GReAT
A Model for the Automatic Generation of Text Summaries

Claudia Gomez Puyana and Alexandra Pomares Quimbaya
Systems Engineering, Pontificia Universidad Javeriana, Bogotá, Colombia

Keywords: Text Mining, Summary Generation, Natural Language Processing.

Abstract: The excessive amount of available narrative texts within diverse domains such as health (e.g. medical records), justice (e.g. laws, declarations), assurance (e.g. declarations), etc. increases the required time for the analysis of information in a decision making process. Different approaches of summary generation of these texts have been proposed to solve this problem. However, some of them do not take into account the sequentiality of the original document, which reduces the quality of the final summary, other ones create overall summaries that do not satisfy the end user who requires a summary that is related to his profile (e.g. different medical specializations require different information) and others do not analyze the potential duplication of information and the noise of natural language on the summary. To cope these problems this paper presents GReAT a model for automatic summarization that relies on natural language processing and text mining techniques to extract the most relevant information from narrative texts focused on the requirements of the end user. GReAT is an extraction based summary generation model which principle is to identify the user's relevant information filtering the text by topic and frequency of words, also it reduces the number of phrases of the summary avoiding the duplication of information. Experimental results show that the functionality of GReAT improves the quality of the summary over other existing methods.

1 INTRODUCTION

During the last thirty years the information systems have stored huge amounts of information in different formats. In some domains such as health (e.g. medical records), justice (e.g. laws, declarations), assurance (e.g. declarations) and research (e.g. research articles) a lot of this information is stored as narrative texts, hindering its use for decision making processes. The process of discovering the knowledge contained in these texts, or creating new hypotheses according to them include human and time expensive tasks (Inniss et al., 2006),(Mohammad et al., 2009) that cannot be afforded by most organizations. To face these problems of narrative information overload, different approaches of summary generation have been proposed, however, our investigation found out that these approaches are not enough to create a summary out of a text. Initially, there are mainly two approaches to perform this task: *Statistical and Linguistic Methods*. The statistical methods are independent of the language. For example, they are based on the frequency of words, or on heuristics such as taking into account the title, headings, position and length of the sentence. On the other hand, linguistic methods include discourse structure and lexical chains (Zhan et al., 2009), for example, some proposals of this method use *Clustering* based strategies that group phrases with similar characteristics, however, they have had accuracy problems due to the ambiguous terms within the language. Some commercial tools are based upon basic statistical approaches, and rely heavily on a particular format or writing style, such as the position in the text or some lexical words(Park et al., 2008). However, most of the tools rely on methods like centroid-based, position-based, frecuency-based summarization or keywords to extract relevant sentences into the summary as MEAD, Dragon ToolKit, LexRank (Gunen, 2004). On the other hand, there are methods and tools based typically on *abstraction*, which are more complicated compared to approaches based on *extraction*, since there are still problems regarding semantic representation, inference and natural language generation (Zhang, 2009). *ML (Machine Learning) and IR (Information Recuperation) Algorithms* need to have a good *similarity measure* between documents, due to ambiguities in used vocabulary. That is, if two documents or publications discuss the same topic, they may use different vocabulary while being semantically similar. Generally, these are some of the

Gomez Puyana C. and Pomares Quimbaya A..
GReAT - A Model for the Automatic Generation of Text Summaries.
DOI: 10.5220/0004546028002888
In *Proceedings of the 15th International Conference on Enterprise Information Systems* (ICEIS-2013), pages 280-288
ISBN: 978-989-8565-59-4

problems seen in the proposals of the consulted literature. To cope these problems this paper presents "GReAT" a model for automatic summarization that relies on natural language processing and text mining techniques to extract the most relevant information from narrative texts focused on the requirements of the end user, and the quality and coherence of the summary. GReAT is an extraction based summary generation model which principle is to identify the relevant categories to the user across the text to use them for reducing the number of phrases included in the summary and to avoid the redundancy of data. Experimental results show that the functionality of GReAT improves the quality and coherence of the summary over other existing methods. This paper is organized as follows: Section 2, presents *Preliminary Concepts* on text automatic summarization techniques, Section 3 shows the *General Process of the Proposed Model - Great* for automatic generation of text summaries, Section 4, compares GReAT with important proposals and illustrates the main strategies for text summary generation, Section 5 evaluates the functionality of GReAT through its application in a Case of Study in the Health domain in Spanish language and finally, Section 6 presents the *Conclusions* and *Future Work*.

2 PRELIMINARY CONCEPTS

The main concepts that must be kept in mind to understand the proposed model are the following:

Summary: The objetive of a summary is to extract a smaller size document but keeping the relevant information from the original document, i.e, automatically create a comprehensible version of a given text, providing useful information to the user (Kianmehr et al., 2009). There are different types of summaries, including: **Single Document:** The summary is generated from a single input text document. **Multi-Document:** The summary is generated from multiple input documents (Zhang, 2009). **Specific Domain:** The summary is generated from multiple input documents, but considering the context or domain in which it was written. **General or Generic:** Its purpose is to try to cover as much content as possible, preserving the organization of general topics of the original text. There are two strategies to extract phrases and generate a summary: (Chang and Hsiao, 2008). **Summary Based on Extraction:** Its purpose is to generate the summary with phrases that are included literally. This strategy produces a summary by selecting a subset of sentences from the original document. **Summary Based on Abstraction:** It is a more difficult task, because the information in the text is re-phrased considering semantic representation and natural language generation (Genest and Lapalme, 2011). Moreover, in the literature there are different approaches to generate summaries out of a text, they include: **Statistical or Probabilistic approaches:** They are based on the frequency of terms to determine the importance of the term. **Semantic-based or Linguistics approaches:** These are based on the incorporation of some form of natural language processing to generate summaries, considering the semantics and linguistics. **Heuristic based approaches:** The most commonly used heuristic techniques are: *Cues-words, Key-words, Title-words, Synonyms y Location / Position* (Dalal and Zaveri, 2011). **Oriented-Questions or Topic approaches:** It focuses on a user's topic of interest, extracting text information that is related to the specific topic. Its approach is to identify significant topics in the data set and generate the topical structure based on these topics. **Cluster-based Approaches:** These are based on forming sentence clusters grouped by similarity measures between sentences. The number of clusters is more or less equal to the number of topics covered in the text (Gunen, 2004). The proposed GReAT model is within the following characteristics: **Multi-Document, Specific Domain, Topic-oriented and Based-Extraction Summary**. There are several challenges and opportunities identified in the literature which will be addressed by the solution proposed in this document regarding to the automatic generation of text summaries, such as: *i)* improve the quality and coherence of the summaries, *ii)* adapt summaries to every user according to their needs and granularities of information, taking into account the different topics that are often important to a person in a specific domain, *iii)* keep the sequentiality of the original document, *iv)* use an alternative way to extract the relevant information instead of training examples from domain of knowledge, and finally, *v)* detect noise in text collections due to the use of natural language. The next section details these challenges and how they are addressed by the proposed model.

3 GReAT

GReAT is a model that produces summaries with the following characteristics: **Multi-Document, Specific Domain, Topic-oriented and Based-Extraction Summary**. It uses several techniques of Information Retrieval and Natural Language Processing (NLP), such as: **Tokenization, Chunking, Named Entity Extraction**, among others. This approach takes into account some considerations for extracting knowl-

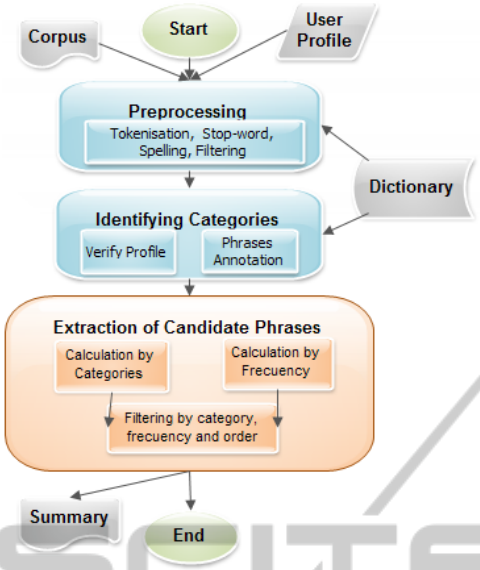

Figure 1: GReAT Model.

edge or generating an adequate summary such as: *quality, consistency, adaptation, duplicity, user profile, size, noise and sequentiality*, which can be seen in Table 1. Additionally, GReAt uses techniques that have not been fully addressed, such as those based on topics. These considerations are addressed as follows: **i) Quality.** to improve the quality GReAT compare similar information in the texts and applies filtering techniques to avoid the duplication of information. **ii) User:** to adapt summaries to every user according to their needs and granularities of information, it takes into account the different topics of interest selected by the user. **iii) Sequentiality:** to keep the sequentiality of the original document it extracts the sentences while taking into account the same chronological order in which they were stored and allowing the user to select the more or less recent text to the summary. **iv) Relevance:** It uses an alternative to extract the relevant information instead of training examples using the knowledge domain and filtering the sentences by topic and frequency of words. **v) Noisy:** to detect noise in text collections it applies spelling techniques.

The steps or phases that are part of the *Great model* are presented in Figure 1. It consists of three main steps: *Preprocessing, Identifying Categories and Extracting Candidate Phrases*. The Preprocessing step is divided into several processes: *Stop-words, Tokenization, Spelling and Filtering* .

3.1 Preprocessing

This phase is divided into four main processes: *Tokenization, Stop-words, Spelling and Filtering*, which are detailed below:

Tokenization. In order to obtain a summary based on the needs of a user, it is necessary to divide the text into phrases that allow us to process information independently (Hotho et al., 2005). The output is the set of sentences.

Stop-words. Some words from the texts do not provide relevant information, such as articles, prepositions, among others. To address the problem of high dimensionality that commonly occurs in a text mining process, it has been proposed to delete the words with low relevance to the language with the technique known as *Stop-Words* (Brun, 2004). This process requires a dictionary of stop-words of the language. The result are the sentences without the *Stop-Words*.

Spelling This phase performs a spelling of narrative text, which involves taking the text input and provides a corrected text, restoring texts spaces without space. Spell Checking is performed on the *Noisy-channel model*, which models user errors (typographical) and expected user input (based on data) (Daumé and Marcu, 2002). The errors are modeled by the weights of the *Edit Distance technique* and the expected input by model language characters. Edit distance measures the minimum number of edit operations to transform one string into another (Wang et al., 2009). The result is the set of correctly spelled phrases.

Filtering. By having the information divided into parts (phrases) without stop-words plus the words spelled correctly and in their roots, it is assumed that the information is in a suitable form, therefore this process proceeds to remove the redundancy of information found in these texts. To achieve this, we use a technique known as *Fixed Weight Edit Distance* where the simplest form of weighted edit distance simply sets a constant cost for each one of the edit operations. This algorithm will maximize the number of matches between the sequences along the entire length of the sequences (Mehdad et al.,), improving the process performance. At this point, *Fixed Weight Edit Distance technique* is used for purposes of eliminating duplication of information, because the comparison in the previous section is used along with a Spanish language dictionary to correct the words orthographically. The output of this stage is the removal of common phrases to each other, leaving only one instance of them. This step is optional.

3.2 Identifying Categories

This phase is divided into two main processes: *Verifying Profile and Phrases Annotation*:

Verifying Profile. As users may have different infor-

mation needs, at this point of the process a set of categories of the specific domain knowledge is defined. Thus the user will be able to select which one these categories he wants to find according to his needs. For example, the categories of Health domain of the case of study that the user can select are: *Drugs, Diseases, Exams*, etc. Once the user selects the categories he wants to find, we proceed to make the information search on these categories. For this, it performs a process consisting of annotating the words contained in the sentences resulting from the preprocessing phase, which is explained in the following subsection:

Phrases Annotation. To assign the annotations, there is a technique known as *Named Entity Extraction*, which involves supervised training of a statistical model, or more direct methods such as dictionary or regular expressions to classify texts or phrases of a document within a category. We will use a dictionary of specific domain with the technique called *Chunking based on Dictionary*, which aims to find adjacent words that make sense being together in a sentence. One example is "diabetes mellitus type I" in the Health domain. This phase will be based upon the implementation of the *Matching Text Strings Aho-Corasick algorithm*, which consists in finding all the alternatives of words against a dictionary independently of the number of matches or the size of the dictionary (Tran et al., 2012). The output of this phase is each phrase, along with the set of annotated words of the corresponding category, applying heuristics to eliminate phrases that have no set of annotated words on the categories selected by the user. See Algorithm 1.

Algorithm 1: Identifying Categories.

Require: $S(f) \leftarrow f \in S$: Corrected sentences set f of a document D. Where every sentence f is composed a set of words p.
Require: $D(c) \leftarrow c \in D$: Dictionary of domain terms t of each category c user-selected.
Ensure: $F(a)$: Set of sentences in the document D with p words annotated belonging to the category c.
  for all $p \in F(c)$ **do**
    To each p applies operation CH "*dictionary-based Chunking*" to $\forall t \in D(c)$
    if $CH(p) = 1$ **then**
      Annotate the word p to the category c and added the annotation-word p to the set $F(a)$.
    end if
  end for
  return $F(a)$

This algorithm requires two input data: the set of corrected sentences of a document and the dictionary of domain terms of each category selected by the user. The output of the algorithm is the set of sentences in the document with annotated words belonging to each category selected by the user. The process starts iterating each word of the sentences of a document, applying the **"Dictionary-based Chunking technique"**, which purpose is to find the category to which the word belongs. If the result of this operation is equal to 1, this word is annotated into the found category and the sentence to which this word belongs is added to the final set of sentences.

3.3 Extraction of Candidate Phrases

This phase involves finding key phrases to form the summary. After the preprocessing phase, which purpose was to debug the information and the phase of identification of phrases were realized, a list of phrases with annotated words in the categories chosen by the user is selected to extract the most relevant and adequate sentences. If the results are more than the expected by the user, it may require further filtering of information depending on the size of the summary. For this, the selection of phrases is based on three steps: *calculation by topics, calculation by frequency and filtering by categories, frequency and order*:

Calculation by Categories. To find the most relevant information, the technique known as *LDA (Latent Dirichlet Allocation)*, will be used to automatically discover topics within the phrases. LDA represents documents as mixtures of topics containing words with certain probabilities. LDA makes the assumption that the number of subjects is the same as the number of items or the number of events describing the corpus, and furthermore, that a complete sentence in a document belongs to one or more subjects, thereby, it calculates the probability that the phrase belongs to the topic (Arora and Ravindran, 2008).

Calculation by Frequency. For each one of the words that were annotated for each sentence, we calculate the *"TF-IDF" frequency* on the complete document, regardless the stop-words. *TF-IDF* is a well known statistic measure used to evaluate how important a word is to a document corpus (Liu et al., 2010).

Filtering by Categories, Frequency and Order. After the process computes the probability of each sentence into a category, the inverse frequency of the frequent words in each sentence in the previous steps, we will calculate the total of annotated words and the number of categories to which the phrase might belong. These calculations of each sentence are summed to form a ranking of sentences. If the maximum number of phrases indicated by the user is greater than the number of annotated sentences obtained in the "Identifying Categories phase", we use the ranking of sentences (probability, the inverse frequency, the number of annotated words and the number of categories) to

filter the phrases, selecting sentences from each category with higher ranking. It is important to mention that it is only allowed to select a phrase once, even if it belongs to more than one category, avoiding duplication of information. Then the phrases are selected by order, i.e. according to the recency of the sentence, note that the sentences are organized from the most to the least recent or vice versa, according to the selection of the user. The result of this phase is the summary of text phrases selected as the most important ones out of the original text. See Algorithm 2.

Algorithm 2: Extraction of Candidate Phrases.

Require: $F(a) \leftarrow a \in F$: Set of phrases f of document D with the annotated words a to c a category.
Require: C: Set of categories c user-defined for the summary.
Require: $|f|$: Total user-defined sentences for the summary.
Require: $|C|$: Total user-defined phrases for each category.
Require: $|a|$: Total of sentences f with annotated words for the selected categories c by the user.
Require: $|p|$: Total words in a sentence belonging to a category.
Require: $|c|$: Total of categories to which a sentence can belong.
Require: $|t|$: Frequency value of each term t on document D, applying the technique "*TF-IDF*" and using phrases with annotated words $F(a)$.
Require: $P(c)$: Value of the likelihood that each sentence f belongs to a category c, applying the technique "*LDA*".
Require: T: Set of values ordered ranking $|t|$ of each sentence obtained by adding: $|t| = P(c) + |p| + |c| + |t|$. If any value is equal to another, the values are organized by the criteria selected by the user (the most recent sentence or less recent).
Ensure: R: Set of phrases selected for the summary R. Since $|R|$ total of added sentences for the summary.
if $|f| > |a|$ {If the total of sentences is greater than annotated sentences maximum total user-defined.} **then**
  for all $c \in C$ {for all user-defined categories.} **do**
    for all $f \in F(a)$ {for all sentences with annotations.} **do**
      for all $|r| \in T$ {for all values ranking of each sentence.} **do**
        while $|R| < |C|$ {while the total of phrases is less than total phrases by user-defined.} **do**
          if $f \ni R$ {if the phrase does not exist in the set of sentences of the summary.} **then**
            The phrase f is added to the set of sentences for summary R.
          end if
        end while
      end for
    end for
  end for
end if
return R

This algorithm requires several input data such as: the set of phrases of a document with the annotated words into a category, the set of categories selected by the user for the summary, the total of sentences for the summary, the total of phrases selected by user for each category, the total of sentences with annotated words for the categories selected by the user, the total of words in each sentence belonging to a category, the total of categories to which each sentence might belong, the value of the frequency of each term of the document, the value of the likelihood that each sentence belongs to a given category, the set of ordered values ranking of each sentence obtained by adding the previous values (specifically, $|t| = P(c) + |p| + |c| + |t|$). If any ranking value is equal to another, the values are organized by the criteria selected by the user (in order from the more or less recent document). The output of this algorithm is the set of phrases selected to the summary. The process starts verifying if the total of sentences selected by the user is greater than the total of annotated sentences by category, then it proceeds to iterate all categories selected by the user, then iterates all sentences with annotations, after that, it iterates all ranking values of each sentence. As long as the total of phrases by category is lower than the total of phrases selected by the user, the process adds up the sentence by category to the final set of sentences if the phrase does not exist in this set. When the iterations are completed, the result will be the set of phrases comprising the text summary.

4 RELATED WORKS

In recent years, *Natural Language Processing (NLP)* has been influential in narrative text extraction. It solved many problems bringing significant benefits from the introduction of robust techniques. Regarding the automatic summarization of narrative texts, there are many efforts based on *heuristics* as an integral part of automatic text summarzation (Dalal and Zaveri, 2011), (Park et al., 2008), (Kianmehr et al., 2009). In addition, *Cross-language* text summaries from trained models (Yu and Ren, 2009) or *Citations-based summaries* that identify the most important aspects of an article or publication (Abu-Jbara and Radev, 2011) as well as the use of *Reduction Rules* (Devasena, 2012). Various tasks such as *Text Mining* and *Information Retrieval* rely on ranking the data items based on their *Centrality or Prestige* in order to summarize a text. Specifically, the work of Radev, which purpose was to evaluate the *Centrality* of each sentence in a cluster and extract the most important to include it in the summary (Gunen, 2004). The hypothesis states that sentences that are similar to many of the other sentences in a cluster are more central (or salient) to the subject. Other type of works such as Qiaozhu (Mei et al., 2010) propose a ranking algorithm called *DivRank*, to balance *diversity and prestige* of the words. Different approaches to

Table 1: Phrase Relevance.

Project	Frequency	Topics	Profile	Orden	Duplicity
GReAT	X	X	X	X	X
(Dalal and Zaveri, 2011)		X	X		
(Ling et al., 2008)	X	X	X		
(Liu et al., 2010)	X	X			
(Chang and Hsiao, 2008)	X				
(Saravanan et al., 2005)	X				
(Bossard et al., 2009)	X				X
(Guelpeli et al., 2011)	X				
(Long et al., 2010)	X	X			
(Devasena, 2012)		X			
(Muthukrishnan et al., 2011)	X	X	X	X	X
(Mohammad et al., 2009)	X				
(Reeve et al., 2006)	X				
(Park et al., 2008)	X	X			
(Gunen, 2004)	X				
(Genest and Lapalme, 2011)		X		X	X
(Kianmehr et al., 2009)	X				
(Zhan et al., 2009)	X	X			

extract summaries from narrative texts have reduced the problem of *information overload*, however, there still are some limitations that should be taken into account to increase the quality of the summaries. The first important issue is to consider the user profile that generates the summary, this aspect will allow generating summaries adapted to the actual requirements of information. The second aspect is to respect the sequentiality of the original document producing well formed summaries. The third aspect is to handle synonyms and avoiding duplication of information to produce compact and right sized summaries. Table 1 shows related works that have used the most common techniques in the text mining area for text summaries generation. The characteristics compared in these tables are: **frequency, topics, profile, order and duplicity** . The symbol "X" indicates that the project contains the characteristics depicted on the table and the meaning of each one is described as follows: **i) Frequency:** It indicates if the project takes into account the frequency of words to extract the relevant phrases to the text summary. **ii) Topics:** It indicates if the project takes into account the topics in the text that are important to the user to create the text summary. **iii) Profile:** It indicates if the project gives importance to the user's information needs. **v) Order:** It indicates if the project keeps the sequentiality of text original to form the text summary. **vii) Duplicity:** It indicates if the project eliminates the duplication of information in the text. In conclusion, no project contains all of the features and these are highlighted by some limitations, which are taken into account in the proposed *GReAT Model*. To emphasize, the paper (Park et al., 2008) proposed a method for summarizing the Web content that attempts to explore the user feedback (comments and tags) in the social bookmarking service. This work applies a feature extraction technique using *TF-IDF method* and heuristics taking into account the titles of the texts. This strategy is not enough to extract the most relevant information in a summary, and has some weaknesses like the handling of the high dimensionality and the absence of methods that consider user needs and sequentiality of original texts. Finally, the (Zhan et al., 2009) approach is based on the text summary regarding the structure of topics from online product reviews that extracts relevant topics. It includes a data preprocessing techniques step such as *Stop-words* and *Lemmatization*, a topic identifying step with the *Text Segmentation technique* based on similarity and frequency of sequences, to finally extract candidate phrases with the *Maximal Marginal Relevance (MMR) method* that reduce the redundancy of information until the summary is presented to the users. However, user requirements, the size of the summary, the sequencing and the duplication of information were not considered.

5 EVALUATION OF GReAT

To evaluate *GReAT*, we applied it in the health domain using data from narrative texts in Spanish language from medical records of patients. The main screen of GReAT System have filter options: **initial date, end date, categories** and **phrases sorted by category**, these may be selected by the user according to his preferences. For the case of study, the options **Gender Category**, **Age Category** and **Keywords Category** were included, as well as the **date range** of a record. As an example, a medical record

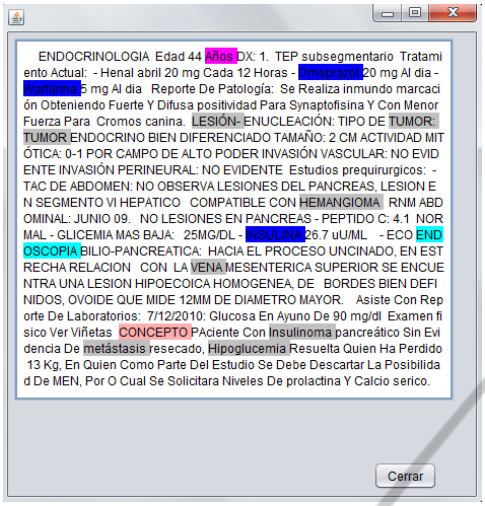

Figure 2: Summary 1.

in chronological order of a patient identified by the number 1679861 will be shown and summarized in a single text following the steps of the proposed GReAT Model: (note that for simplicity and space the result of each step is not shown, just the final result (the text summary)).

Original Medical History: 2011-01-25 13:49:00.000

ENDOCRINOLOGIA Edad 44 aos DX: 1. POP Reseccin de masa pancretica insulinoma 7/10/2010 2. Hipoglicemia hiperinsulinemica 2.1 Insulinoma 3. Sobrepeso IMC 28 3. Hipertensin arterial 4. TEP subsegmentario Tratamiento actual: - Enalapril 20 mg cada 12 horas - Omeprazol 20 mg al dia - Warfarina 5 mg al dia Reporte de patologa: Se realiza inmunomarcacin obteniendo Fuerte y difusa positividad para Synaptofisina y con menor fuerza para Cromogranina. El Ki67 muestra actividad mittica en menos del 5% de las clulas tumorales, el CEA es negativo. El marcador de Insulina fue negativo. Diagnostico: pncreas. lesinenucleacin: tipo de tumor: tumor endocrino bien diferenciado tamao: 2 cm actividad mittica: 0-1 por campo de alto poder invasin vascular: no evidente invasin perineural: no evidente estudios prequirurgicos: - tac de abdomen: no observa lesiones del pancreas, lesion en segmento vi hepatico compatible con hemangioma rnm abdominal: juni 09. no lesiones en pancreas - peptido c: 4.1 normal - glicemia mas baja: 25mg/dl - insulina 26.7 uu/ml - eco endoscopia biliopancreatica: hacia el proceso uncinado, en estrecha relacion con la vena mesenterica superior se encuentra una lesion hipoecoica homogenea, de bordes bien definidos, ovoide que mide 12mm de diametro mayor. Paciente que no ha vuelto a presentar hipoglucemia sintomatica, actualmente asintomtico. Asiste con reporte de laboratorios: 7/12/2010: glucosa en ayuno de 90 mg/dl Examen fisico ver vietas CONCEPTO PAciente con Insulinoma pancreatico sin evidencia de metastasis resecado, hipoglucemia resuelta quien ha perdido 13 Kg, en quien como parte del estudio se debe descartar la posibilidad de MEN, por o cual se solicitara niveles de prolactina y calcio serico. Control en 3 meses con glucosa.

Step 1 - Preprocessing. This phase executes five processes *Tokenization, Stop-words, Spelling and Filtering*, and its result is a shorter text, corrected and without duplication of information.

Step 2 - Identifying Categories. In this phase, the user profile was initially verified, specifically the categories that the user selected to be considered in the summary. For example, the selected categories were: *Diseases, Drugs and Exams*, the *Keywords "Diagnóstico", Age Category* and *Gender Category*. The date range selected was '2011-01-25'. We assume that the user profile selected at most (6) phrases, from which the user wants to see: (1) phrase by category selected. As a result of this step, the phrases belonging to these selected categories by the user will be extracted.

Step 3 - Extraction of Candidate Phrases. This phase is characterized by filtering the sentences annotated in the previous step, following three steps: Calculation by Categories, Calculation by Frequency and Filtering by Categories, Frequency and Order. The end result of the filtering process is presented in Figure 2. Each color highlighting the words in this figure represents a word that belongs to a category; for example, the grey color represents a word annotated within the *Disease Category*, the fuchsia color - *Age Category*, the blue color - *Drug Category*, the pink color - the *Keywords Category* and the aquamarine color - the *Exam Category*. We can see that the generated summary achieved to extract the existing information out of the categories selected by the user, with the adequate coherence, in a shorter text, in the order as they exist chronologically and without duplication of information. We also compare the results obtained with the *System Dragon ToolKit*, using the same clinical records. In terms of metrics, the results regarding amount of information (in terms of words that belong to the categories selected by the user), comparing both systems (Dragon ToolKit and GReAT System), indicates that GReAT gets more relevant information to the user, because it gets more categories than the Dragon Toolkit System. See table 2. Several tests were conducted with 100 records, which results are shown in Table 2. This table shows the metrics used for the evaluation of both systems, for example, the metrics regarding to the quality were: **Duplicity**indicates if a system duplicates data or not, **Noise** indicates whether a system corrects or not the spelling of the text within the records and **Sequentiality** indi-

Table 2: Comparison of Results.

Date Range	# HCE	Systems GReAT	Systems DragonTool
Duplicity			
01/09/2011-30/09/2011	43	No	Yes
01/12/2011-30/12/2011	18	No	Yes
01/07/2011-30/07/2011	15	No	Yes
01/10/2011-30/10/2011	13	No	Yes
01/08/2011-30/08/2011	11	No	Yes
Noise			
01/09/2011-30/09/2011	43	No	Yes
01/12/2011-30/12/2011	18	No	Yes
01/07/2011-30/07/2011	15	No	Yes
01/10/2011-30/10/2011	13	No	Yes
01/08/2011-30/08/2011	11	No	Yes
Sequentiality			
01/09/2011-30/09/2011	43	Yes	No
01/12/2011-30/12/2011	18	Yes	No
01/07/2011-30/07/2011	15	Yes	No
01/10/2011-30/10/2011	13	Yes	No
01/08/2011-30/08/2011	11	Yes	No
User			
01/09/2011-30/09/2011	43	58	14
01/12/2011-30/12/2011	18	22	4
01/07/2011-30/07/2011	15	12	2
01/10/2011-30/10/2011	13	40	16
01/08/2011-30/08/2011	11	71	17
Performance			
01/09/2011-30/09/2011	43	48384	2550
01/12/2011-30/12/2011	18	43353	1306
01/07/2011-30/07/2011	15	34192	1316
01/10/2011-30/10/2011	13	20244	2393
01/08/2011-30/08/2011	11	16395	2178

cates if a system maintains or not text sequentiality found in the medical record when it summarizes the texts. On the other hand, the metric called **User** is a measure in terms of #Total words by category found in the final summary for each system, and finally, the metric **Performance** measures the total milliseconds to finish the summary process for each system. The results show that the GReAT improves the quality of summaries, eliminating noise and duplicity of information within the text, preserving the sequentiality of the original text and providing more information according to the user's needs. This table shows that GReAT System extracts more relevant information to the user than the DragonTool System. However, in terms of performance, the Dragon Toolkit System has better results than the System GReAT, this will be an important challenge for the future work.

6 CONCLUSIONS

From the review of the literature about the existing solutions for automated generation of text summaries, GReAT seeks to address some weaknesses and forward challenges. Although it is in a process of evaluation and validation, initial results show an improvement in quality and consistency of summaries obtained, taking into account the needs of users, and topics that are often important in the domain, the sequentiality of the original text and the noise that natural language presents. As future work we will focus on improving the performance and reducing the computational cost of the analysis and data mining on text documents, being able to solve the existing problems to compute the similarity between documents when not using the same vocabulary, covering shortcoming of missing words such as" Internet" and" World Wide web" as the same concepts and finally achieving a suitable length to present summaries of texts depending on the size of the input document.

ACKNOWLEDGEMENTS

This work was supported by the project *"Extraccion semi-automatica de metadatos de fuentes de datos estructuradas: Una aproximacion basada en agentes y mineria de datos"* made by Banco Santander S.A. and Pontificia Universidad Javeriana.

REFERENCES

Abu-Jbara, A. and Radev, D. (2011). Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 500–509, Stroudsburg, PA, USA. Association for Computational Linguistics.

Arora, R. and Ravindran, B. (2008). Latent dirichlet allocation based multi-document summarization. In *Proceedings of the second workshop on Analytics for noisy unstructured text data*, AND '08, pages 91–97, New York, NY, USA. ACM.

Bossard, A., Généreux, M., and Poibeau, T. (2009). Cbseas, a summarization system integration of opinion mining techniques to summarize blogs. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, EACL '09, pages 5–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brun, Ricardo Eto, S. J. A. (2004). Minería textual. *El profesional de la informacin*, 13(1).

Chang, T.-M. and Hsiao, W.-F. (2008). A hybrid approach to automatic text summarization. In *Computer and Information Technology, 2008. CIT 2008. 8th IEEE International Conference on*, pages 65–70.

Dalal, M. K. and Zaveri, M. A. (2011). Heuristics based automatic text summarization of unstructured text. In *Proceedings of the International Conference &*

Workshop on Emerging Trends in Technology, ICWET '11, pages 690–693, New York, NY, USA. ACM.

Daumé, III, H. and Marcu, D. (2002). A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 449–456, Stroudsburg, PA, USA. Association for Computational Linguistics.

Devasena, C. (2012). Automatic text categorization and summarization using rule reduction. In *Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on*, pages 594–598.

Genest, P.-E. and Lapalme, G. (2011). Framework for abstractive summarization using text-to-text generation. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 64–73, Portland, Oregon. Association for Computational Linguistics.

Guelpeli, M. V. C., Garcia, A., and Branco, A. (2011). The process of summarization in the pre-processing stage in order to improve measurement of texts when clustering. In *Internet Technology and Secured Transactions (ICITST), 2011 International Conference for*, pages 388–395.

Gunen, Erkan, D. R. R. (2004). Lexrank: Graph.based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research 22 (2004) 457-479*, 22.

Hotho, A., Nrnberger, A., and Paa, G. (2005). A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*.

Inniss, T. R., Lee, J. R., Light, M., Grassi, M. A., Thomas, G., and Williams, A. B. (2006). Towards applying text mining and natural language processing for biomedical ontology acquisition. In *Proceedings of the 1st international workshop on Text mining in bioinformatics*, TMBIO '06, pages 7–14, New York, NY, USA. ACM.

Kianmehr, K., Gao, S., Attari, J., Rahman, M. M., Akomeah, K., Alhajj, R., Rokne, J., and Barker, K. (2009). Text summarization techniques: Svm versus neural networks. In *Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services*, iiWAS '09, pages 487–491, New York, NY, USA. ACM.

Ling, X., Mei, Q., Zhai, C., and Schatz, B. (2008). Mining multifaceted overviews of arbitrary topics in a text collection. In *In Proc. SIGKDD08*, pages 497–505. ACM.

Liu, H.-H., Huang, Y.-T., and Chiang, J.-H. (2010). A study on paragraph ranking and recommendation by topic information retrieval from biomedical literature. In *Computer Symposium (ICS), 2010 International*, pages 859–864.

Long, C., Huang, M.-L., Zhu, X.-Y., and Li, M. (2010). A new approach for multi-document update summarization. *J. Comput. Sci. Technol.*, 25(4):739–749.

Mehdad, Y., Negri, M., Cabrio, E., Kouylekov, M., and Magnini, B. EDITS: An Open Source Framework for Recognizing Textual Entailment.

Mei, Q., Guo, J., and Radev, D. (2010). Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1009–1018, New York, NY, USA. ACM.

Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D., and Zajic, D. (2009). Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 584–592, Stroudsburg, PA, USA. Association for Computational Linguistics.

Muthukrishnan, P., Radev, D., and Mei, Q. (2011). Simultaneous similarity learning and feature-weight learning for document clustering. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, TextGraphs-6, pages 42–50, Stroudsburg, PA, USA. Association for Computational Linguistics.

Park, J., Fukuhara, T., Ohmukai, I., Takeda, H., and Lee, S.-g. (2008). Web content summarization using social bookmarks: a new approach for social summarization. In *Proceedings of the 10th ACM workshop on Web information and data management*, WIDM '08, pages 103–110, New York, NY, USA. ACM.

Reeve, L. H., Han, H., Nagori, S. V., Yang, J. C., Schwimmer, T. A., and Brooks, A. D. (2006). Concept frequency distribution in biomedical text summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 604–611, New York, NY, USA. ACM.

Saravanan, M., Raman, S., and Ravindran, B. (2005). A probabilistic approach to multi-document summarization for generating a tiled summary. In *Computational Intelligence and Multimedia Applications, 2005. Sixth International Conference on*, pages 167–172.

Tran, N.-P., Lee, M., Hong, S., and Shin, M. (2012). Memory efficient parallelization for aho-corasick algorithm on a gpu. In *High Performance Computing and Communication 2012 IEEE 9th International Conference on Embedded Software and Systems (HPCC-ICESS), 2012 IEEE 14th International Conference on*, pages 432–438.

Wang, W., Xiao, C., Lin, X., and Zhang, C. (2009). Efficient approximate entity extraction with edit distance constraints. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, SIGMOD '09, pages 759–770, New York, NY, USA. ACM.

Yu, L. and Ren, F. (2009). A study on cross-language text summarization using supervised methods. In *Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on*, pages 1–7.

Zhan, J., Loh, H. T., and Liu, Y. (2009). Gather customer concerns from online product reviews - a text summarization approach. *Expert Syst. Appl.*, 36(2):2107–2115.

Zhang, Pei-ying, L. C.-h. (2009). Automatic text summarization based on sentences clustering and extraction. *IEEE*.