and Spanish language for specific purposes, involving 200 teaching hours per year.