Ontology-based text summarization. The case of Texminer

Pedro Hípola, José A. Senso, Amed Leiva-Mederos and Sandor Domínguez-Velasco
Departamento de Informacion y Comunicacion, Universidad de Granada, Granada, Spain

Received 9 January 2014
Revised 2 February 2014
Accepted 3 February 2014

Abstract

Purpose – The purpose of this paper is to look into the latest advances in ontology-based text summarization systems, with emphasis on the methodologies of a socio-cognitive approach, the structural discourse models and the ontology-based text summarization systems.
Design/methodology/approach – The paper analyzes the main literature in this field and presents the structure and features of Texminer, a software that facilitates summarization of texts on Port and Coastal Engineering. Texminer entails a combination of several techniques, including: socio-cognitive user models, Natural Language Processing, disambiguation and ontologies. After processing a corpus, the system was evaluated using as a reference various clustering evaluation experiments conducted by Arco (2008) and Hennig *et al.* (2008). The results were checked with a support vector machine, Rouge metrics, the *F*-measure and calculation of precision and recall.
Findings – The experiment illustrates the superiority of abstracts obtained through the assistance of ontology-based techniques.
Originality/value – The authors were able to corroborate that the summaries obtained using Texminer are more efficient than those derived through other systems whose summarization models do not use ontologies to summarize texts. Thanks to ontologies, main sentences can be selected with a broad rhetorical structure, especially for a specific knowledge domain.

Keywords Information retrieval, Software evaluation, Ontologies, Indexing, Programming, Automatic summarization systems, Texminer
Paper type Research paper

1. Introduction

Advances in Natural Language Processing (NLP), a discipline that goes back to the 1950s, continue to give rise to new summarization techniques that combine different lexical processing methods. The quality obtained in the process depends on a number of factors, such as:

- the quality of the corpus;
- the subject area of the documents; and
- the means of knowledge organization used to disambiguate the texts.

Automatic production of abstracts was initially based on statistical methods, in the wake of research by Luhn (1958), and progress today is largely due to diverse methodologies developed in the field of NLP. Such is the case of the automatic analyzers and socio-pragmatic discourse models, based on techniques for the extraction of terms or strings of significant words. These systems have been used since the 1990s to identify "rhetorical structures" highly related with the content of documents, and which give indications about the conceptual and organizational scheme of the different units making up a text (Ono *et al.*, 1994).

Library Hi Tech
Vol. 32 No. 2, 2014
pp. 229-248
© Emerald Group Publishing Limited
0737-8831
DOI 10.1108/LHT-01-2014-0005

Meanwhile, the growth of cognitive science has allowed the incorporation of semantic-conceptual models. When used in conjunction with knowledge bases, these models vastly improve the process of summarizing texts on specific topics.

Information regarding specific topics has recently experienced exceptional growth, and models for the extraction and disambiguation of texts should be developed accordingly. Depending on the specific subject matter of a document, it may prove more or less complicated to locate tools that allow knowledge to be processed and summarized.

One clear example would be Port and Coastal Engineering. Although it is a relatively new discipline, mixing classical Civil Engineering together with Ocean or Naval Engineering, the scientific evolution undergone by this field is noteworthy, encompassing new techniques for the construction, use and management of marine structures and resources. Notwithstanding, to date there is no specific tool available for the purpose of summarizing texts in this field of expertise.

The lack of specific dictionaries, the absence of a defined theory and the dire need for professionals in the sector to have summarization tools to carry out their work appear to plague this area – at least, seen from the standpoint of documentation, terminology and NLP. With the software presented here, Texminer, this problem may be solved. Our proposal is based on the conviction that summarizing texts within a specialized domain requires a model capable of processing its semantic and socio-cognitive components.

2. Summarization systems

2.1 Linguistic techniques

Summarization systems based on discourse structure analysis resort to the Rhetorical Structure Theory (RST) of Mann and Thompson (1988) to generate abstracts (Ono *et al.*, 1994; Marcu, 1998, 2000). They attempt to identify the internal structure of the text and the relations of discourse formed within it, giving priority to the nuclear components of these relations. Departing from the RST, Marcu (2000) segments a text into small units of discourse. Analyzing the set of relations that exist among all of them, he builds a rhetorical structure in the shape of a tree, with an orientation towards automatic summarization. Once the discourse structure of a text has been created, an algorithm assigns a weight and an order to each element of the structure – the higher the element within the structure, the greater its weight, and vice versa – to choose for the summary the most weighted elements, while excluding those with a low weight. In view of the length desired for the summary, it is possible to select an amount of elements in consonance with the organization previously established by the algorithm.

Other researchers use the text discourse structure as well, but in different ways. Teufel and Moens (2002) put forth a method for summarizing scientific articles from the domain of computational linguistics that uses the rhetorical status of affirmations contained in documents to identify their internal structure. The main contribution of these authors lies in the algorithm that deals with the non-hierarchical structure: given seven fixed categories (aim, textual, own, background, contrast, basis, other), it is capable of distributing the contents of articles within each category.

In turn, Gaizauskas proposes the use of templates to generate summaries and retrieve information (Gaizauskas *et al.*, 2001). This technique can only be applied when the text is previously structured. It has been used, above all, for extracting news items, in systems such as Scissor and Jasper, which use input templates subjected to partial analysis processes. In the financial field, systems such as Fies are similarly oriented to extract financial information from digital press articles.

The contributions of Mateo combine superficial with deep structure analysis, such as the detection of pronominal anaphora, and the use of discourse connectors to enhance the coherence and cohesion of the abstract (Mateo *et al.*, 2003). The computational complexity of these techniques may be considerable. Alonso and Fuentes focus their research on summarizing news stories, and demonstrate the implementation of a system that locates the cohesive properties of a text, using lexical chains plus the rhetorical and argumentative structure, derived using discourse markers (Alonso and Fuentes, 2003).

Along these same lines, involving a combination of complex linguistic techniques, we find the works by Aretoulaki (1997) and D'Cunha (2006), who develop a model for automatic summarization called Cosy-Mats. The system selects sentences using content features of a pragmatic and rhetorical nature, obtained by means of superficial linguistic analysis. The identification of features in Cosy-Mats is applied to a corpus comprising 160 newspaper articles and 170 scientific articles accompanied by their abstracts, the topic being computer science, natural sciences, philosophy and linguistics. Analysis is carried out by means of superficial level techniques, supported by the Theory of Speech Acts (Austin, 1962; Searle, 1969), the RST (Mann and Thompson, 1988) and theories centered on cohesion and coherence, such as Systemic Functional Linguistics (Halliday and Hasan, 1976), among others.

2.2 Use of ontologies

In addition to the statistical or superficial analysis (frequency of terms, words in titles, other textual positions or cue phrases) to determine the relevance of words in the original title, new summarization systems have incorporated additional procedures and systems, such as the use of ontologies, knowledge bases and databases. These innovations are founded on research by Lin and Hovy (1997) as well as Mani and Bloedorn (1999).

Ontologies have been applied in numerous research areas, above all in text mining (Yoo *et al.*, 2006). Having demonstrated their efficacy to cope with the tasks of documental abstracting, the use of ontologies is becoming more and more widespread for summarization methods.

Zhang proposes a semantic data processing focus that can be scaled to summarize documents (Zhang *et al.*, 2010). The key notion is that, aside from producing scalable semantic data for text extraction, it also serves for the different levels of the summary, in order to reduce its cardinality and enhance efficiency in information processing. To this end, it applies so-called granular computing techniques.

Another variant of ontology use can be found in the work by Hu, who describes a method of spatial integration to generate summaries of a single document with maximum topic completeness and minimum redundancy (Hu *et al.*, 2004). In the first place, a semantic vector is used to represent each class in the ontology as a linguistic unit, which may improve the quality of representation of the terms by means of vectors. Then, application of the K-means algorithm, along with a clustering analysis algorithm, makes it possible to identify different latent regions where the topic of a document appears. Finally, from each thematic region, the most representative sentences are chosen to generate the summary.

Likewise important is the contribution by Yuan and Sun, who apply an algorithm for text categorization, implementing cosine similarity to generate the summary of a document (Yuan and Sun, 2004). The method accounts for the structure of the terms in the documents, thereby improving the quality in terms of document grouping.

On the basis of ontologies, Popescu creates a functional summary of gene products, assigning them semantic annotations (Popescu *et al.*, 2004). The process begins when the Blast system carries out a hierarchic grouping of clusters, noting which are the most representative terms (MRT) for each cluster. The application then proposes more specific MRTs using weights from the fuzzy partition matrix generated by a relational fuzzy clustering algorithm. Finally, the weight of the terms is further increased in view of their content, which makes for greater specificity in the functional annotation of the groups and generates a higher quality summary. Havens, following Popescu, develops a summarization system that organizes the clusters in maps to facilitate their visualization, and produces a summary of gene products using similarity measures (Havens *et al.*, 2008).

Authors Chen and Verma propose text consultation using summarization techniques and the UMLS (Unified Medical Language System), which manages the ontologies of the National Library of Medicine. This method may be used to retrieve information from other areas of knowledge (Chen and Verma, 2006).

To help experts evaluate the evolution of catastrophes, Telesum, a system developed by Wu, provides informative summaries from a vast collection of documents (Wu *et al.*, 2013). The aim of this application is to locate the submodularity hidden among ontological concepts.

In turn, Andreasen and Bulskov propose a conceptual focus for information queries: the objective is to achieve an abstraction of the concepts appearing in texts from a collection of objects or documents (Andreasen and Bulskov, 2009). Two ontologies are introduced for this purpose. The second is conceived as an "instantiated ontology," a structure reflecting contents in the document collection. Assuming ontology-based similarity, they describe language constructs for direct navigation and retrieval of concepts in the ontology.

Probably the institutions that most use ontologies for summarization are news agencies. Lin and Liang have come up with a system that enables one to derive numerous reports to cover news areas about diverse events (Lin and Liang, 2008). This means of summarization identifies events while also presenting the news headlines and keywords, based on superficial processing techniques. Using TDT procedures (Topic Detection and Tracking), a summary is generated by means of a topic retrospection process applied to the SToRe system (Story-line based Topic Retrospection), which identifies several relevant events and drafts a summary that allows readers to infer the evolution of the event.

The system works along three basic processes: the identification of events, elaboration of a main storyline and a summary in which irrelevant aspects are eliminated. By extracting the most representative sentences, we arrive at a structured template with which to build the summary.

In the work by Hennig, ontologies are used to extract sentences taking into account the hierarchic structure of their nodes. In light of the ontology attributes, the semantic representation of the contents of sentences is improved by means of a classifier that uses search engines, and that produces satisfactory results, regardless of the topic being processed (Hennig *et al.*, 2008).

Similarly, Lee *et al.* (2003) propose an Ontology-Based Fuzzy Event Extraction agent to summarize news in Chinese. First, a Retrieval Agent (RA) retrieves internet e-news items regularly and stores them in a repository. Then, a Document Processing Agent utilizes the Chinese Part-of-Speech tagger to process the e-news and filter the Chinese term set. The Fuzzy Inference Agent and the Event Ontology Filter extract the e-news event ontology, and finally, a Summarization Agent (SA) sums up the e-news by the extracted-event ontology.

The many uses of ontologies to construct knowledge domains have also been addressed by Huang and Kuo (2007). These researchers propose that each sentence of a document may be represented by a set of WordNet senses and would constitute a fuzzy transaction for mining the conceptual lexicon and for ranking relevance. This form of automatic text summarization relies on information-retrieval criteria to assess the summary quality.

3. How Texminer works

3.1 System design

The process of developing the Texminer system over the years, from its general architecture to the computing tools used, and the software modules put into place successively, is described in previous studies (Leiva *et al.*, 2009; Leiva-Mederos, 2012; Leiva-Mederos *et al.*, 2012). Both the overall functional framework of Texminer and its respective modules have been in the beta stage since early 2012. Our objective is to have a candidate for the definitive version in the first semester of 2014. The whole system is intended to be developed as an open source platform. After the trial stage, all the modules will be available in Google Code under GNU General Public License v. 3.0, and processed by means of Git version control.

3.2 Cognitive processes

As we mentioned above, the summarization process within the domain of Port and Coastal Engineering as performed by Texminer entails a combination of summarization techniques, including: socio-cognitive user models, NLP, disambiguation and ontologies.

The socio-cognitive user paradigm, in the framework of cognitive psychology, takes into account historic, social and cultural factors. To carry out domain analysis, socio-cognitive techniques consider a field as a community developing and sharing common concepts, terms and knowledge (Endres-Niggemeyer *et al.*, 1995; Hjørland, 2002).

To develop the summarization process, we created an observers' guide that allowed us to register the actions undertaken by 12 librarians who summarized articles on Port and Coastal Engineering. Each of these librarians was observed by 12 psychologists, who took notes to fill in a form about the specific actions performed to draft and abstract. Finally, their abstracts were assessed and put into an order of priority and relevance as indicated in the works by Endres-Niggemeyer *et al.* (1995) and Pinto (2001). The observers' guide establishes the cognitive strategies enumerated in Figure 1.

The library professionals first read the document to determine whether or not it pertained to the specialized topic of Port and Coastal Engineering. They then segmented the text in view of its structure, locating the different parts. For instance:

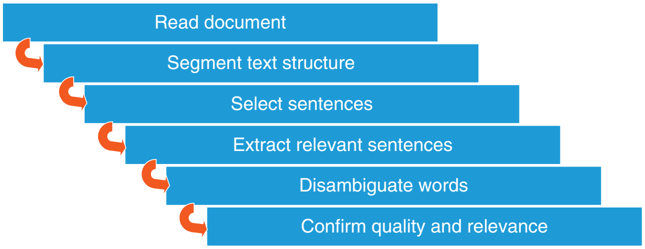

Figure 1.
Cognitive strategies involved in summarization

methodology, results, conclusions and discussion. In each one of these sections, they identified the sentences of greatest relevance, bearing in mind the domain terminology and rhetorical relations. After this stage, using specialized dictionaries, the words were disambiguated and pronominal anaphora were identified. Finally, the quality of the abstract and its relevance were corroborated.

3.3 Ontologies

In our research, ontologies are understood to be algebraic descriptions of a conceptual network in which the elements are binary relations established through concepts (Hung, 2008).

The ontology used in our project contains all the essential concepts of the topic of Port and Coastal Engineering appearing in the documental corpus or in the queries made using the PuertoTex system.

The conceptual network, founded on terminology provided by the LexiCon research group, is available at http://ecolexicon.ugr.es/visual/index_en.html. This network gave us abundant conceptual and semantic information to facilitate summarizing documents in this specific topic. Each segment of the ontology is a concept that describes one facet or hierarchy (Figure 2).

It is made up of 6,123 instances corresponding to 3,006 concepts obtained using WorldSmith tools. The semantic network classifies the concepts contained in the thesaurus under different frames: agent (natural agent/artificial agent), process (natural process/artificial process), patient/result, description (attributes, representation of, disciplines for study), instruments/procedures of description …. The main taxonomic relation in the ontology is defined as is_a, to establish the internal hierarchy. There are other relations for the objects, however. They include: caused_by, affected_by, produces, changes, affects, composed_of …. In addition to these vertical relations, we have others that define the properties of each datum: part_of, result_of, made_of, has_function, has _location, type_of, located_in ….

3.4 The agents of Texminer

Texminer uses agents for processing information, as does the system by Lee *et al.* (2003). In our case, there are three agents.

3.4.1 Reading agent. According to Grimes, reading is the first process in text construction (Grimes, 1975). Consequently, the process begins with an agent who goes over each section of the text (e.g. objectives, methodology, results and conclusions). The basic prerequisite for a document to be read by the agent is that its main sections and sentences are labeled in XML.

The initial task is to discern whether a document belongs to the domain of Port and Coastal Engineering. To confirm this, all stopwords are eliminated, whereas the terms that appear most often in the text are confronted with the ontology. If at least 50 percent of the terms are located in the structure of the ontology, the document is accepted as forming part of the domain, and it is sent to the processing subagent, who finalizes the relevance filter.

3.4.2 SA. The second stage of our model is founded on the approach of Marcu (2000), on the practical contributions by Mann and Thompson (1988), on those of D'Cunha (2006) and on the research works of Berry (2004) and Hu *et al.* (2004). The SA carries out a system of automatic tagging of the syntactic-discourse structure of the text. The conceptions used for the design of the tags obey three essential criteria:

- the sequence in which the levels are tagged in practice;

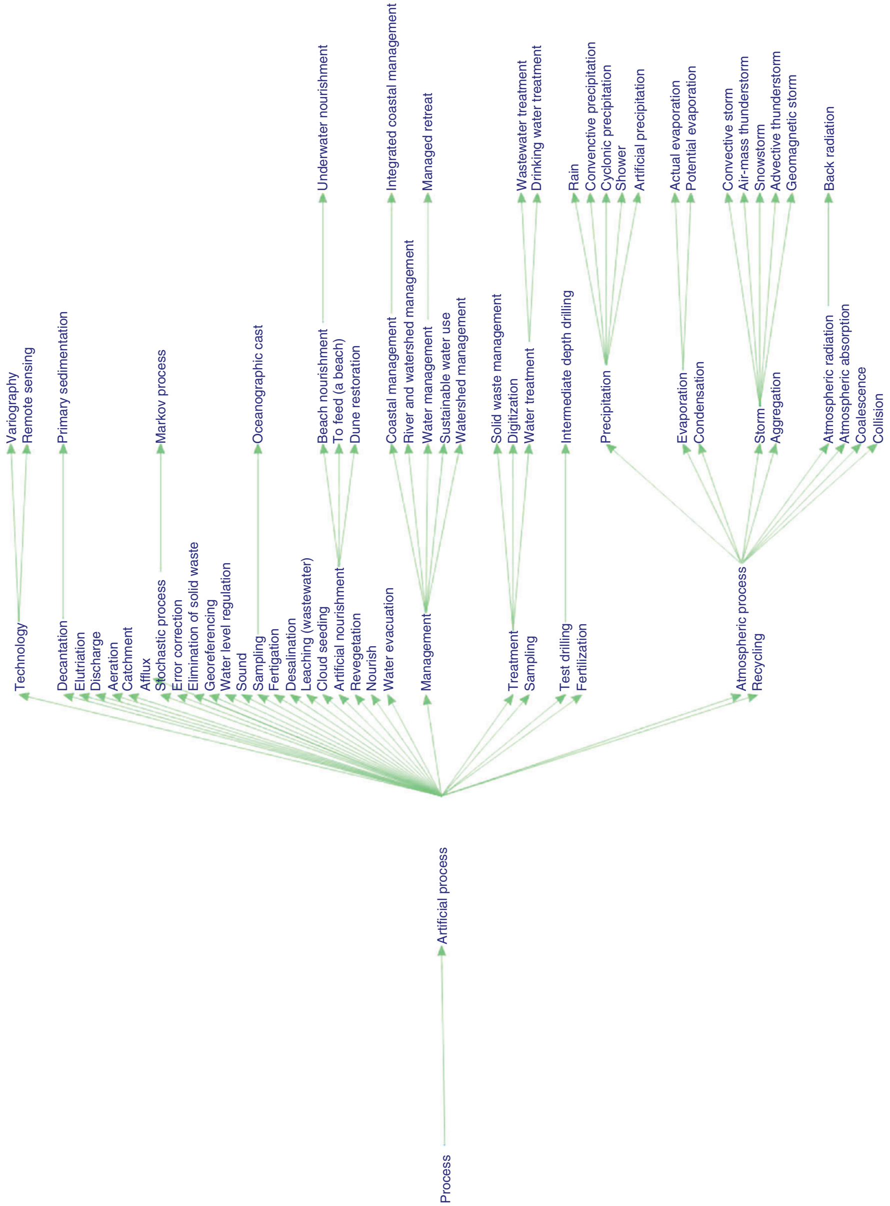

Figure 2.
Example of ontology from EcoLexiCon

- the symbols used to denote cohesive elements of the text; and
- the presence in the text of geographic elements, verbs, statistical data, formulae and processes.

Software developed by Domínguez-Velasco performs the tagging (Domínguez-Velasco, 2010). This tool assigns tags by means of a learning algorithm fed by discourse studies carried out previously (Leiva *et al.*, 2009). It identifies nuclei and satellites, at both the discourse and the syntactic-communicative level.

The agent follows a four-stage tagging strategy:

(1) It marks the discourse relations of communication, detecting within the sentences any textual elements that describe elaboration, interpretation, evidence, background, justification, results, condition, summary, reformulation, circumstance, list, sequence or union. One example:
<C><circu_s></circu_s></C>

(2) It tags the communicative structure of the text, marking the relations of theme and rheme.

(3) It marks all the bibliographic elements of the text, among them the name of the authors, using the schemes FOAF (Friend of a Friend) and Dublin Core.

(4) It records all the syntactic elements of the text by means of EuroWordNet and it disambiguates the text. After the tagging process there is text processing to select sentences from the potential ones offered by the ontology, and the rules described by Leiva-Mederos *et al.* (2012) are used to process the communicative structure of the text (Figure 3).

In generating automatic summarization, unsupervised algorithms are used in two ways. First, they calculate the weighting of each text element separately, analyzing statistical and linguistic-semantic characteristics of the text, so as to detect the greatest weight elements. Goldstein *et al.* (1999, 2000) describes research conducted with this type of algorithm, used very extensively in commercial systems.

Other algorithms use textual discourse for extracting the resulting text. The most important are:

- Barzilay's algorithm, based on lexical cohesion (Barzilay and Elhadad, 1997). This algorithm selects the first sentence of the text containing a representative chain and includes it in the extract as a strong lexical chain of the source text. When identifying lexical chains, calculation is based on the relationships between words following a knowledge base pattern of an ontological nature such as WordNet or EuroWordNet.
- The Nomoto and Matsumoto algorithm (Nomoto and Matsumoto, 2001). This algorithm performs a partition of the set of sentences of the source text through a clustering procedure, to generate an extract consisting of the most important sentences in each subset. The selection of such sentences takes into account the frequency of the terms in the document. According to Anaya *et al.* (2006) this algorithm has a quadratic type of complexity.

In text mining, the process of comparing classes and sentences in order to group them and determine their similarity depends on the representation of the sentence and the weight assigned. Several studies shown that grouping documents, sentences and

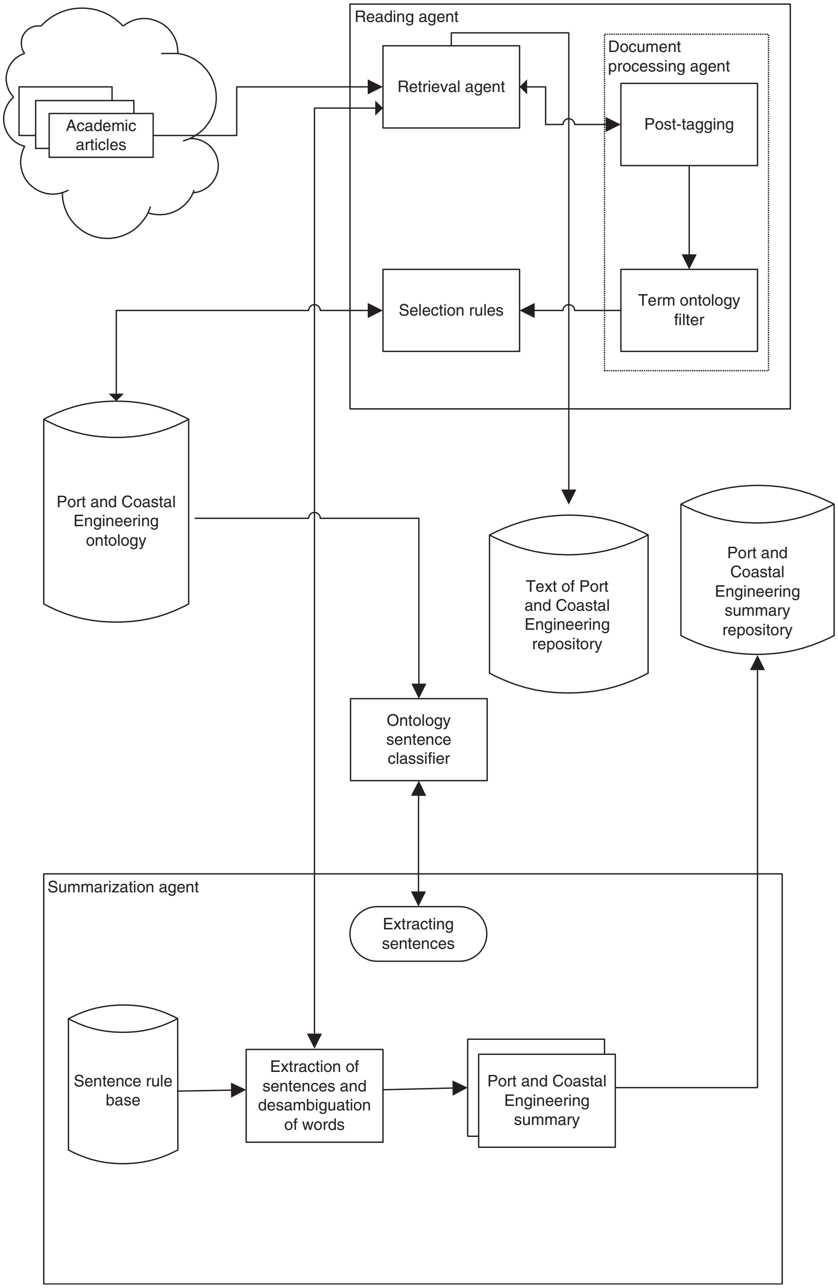
Academic articles
Reading agent
Retrieval agent
Document processing agent
Post-tagging
Selection rules
Term ontology filter
Port and Coastal Engineering ontology
Text of Port and Coastal Engineering repository
Port and Coastal Engineering summary repository
Ontology sentence classifier
Summarization agent
Extracting sentences
Sentence rule base
Extraction of sentences and desambiguation of words
Port and Coastal Engineering summary

Figure 3.
Process of summarizing

classes of ontologies would produce the best results with the coefficients Dice, Jaccard and Cosine.

In Texminer, a classification tool groups the sentences in the ontology. Hence, all data, organized in a tree, can be distributed in classes with increasingly thinner granularity (a divisive, top-down tree). Another tool builds the tree in a reverse way, combining small ontology classes into others with a thicker granularity (an agglomerative, bottom-up tree). These classification tools, developed ad hoc by Domínguez-Velasco (2013), implement an algorithm that performs cluster combinations proposed by Nomoto and Matsumoto (2001) and Alpcan *et al.* (2007). To group the main sentences, the software works with a classification model supported by Arco (2008) and improved for this research. These are the steps performed:

- tag the text according to the rules of the domain with the help of the ontology;
- vector space model (VSM) representation;
- calculate the quality of the terms;
- normalize the matrix;
- assign weights to terms and sentences;
- reduce the resulting matrix;
- calculate the similarity between sentences and ontology nodes;
- identify keywords and main sentences;
- classify or group sentences in the nodes of the ontology; and
- generate a summary extract mono-document (Figure 4).

Our system uses cosine similarity metrics because they produce the best results to estimate the similarity between ontology nodes and sentences. As in Hennig *et al.* (2008), the algorithm determines σ, the standard deviation of the resulting similarities,

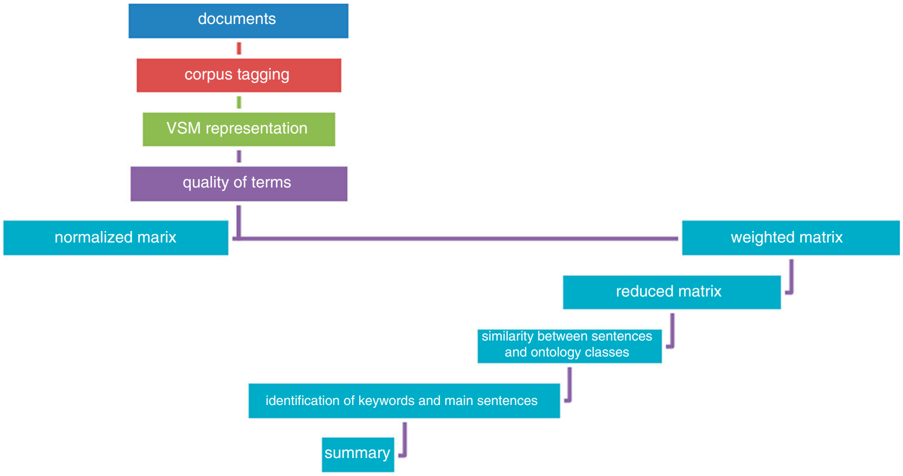

Figure 4. Metric operational framework

Source: Domínguez-Velasco (2013)

and selects all the nodes whose similarity with the sentence is sim (sentence $+$ node) $> \mu + \alpha\sigma$. This coefficient is widely used to determine the similarity between groups or nodes and is based on the cosine of the angle between them. If the cosine of the angle is close to one, the terms are considered similar, and if it is close to zero, the terms or sentences are considered different. Other coefficients, such as those applied by Jaccard or Dice, are not used in the study because their results vary widely if the ontology has many nodes or a highly branched structure. This leads to substantial differences in the results for similarity measures.

If the similarity is high, the system selects a single class or node to build the grouping or classification. Sentences are represented by nodes in a hierarchy. Thus, if a child node achieves greater similarity than a parent node, it will not be used as a candidate for the summary.

The authors of this paper are aware that the calculation of similarity could be based on classification algorithms commonly used in text mining such as:

- Simultaneous Keyword Identification and Clustering of Text Documents (SKWIC), which performs hard, deterministic cluster analysis; and
- Simultaneous Keyword Identification and Fuzzy Clustering of Text Documents (FuzzySKWIC), which uses a technique of fuzzy cluster analysis (Berry, 2004).
- Extended Star, which applies a hard overlapping cluster analysis technique (Gil-García and Pons-Porrata, 2008).

These algorithms are perfectly suited to interaction with VSM, and produce a collection of clusters. The magnitude of their operation varies in size, however, depending on the classification algorithms used.

Having selected the relevant sentences by means of the ontology, the following rules are applied:

- weighting rules, which allow us to choose the sentences already classified by the ontology because they contain terms that make reference to: statistical elements, chemical elements, geographic names and terms from the discourse list;
- syntactic-communicative rules, in charge of text coherence, managing criteria that allow for the detection and elimination of nuclei and satellites in the sentences, and to omit and eliminate elements of theme and rheme; and
- disambiguation rules, to lend meaning to the text.

Disambiguation of the meaning of words relies on the approach first described by Lesk (1986). The algorithm assigns a meaning to a specific word by comparing its different meanings with those of other words in the same context. The steps involved in the disambiguation process are:

(1) Let $(e_1, e_2, \ldots, e_n)$ be an input sentence when E_i is a word. Let $\{s_k(e_i) | k = 1, \ldots, q\}$ be the set of meanings of EuroWordNet e_i given. The "direct score" of the meaning candidate $S_k(e_i)$ is obtained with the following formula:

$$D(S_{k(e_i)}) = \sum_{j=1,i\neq j}^{n} L(S_{k(e_i)},S_{l(i,j)})$$

where $L(S_{k(e_i)},S_{l(i,j)})$ is the Lesk algorithm that registers the semantic relation between $s_k(e_i)$ and $s_l(e_j)$.

(2) One must bear in mind that each meaning in EuroWordNet is semantically related with a set of similar meanings. The score calculated in the above step reflects the direct relation between two meanings. Though it acts like a local link to plainly express the relationship between two meanings, its indirect relationship must also be taken into account. The indirect semantic relation of the two meanings is the sum of the paired affinity scores of the two sets of similar meanings. The final score of each meaning of a word is the sum of the score obtained in step 1 plus the mean value (heuristically selected) of the score obtained in step 2. Of all the potential meanings of the word in question, the one with the highest score is selected as the candidate meaning.

This phase of the operation determines the meaning of the words, hence their disambiguation, to avoid possible errors in text comprehension.

3.4.3 Information RA. Texminer allows for searches to be made through the ontology, as well as through the lexical database developed, to take advantage of the functionalities of summaries for the purposes of text retrieval.

4. Results

4.1 Experiment

To evaluate the model we pre-processed texts, using as a reference various clustering evaluation experiments conducted by Arco (2008) and Hennig *et al.* (2008). First, the bag of words technique was applied to the sentences. Then, we calculated the mean scores for term frequency (TF) and term frequency-inverse document frequency (TF-IDF) (Salton and Buckley, 1988) of the bag of words for each sentence.

According to the VSM a document is $d \in D$, and is represented as a vector $d = \rho(d) = (w_1, \ldots, w_m)T \in R = R_m +$, where each dimension corresponds to a term in the document, and w_i denotes the weight of the ith term. The set of the m indexed terms is $V = \{t_1, \ldots, t_m\}$ (Lanquillon, 2002). A modification of the formula allows each sentence of the document to become the object of the VSM, so as to provide a proper VSM representation of each sentence structure. Vector representation ignores the sequence in which the sentences appear in a document, and is able to group sentences and their principal terms without taking into account the rhetorical structure of the text.

This technique is computationally very efficient. However, afterwards it is necessary to implement dimensionality reduction techniques, as each inflection of a word is a possible feature, and the number of potential features may be unnecessarily large. The goal of dimensionality reduction is to reduce the number of features that are finally used to represent ontology classes and sentences in each document. The bag of words assigns weights to sentences and adds TF and TF-IDF (Montalvo *et al.*, 2006). The idea of using the TF-IDF is to reflect the relative importance of a term or sentence compared to other sentences in the summary of a document in the bag of words. The expression used for calculating the TF-IDF herein, proposed by Berry (2004), allows us to obtain the weight of the ith term in d (the document or summary) according to the following expressions:

$$w(d,t) = tf_d\left(1 + \log_2\left(\frac{n}{n(t)}\right)\right)$$

The calculation procedure for the bag of words can be easily used in this experiment by adding each value of TF-IDF and assigning weights to sentences underlying the bag of words. No evaluation within text mining – particularly in the field of automatic summarization – is complete without the assessment of the characteristics of the texts that make up the corpus of the experiment. Therefore, it is essential to define thresholds and weight designators regulating positions of sentences. In Texminer these weights are assigned in accordance with the rules of discourse, combined with selection criteria that are explained below (Leiva Mederos, 2012).

The sentences in each section are labeled with values [1, 2, 3, 4]: 1 for the sentence with more semantic meaning, 2 for the subordinate clause to the sentence with more semantic meaning, 3 for substantive subordinate clauses, with more semantic meaning, and 4 for a synthesis sentence of each paragraph of text. This score is achieved by the labeling process conducted by means of the metric classification tool (Domínguez-Velasco, 2013), the text and ontology. The sentence length is also measured, ignoring binary features but counting XML tags. All elements of the sentence can be included and used in its measurement.

After this process, statistical thresholds are assigned to each sentence to build a strategy of exclusion. It is important to note that in this experimental phase, contrast corpuses produced by the PuertoTex system (Leiva-Mederos *et al.*, 2012) were used. This tool performs a summary extract from the main sentences of each section of the text, thanks to its labeling system, which makes it easier to obtain summaries of average size. The corpus of evaluation was Klep, a set of documents in Spanish on Port and Coastal Engineering. It provides extracts created by experts in the field, obtained from the PuertoTerm project and built specifically for this research.

4.2 Evaluation

Following Hennig *et al.* (2008), the results were analyzed using an support vector machine (SVM) to classify all the chains extracted from the text. In order to classify each chain, it is vital to have previously annotated texts by librarians with expertise in the Port and Coastal Engineering domain. When SVM is applied, the vectors that characterize sentences adopt values between [0, 1]. Moreover, we established the parameter SVM C to ensure equity between the adjustment of the training data set and the generalization of the model. The adjustment value taken to achieve this balance is 2. To quantify J (weighting errors in positive examples), we adopted a value of 6, as quantifying J makes it possible to achieve a greater retrieval of sentences constructed by the expert librarians, a point brought out previously by Leskovec *et al.* (2005).

Meanwhile, an overall cross-evaluation was undertaken, using topics from the ontology. Automatic grouping was thus generated from a set of concepts or sentences selected from the topic. The sentences were ordered according to the value obtained in the SVM classification, and all those classified as positive grouping cases, with the length desired for the summary, were extracted. Sentences could then be located in the documents according to their hierarchy.

Evaluation of the model entailed the use of the F-measure, a statistical indicator of precision and recall values in the field of information retrieval, adapted to validate the classification of sentence groups. Precision (P_r) and recall (R_e) are calculated for a given group j and class i using the expressions $P_r(i, j) = n_{ij}/n_j$ and $R_e(i, j) = n_{ij}/n_i$, respectively. The F-measure is computed from the values of precision and recall. In turn, the influence of precision and recall in its calculation depends on a threshold α ($0 \leqslant \alpha \leqslant 1$) (Frakes and Baeza-Yates, 1992). A global value, overall F-measure, is calculated using the weighted average of the maximum values of F-measure for all groups (Steinbach *et al.*, 2000). The F-measure registers the grouping coincidence when comparing the classification obtained by the ontology and the baseline or reference corpus (Rosell *et al.*, 2004). Larsen and Aone (1999) proposed a variant of F-measure for a hierarchical clustering, taking a per class maximum value of F-measure for all groups at all levels of the hierarchy. Variants of precision and recall – micro-averaged precision and micro-averaged recall – are used to evaluate the clustering (Hu *et al.*, 2004). The expressions for the calculation match if each object belongs to only one group and the reference classification also has a unique classification for each object.

The assessment of the VSM, as in the work of Hennig *et al.* (2008), requires that data be organized by subject area to evaluate precision, recall and $F-1$. Rouge (Lin, 2004a) is calculated for the measurement of the quality of automatic summaries, with respect to extracts made by librarians. Rouge metrics also include three contrast tests for the automatic evaluation of summaries: Rouge L, Rouge W and Rouge-SU. We did not include Rouge-SU because it is focused on the operational disabilities of Rouge-S; moreover, it ignores sentences where no pair of words co-occurs in a given reference, unusual for a text in Spanish.

The metrics used in this work, then, are:

- Rouge-L (Longest Common Subsequence): it allows evaluation of text summarization through the length of the word sequences generated from the sub-sequences of two summaries.
- Rouge W (Weighted Longest Common Subsequence): this measure is based on classical LCS. It calculates the spatial relationships that occur when there are sub-sequences of *n*-grams located within other sub-sequences.
- Rouge-S (Skip-Bigram Co-occurrence Statistics): used to obtain the bi-grams that occur in two texts, one reference and one candidate. Comparison of Rouge-S and Rouge LCS shows that LCS only analyzes common sub-sequences, omitting expressions from the analysis that might promote terminological combinations.

Finally, the evaluation includes assessment of abstracts by 12 librarians, considering textual cohesion, coherence of the sentences and domain adequacy. Each expert evaluated 12 summaries developed by the ontology, the lead, baseline and Klep. The mean score for each category was in the range between 5 and 10. While clearly subjective, this procedure of a human nature facilitated pragmatic assessment of the study. As explained above, Klep is a corpus labeled by domain experts, the collection prepared for this research having summaries varying in length from 100 to 200 words.

Table I shows the averages obtained when evaluating the precision, recall and F-measure. It is clear that the summaries created with the help of the ontology are superior to those generated by baseline. The micro and macro dimensions of precision and recall make it possible to verify the effectiveness of the ontology in terms of these two parameters. The F-measure metric recorded better results in this analysis, related to the precision and recall values obtained. These values are superior in the ontology, and therefore affect the difference between the results of the F-measure and baseline. It is also evident that the lowest values are obtained when evaluating the average

Features	Avg.	Prec.	Rec.	F-measure
Baseline	Macro	0.789	0.804	0.799
	Micro	0.736	0.827	0.722
Ontology	Macro	0.719	0.890	0.804
	Micro	0.653	0.734	0.693

Table I.
Precision, recall and F-measure

macro, this being the case for both baseline and ontology. However, regarding this parameter, the ontology gives better values for recall, precision and *F*-measure.

Table II shows the result of evaluating summaries for this median length using the metrics called Rouge (Lin, 2004b), a measure able to automatically evaluate the production of *n*-grams between a candidate summary and a set of reference summaries. Supported by the functionality of its knowledge structure system, the ontology displays better values obtaining *n*-grams. Rouge assessment conducted in the reference corpus from the *n*-grams of sentences that hold more semantic meaning, denominated in computational linguistics as core sentences, shows that the results obtained with the model surpass the quality of the ones obtained with the data set Klep, unlike the findings from Hennig *et al.*'s (2008) study.

It is clear that, in order to obtain better results, the classifier must be prepared to work with any data set or corpus. Otherwise, the results will be poor, even if the corpus (in this case Klep) offers flexibility and accuracy when working with the sentences of the ontology.

In assessing the variations of Rouge, Table III illustrates the superiority of abstracts obtained through the assistance of the ontology. When evaluating summaries with Rouge L, we see that high averages for matching in string summaries are obtained with respect to the baseline summaries. Klep corpus summaries have lower levels of matching. The variant Rouge W shows that the ontology achieved a number of relationships in the summaries that evidence the existence of sub-sequence *n*-grams, meaning a higher quality summary when compared with baseline and Klep corpus.

	Rouge-1 *F*-measure		Rouge-2 *F*-measure	
Feature set length	100	200	100	200
Lead sentences	0.4676	0.6203	0.1204	0.2333
Baseline	0.5166	0.6218	0.1237	0.2122
Ontology	*0.5836*	*0.6716*	*0.2955*	*0.3823*
Klep	0.5063	0.5821	0.28779	0.36922

Table II.
Rouge metrics

Feature set	*Rouge L1F1*		*Rouge L2F1*	
Length	100	200	100	200
Lead	0.6518	0.6822	0.615	0.7733
Baseline	0.6	0.5478	0.6123	0.7342
Ontology	0.8223	0.8721	0.8671	0.8944
Klep	0.6821	0.8937	0.7759	0.8146
	Rouge W1F1		*Rouge W2F1*	
Length	100	200	100	200
Lead	0.769	0.6523	0.6945	0.792
Baseline	0.7667	0.742	0.745	0.7111
Ontology	0.833	0.8425	0.89	0.8826
Klep	0.6239	0.6013	0.6123	0.5621
	Rouge S1F1		*Rouge S2F1*	
Length	100	200	100	200
Lead	0.8123	0.7	0.8417	0.83
Baseline	0.8596	0.863	0.8693	0.8397
Ontology	0.8894	0.8961	0.8913	0.8956
Klep	0.6225	0.711	0.8215	0.866

Table III.
Evaluation of variants of Rouge

Evaluating Rouge-S demonstrates that the ontology obtains more bi-grams within the summaries generated by the ontology and the contrast corpus, and the results are lower in the corpus Klep. In all cases, the assessment of variants Rouge, lead and baseline shows favorable results for the Klep collection. This comes to demonstrate that the corpus labeled for this experiment was not able to generate summaries of quality.

When evaluating the categories cohesion, coherence and domain perspective, we note that abstracts generated by the ontology have more textual qualities because the ontology features a disambiguation system that enhances the pragmatic quality of its summaries. The lead also exhibits very good results in consistency. Meanwhile, the baseline corpus and Klep yield very low rates of textual quality, according to the opinion of the evaluators (Figure 5).

5. Conclusions

The paper emphasizes that automatic summarization of texts can be achieved using rich knowledge sources. In this case, thanks to ontologies, main sentences can be selected with a broad rhetorical structure, especially for a specific knowledge domain. It is therefore feasible to generate summaries or extracts of acceptable value, because the quality of these products can be checked in terms of precision, recall, *F*-measure and Rouge, all useful statistical metrics for the automatic evaluation of summaries. Still, this does not ensure that the summary obtained will be legible and understandable, i.e. endowed with pragmatic quality. This is a highly complex issue when dealing with documents written in Spanish. For this reason, summarization methods are usually complemented by anaphora resolution systems.

The summaries produced by selecting sentences have strengths and weaknesses. On the one hand, the type of text to be summarized and the desired summary length are key. With these procedures it is possible to summarize news items or relatively brief texts. Yet the resulting summaries are almost always incoherent, unbalanced and hardly pragmatic.

Such a summarization method is useful only when it is meant to process a single document, and the conditions for effectiveness are reduced to the existence of a tagged corpus, a lexical knowledge base and a set of algorithms that treat texts to obtain further summary extracts. Even so, the results obtained are still far below the requirements of humans. This technique cannot be considered a substitute for the classic text extraction models we have been using for some time.

When using a classifier trained in the selection of sentences, the results obtained with the ontology can surpass those achieved with a corpus labeled for a specific

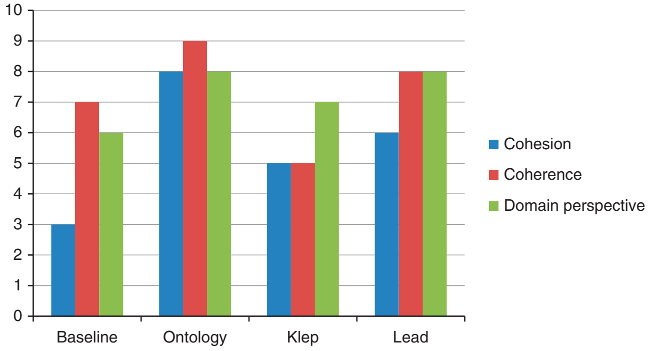

Figure 5.
Cohesion, coherence and domain perspective

domain. In this case, we would not have to look toward improvements in precision and recall to achieve better results in the selection of sentences.

The combination of purely statistical strategies together with socio-cognitive techniques facilitates automatic summarization of scientific texts in the mentioned domain.

The results of Rouge metrics for evaluating an *n*-gram demonstrate the capabilities of our system in summarizing texts in the specialized domain described here. At the same time the results highlight its inefficiency when summarizing texts in different knowledge domains subjected to unmatched rhetoric features.

References

Alonso, L. and Fuentes, M. (2003), "Integrating cohesion and coherence for text summarization", *Proceedings of the EACL'03 Student Session, Budapest*, pp. 1-8.

Alpcan, T., Bauckhage, C. and Agarwal, S. (2007), "An efficient ontology-based expert peering system", *Proceedings of the IAPR Workshop on Graph-Based Representations*, pp. 273-282.

Anaya, H., Pons, A. and Berlanga, R. (2006), "Una panorámica de la construcción de extractos de un texto", *Revista Cubana de Ciencias Informáticas*, Vol. 1 No. 1, pp. 55-65.

Andreasen, T. and Bulskov, H. (2009), "Conceptual querying through ontologies", *Fuzzy Sets and Systems*, Vol. 160 No. 5, pp. 2159-2172.

Arco, L. (2008), "Agrupamiento basado en intermediación diferencial", PhD thesis, Universidad Central "Marta Abreu" de las Villas, Santa Clara.

Aretoulaki, M. (1997), "COSY-MATS: 'an intelligent and scalable summarisation shell'", *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, Madrid*, pp. 74-81.

Austin, J.L. (1962), *How To Do Things With Words*, Clarendon Press, Oxford.

Barzilay, R. and Elhadad, M. (1997), "Using lexical chains for text summarization", *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, Madrid*, pp. 10-17.

Berry, M. (2004), *Survey of Text Mining: Clustering, Classification, and Retrieval*, Springer, New York, NY.

Chen, P. and Verma, R. (2006), "A query-based medical information summarization system using ontology knowledge", *Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, pp. 37-42.

D'Cunha, I. (2006), "Hacia un modelo lingüístico de resumen automático de artículos médicos en español", PhD thesis, Universitat Pompeu Fabra, Barcelona.

Domínguez-Velasco, S. (2010), *Protex Beta Software*, Departamento de automática, Universidad Central de las Villas, Santa Clara.

Domínguez-Velasco, S. (2013), *Metric Beta Software*, Universidad Central "Marta Abreu" de las Villas, CDICT, Santa Clara.

Endres-Niggemeyer, B., Maier, E. and Sigel, A. (1995), "How to implement a naturalistic model of abstracting: four core working steps of an expert abstractor", *Information Processing & Management*, Vol. 31 No. 5, pp. 631-674.

Frakes, W.B. and Baeza-Yates, R. (Eds) (1992), *Information Retrieval: Data Structures & Algorithms*, Prentice Hall, New York, NY.

Gaizauskas, R., Herring, P., Oakes, M., Beaulieu, M., Willett, P., Fowkes, H. and Jonsson, A. (2001), "Intelligent access to text: integrating information extraction technology into text browsers", *Proceedings of the Human Language Technology Conference, San Diego, CA*, pp. 189-193.

Gil-García, R. and Pons-Porrata, A. (2008), "Hierarchical star clustering algorithm for dynamic document collections", *CIARP 2008*, pp. 187-194.

Goldstein, J., Kantrowitz, M., Mittal, V. and Carbonell, J. (1999), "Summarizing text documents: sentence selection and evaluation metrics", *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99), ACM, New York, NY*, pp. 121-128.

Goldstein, J., Mittal, V., Carbonell, J. and Callan, J. (2000), "Creating and evaluating multi-document sentence extract summaries", *Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM '00), ACM, New York, NY*, pp. 165-172.

Grimes, U.E. (1975), *The Thread of Discourse*, Mouton, The Hague.

Halliday, M.A.K. and Hasan, R. (1976), *Cohesion in English*, Longman, Essex.

Havens, T.C., Keller, J.M., Popescu, M. and Bezdek, J.C. (2008), "Ontological self-organizing maps for cluster visualization and functional summarization of gene products using gene ontology similarity measures", *IEEE International Conference on Fuzzy Systems (FUZZ 2008), Hong Kong, June 1-6*.

Hennig, L., Umbrath, W. and Wetzker, R. (2008), "An ontology-based approach to text summarization", *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 291-294.

Hjørland, B. (2002), "Epistemology and the socio-cognitive perspective in information science", *Journal of American Society of Information Science*, Vol. 53 No. 4, pp. 257-270.

Hu, P., He, T., Ji, D. and Wang, M. (2004), "A study of Chinese text summarization using adaptive clustering of paragraphs", *Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04), IEEE, ACM, Wuhan, September 14-16*.

Huang, H.H. and Kuo, Y.H. (2007), "Towards auto-construction of domain ontology: an auto-constructed domain conceptual lexicon and its application to extractive summarization", *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, IEEE, Hong Kong*, pp. 2947-2952.

Hung, J. (2008), "RETRACTION: a new WSD approach using word ontology and concept distribution", *Journal of Information Science*, Vol. 34 No. 22, pp. 231-253.

Lanquillon, C. (2002), "Enhancing text classification to improve information filtering", PhD thesis, Otto-von-Guericke-Universität Magdeburg, Magdeburg.

Larsen, B. and Aone, C. (1999), "Fast and effective text mining using linear-time document clustering", *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, NY*, pp. 16-22.

Lee, C.S., Chen, Y.J. and Jian, Z.W. (2003), "Ontology-based fuzzy event extraction agent for Chinese e-news summarization", *Expert Systems with Applications*, Vol. 25 No. 3, pp. 431-447.

Leiva, A., Senso, J.A., Domínguez, S. and Hípola, P. (2009), "An automat for the semantic processing of structured information", *ISDA 9th International Conference of Design of Software and Application, IEEE, Pisa, November 30-December 3*.

Leiva-Mederos, A. (2012), "Texminer: un modelo para la extracción y desambiguación de textos científicos en el dominio de Ingeniería de Puertos y Costas", PhD thesis, Universidad de Granada, Granada.

Leiva-Mederos, A., Domínguez-Velasco, S. and Senso, J.A. (2012), "PuertoTex: un software de minería textual para la creación de resúmenes automáticos en el dominio de ingeniería de puertos y costas basado en ontologías", *TransInformação*, Vol. 24 No. 2, pp. 103-115.

Lesk, M. (1986), "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone", *Proceedings of SIGDOC*.

Leskovec, J., Milic-Frayling, N. and Grobelnik, M. (2005), "Impact of linguistic analysis on the semantic graph coverage and learning of document extracts", in Veloso, M. and Kambhampati, S. (Eds), *Proceedings of the 20th national Conference on Artificial Intelligence*, ACM Press, Pittsburgh, Vol. 3, pp. 1069-1074.

Lin, C. (2004a), "Looking for a few good metrics: automatic summarization evaluation – how many samples are enough?", *Proceedings of the NTCIR Workshop 4, Tokyo.*

Lin, C. (2004b), "Rouge: a package for automatic evaluation of summaries", *Proceedings of the Workshop on Text Summarization Branches Out (WAS'04), Barcelona*, pp. 25-26.

Lin, C. and Hovy, E. (1997), "Identifying topics by position", *Proceedings of the ACL Applied Natural Language Processing Conference, Washington DC*, pp. 283-290.

Lin, F.R. and Liang, C.H. (2008), "Storyline-based summarization for news topic retrospection", *Decision Support Systems*, Vol. 45 No. 3, pp. 473-490.

Luhn, H. (1958), "The automatic creation of literature abstracts", *Journal of Research and Development*, Vol. 2 No. 2, pp. 159-165.

Mani, I. and Bloedorn, E. (1999), "Summarizing similarities and differences among related documents", *Information Retrieval*, Vol. 1 Nos 1-2, pp. 35-67.

Mann, W.C. and Thompson, S.A. (1988), "Rhetorical structure theory: toward a functional theory of text organization", *Text*, Vol. 8 No. 3, pp. 243-281.

Marcu, D. (1998), "The rhetorical parsing, summarization, and generation of natural language texts", PhD thesis, University of Toronto, Toronto.

Marcu, D. (2000), *The Theory and Practice of Discourse Parsing Summarization*, Massachusetts Institute of Technology, Cambridge, MA.

Mateo, P., González, J.C., Villena, J. and Martínez, J.L. (2003), "Un sistema para resumen automático de textos en castellano", *Procesamiento del Lenguaje Natural*, Vol. 31, pp. 29-36.

Montalvo, S., Navarro, A., Martínez, R., Casillas, A. and Fresno, V. (2006), "Evaluación de la selección, traducción y pesado de los rasgos para la mejora del clustering multilingüe", *Campus Multidisciplinar en Percepción e Inteligencia (CMPI 2006) – 50 Años de Inteligencia Artificial*, Vol. 2, pp. 769-778.

Nomoto, T. and Matsumoto, Y. (2001), "A new approach to unsupervised text summarization", *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01), ACM, New York, NY*, pp. 26-34.

Ono, K., Sumita, K. and Miike, S. (1994), "Abstract generation based on rhetorical structure extraction", *Proceedings of the International Conference on Computational Linguistics, Kyoto,* pp. 344-348.

Pinto, M. (2001), *El resumen documental: principios y métodos*, Fundación Germán Sánchez Ruipérez, Madrid.

Popescu, M., Keller, J.M., Mitchell, J.A. and Bezdek, J.C. (2004), *Functional Summarization of Gene Product Clusters Using Gene Ontology Similarity Measures*, ISSNIP, *Proceedings of the 2004 Intelligent Sensors, Sensor Networks and Information Processing Conference*, *IEEE, New York.*

Rosell, M., Kann, V. and Litton, J. (2004), "Comparing comparisons: document clustering evaluation using two manual classifications", *Proceedings of ICON 2004, 3rd International Conference on Natural Language Processing, Hyderabad, December 19-22.*

Salton, G. and Buckley, C. (1988), "Term weighting approaches", *Automatic text Information Processing and Management*, Vol. 24 No. 5, pp. 513-523.

Searle, J. (1969), *Speech Acts. An Essay in the Philosophy of Language*, Cambridge University Press, Cambridge.

Steinbach, M., Karypis, G. and Kumar, V.A. (2000), "Comparison of document clustering techniques", *KDD Workshop on Text Mining*, Vol. 400 No. 1, pp. 525-526.

Teufel, S. and Moens, M. (2002), "Summarizing scientific articles: experiments with relevance and rhetorical status", *Computational Linguistics*, Vol. 28 No. 4, pp. 409-445.

Wu, K., Li, L., Li, J. and Li, T. (2013), "Ontology-enriched multi-document summarization in disaster management using submodular function", *Information Sciences*, Vol. 224, pp. 118-129.

Yoo, I., Hu, X. and Song, I. (2006), "Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering", *Proceedings of ACM SIGKDD, ACM*, pp. 791-796.

Yuan, S.T. and Sun, J. (2004), "Ontology-based structured cosine similarity in speech document summarization", *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), IEEE, ACM, Beijing, September 20-24*.

Zhang, Z., Huang, Z. and Zhang, X. (2010), "Knowledge summarization for scalable semantic data processing", *Journal of Computational Information Systems*, Vol. 6 No. 12, pp. 3893-3902.

Corresponding author
Professor Pedro Hípola can be contacted at: phipola@ugr.es