On Automatic Text Summarization: Some Ideas and an Illustration

Everardo Garcia-Menier

In this work, text summarization task is discussed. It is mentioned that it would be advisable to summarize a text before any other treatment and to work with the summary instead of the original document, since the former is smaller and therefore, easier to handle. This would be valid under the assumption that the summary contains the same important information that the original text. The use of Domain-dependent sublanguages is proposed as a very useful tool for obtaining quality summaries. Context factors and summary purpose are mentioned as important parts of the text summarization process. A real-world example, belonging to the medical domain, is given to illustrate the ideas developed throughout the paper.

1 INTRODUCTION

Due to the great amount of information that can be electronically stored nowadays, it has become necessary to develop computational tools to manage this information. Many of these tools have to deal with free format text, that is, text written in Natural Language. We can group these tools as Automatic Text Processing Tools. As examples, we can find systems for Information Retrieval, Information Extraction, Text Classification, Text Understanding, Text Summarization, etc.

Although some of the areas mentioned above (as Information Retrieval for instance) use tools other than merely text-based ones, we should agree on the importance of the role that text management plays in all of them.

We think that Text Summarization is a very important area of Automatic Text Processing because a summary is a reduced version of a text that preserves its more important parts. Therefore, on one hand we have extracted the main ideas from a text and; on the other hand we have now a smaller text. That's why we consider that it would be advisable first to summarize the text and then to use the summary as an input for the Automatic Text Processing task we have in mind. This emphasizes the importance of text summarization task. In this work, we discuss some ideas about it.

The main purpose of this paper is to propose the use of Domain-dependent sub-languages as very important tools for text summarization tasks. Section 2 is devoted to giving some background knowledge about text summarization, some related works are mentioned and some definitions are given.

In section 3, we analyze how people make summaries and context factors are introduced as important parts of summary production process. Some considerations about Domain-dependent sublanguages and how summary purpose can be included in the sublanguage to improve this task are the subjects of section 4.The problem of automatic text summarization is discussed in section 5. In section 6 some considerations about the Domain are given. In section 7, we give a practical example of how a text can be automatically summarized using the ideas developed so far. Future work is mentioned in section 8. Finally, in section 9 some conclusions are given.

2 BACKGROUND

We will use the definition of summarization given in [Sparck-Jones 98]. A summary is defined here as "a reductive transformation of source text to summary text through content reduction by selection and/or generalization on what is important in the source". According to this definition, we need to extract from the source text the important parts or ideas. The mentioned work states that practically all summarizer systems proposed so far fall under two headings: "text extraction" that is, to choose some parts from the source text based on linguistic criteria of importance. On the other hand is "fact extraction", in this case the search of parts to be extracted is driven by certain knowledge about the important information we want to find. In both cases we need to face the following problem: How can we decide whether certain part of the text is important, so we have to chose it; or irrelevant, so we have to discard it.

It could seem that the task of marking up important parts from a text, for a later extraction, depends on the text. However there are a number of domain-independent text summarizers for instance Microsoft's Word 97 AutoSummarize Feature[1], Verity's Search 97: The Summarize Feature, NCR Extractor[2] or the one described in [Barker et. al. 98]. All of them use general linguistic knowledge to make summaries.

We strongly believe, as it is stated in [Sparck-Jones 98], that a very important feature that a summarizer system must have is "to recognize the role of summary purpose in determining the nature of the content condensation". That is, the way

[1] Available from the "tools" menu in Word 97.
[2] A brief description and an evaluation of these systems can be found in [Turney 97]

in which source text will be condensed must be driven mainly by the requirements the summary must fill in.

To achieve this, it is necessary to recognize which phrases of the source text are related with the summary purpose, and for this, we need to understand the text. This leads us into fields such as Text Interpretation. This is a difficult task, actually. A system that can process unrestricted natural language is seen as unreachable at least in the middle term, [Sparck-Jones 98; Mallery 94]. For this reason, one approach in text interpretation is to use domain-specific knowledge.

As examples of this approach, we can mention PDS, a system that interprets the letters sent by a Hospital to the patient's doctor [Mikheev 96]. A system that makes diagnostics about car faults [Ciravegna] and a system that classifies newspaper notes in two groups namely the “good” ones and the “bad” ones for a University's prestige [Garcia 98]. All of these systems use a language that depends on the topic of the text, i.e. a Domain-dependent language, in order to extract some knowledge from texts.

The constraints on language and its dependence on domain are important factors in the emergence of what is called a sublanguage. As it is stated in [Lehrberger 86] “A sublanguage is a language resulting from restriction on and deviation from the standard grammar of a natural language, albeit in special circumstances”.

In the system for newspaper notes, described in [Garcia 98] a main assumption is stated: “The lexicon necessary to cause a good or bad opinion in a public, about a topic is a restricted subset of the language lexicon”. Although it is not mentioned in an explicit way, it follows from text that the grammar used to cause such an opinion is also restricted. So, in that system a sublanguage is used. The same is true for the other systems mentioned above.

In the following, we will give some ideas on how domain-dependent sublanguages can help in the text summarization task.

3 HOW DO WE HUMANS SUMMARIZE TEXT?

As we stated before, text summarization is a difficult task, even for human beings. Summary quality depends on a set of heterogeneous factors that are inherent to the person who makes the summary. For instance education level, the extent of knowledge s/he has about the subject of the text, how much abstraction capacity s/he has, if s/he is a trained or a novel person, etc.

However, people can be trained to summarize text and this increases their performance as shows an experiment held in Spain. In this experiment, two groups of Spanish-speaking people, previously divided according their proficiency in the English language, were asked to make a summary from an English text. The group with less proficiency in English received a special training program to learn how to summarize text. The other group did not receive any previous training.

At the end of the experiment, the qualities of the summaries obtained by both groups were virtually the same. The complete experiment is described in [Juan and Palmer]. The training program mentioned in this work, contains a set of tips for to the students, among these tips we can find the following:

> 3.- "We should start by finding the main topic of the summary",
> 4.- "Read the text thoroughly once in order to see what is the main topic. Read it again starting to underline all the important information"[3]

The tasks of finding the main topic of a text, and underlining the important information, are supposed to be very easy for a person. However, for an automatic system they represent a real challenge. A way to cope with the former could be the system CLASITEX that finds the main themes in a document [Guzman 98]. The second task mentioned above, underlining (recognizing) the important information is a main problem in text summarization.

To face this problem it is necessary to take into account the context for making a quality summary, especially purpose factors. One should use these factors to answer questions like: What context will the summary be used in? What is the kind of readers for whom this summary is intended? Or what is the summary for? For a more detailed description of context factors, see [Sparck-Jones 98].

So, if we want a system that automatically makes summaries from texts, with a quality similar to the one achieved by a person. We need to provide such a system with some "tools" that take into account context factors, especially purpose factors. Some ideas to accomplish this are given below.

4 DOMAIN-DEPENDENT SUBLANGUAGES

As stated above a sublanguage can be viewed as a subset of natural language, i.e. a sub-lexicon together with a subset of grammar rules defined mainly by the Domain. We will clarify this idea below.

[3] The numbers here are those used in this work.

Language is used when a person wants to send a certain message to others. This message belongs to a specific domain. Moreover, the person has a purpose when s/he sends the message. We think that the lexicon and the grammar rules that this person needs to make other people understand his/her message are strongly related to the domain the message belongs to. They are also related to the purpose this person has in mind when s/he sends the message. Both of these entities: **Domain** and **purpose** define the sublanguage to be used.

For example, let us suppose a student wants to cause a bad opinion on people about her teacher. So she says: "This teacher visits many bars every night, and he has been involved in sexual harassment cases more than once. He has also been arrested for driving when he is drunk and is suspected for rape. It is clear that this student wants to destroy her teacher's reputation, but we can see this message from different points of view:

If the domain were the teacher's sexual behavior and the purpose of the student were to exhibit her teacher as a person with an immoral conduct. Then the sublanguage should contain words (or phrases) like "sexual harassment" or " rape".

On the other hand, if the domain were the teacher's bad habits and the purpose of the student were to exhibit him as an alcoholic, then the sublanguage should contain words like "bars" or "drunk"[4]. For a more detailed description of the idea that a restricted language is enough to communicate certain facts that will cause an opinion, see [Garcia 98].

The lexicon and grammar rules that are defined by the message's domain and by the speaker's purpose are what we call a Domain-Dependent sublanguage. We think that this kind of sublanguages will be very useful for the text summarization task.

5 HOW CAN A TEXT BE AUTOMATICALLY SUMMARIZATED?

Let's suppose that we have a text, and that we know which its main topic is (tip 3 mentioned above). If we now underline the important information (i.e. tip 4), then we can extract these underlined parts in order to generate a summary.

The strategy proposed here to "underline important information" is closely related with the use of a suitable Domain-Dependent sublanguage. As a first step, we will erase from the text all those words not present in the sub-lexicon.

[4] Words like "suspect" or "arrested" belong to both sublanguages, that is, sublanguages not need to be disjoint.

Consequently, we will have a skimmed text that contains only words that are relevant to the Domain. As a second step, we will mark up all those sentences that are important to the Domain. This can be achieved by using the sub-grammar defined by the sublanguage. In this way, a summary that satisfies the purpose requirements will be obtained.

Domain-Dependent dictionaries have been successfully used in a number of systems that cope with text interpretation. It has even been said that an expert can build such a dictionary in a reasonable time and systems have been developed for automatically constructing such dictionaries. For example, see [Riloff 93; Chen et al.94].

What it is not that obvious in the use of a sub-grammar is that, from a point of view, it includes expressions and structures belonging specifically to the domain as well as general expressions and structures common to any domain. Moreover, this sub-grammar must contain, inside its structure, the purpose factors needed to make a quality summary. We will clarify this idea later.

6 SOME CONSIDERATIONS ABOUT THE DOMAIN

Since the Domain determines both the sub-lexicon and the sub-grammar that will be used in the construction of a text summarizer system, we need to make some considerations about it before we go on.

It seems obvious that the broader the Domain is, the bigger the sub-lexicon and the more complicated the sub-grammar will be. Therefore, we prefer to choose the Domain as small as we can without loosing generality. For instance, a very general Domain could be the medical one. This is too general. On the other hand, we can think of a sub-domain of it like, for example, the diseases of the last section of the left coronary artery, which is a too specific one. We can mediate this situation by choosing, for example, the domain of heart diseases due to faults in coronary arteries[5].

Following with the previous example, we can think of the diseases of the left coronary artery last section domain, as a sub-domain of the previously chosen one. Moreover, the medical domain would be a super-domain. This gives us the idea that a whole hierarchy of domains could be constructed. This idea will show up later.

Once the Domain has been properly chosen, and in order to define its sublanguage, we need to know the summary purpose. That is, we need to answer

[5] This is the Domain used in Mikheev's PDS.

the question: What is the summary for? We can think of the purpose as a part of the Domain, or to take it as a separated concept. The important thing is that the sublanguage necessary to summarize the text must include both of them.

We will illustrate these ideas with an example. We have chosen as domain a subfield of Medicine, namely, medical records. As stated above, we need to circumscribe our analysis to a restricted kind of medical record. We chose Otorhinolaringology as the topic for the records. Therefore, our Domain consists of medical Otorhinolaringologic records.

7 AN ILLUSTRATION

Here we give an example of how a Domain-Dependent sublanguage may be used to recognize important information and, therefore, to make a summary. As stated above we will use the medical Otorhinolaringologic records Domain.

In order to just illustrate the main ideas of this work, we will oversimplify the example. A more detailed description of how to construct such a sublexicon is under preparation now.

This work is being developed to treat medical records written in Spanish language; so, the example we will analyze here, although a real world one, had to be adapted to the English language. We expect that the translation has kept the main features of the original note.

The medical record is the following:

> November 5th, 1999 9:54.
>
> Male patient, 74 years old, musician, married, ed: elementary school. He has been a smoker during the last 50 years, about a half package a day, doesn't drink. He's diabetic since 4 years ago and had a Qx due to appendicitis ten yrs. ago. IMSS[6]. The pat. presents a sensation of obstructed le. ear since he had water penetration during a bath; he attempted to manipulate with a cotton swab and the sensation got worse. Yesterday he presented the same sensation in right ear. Important hypoacousy[7] is noted. Dx. R EAC

[6] This means that the patient is a user of IMSS, acronym that stands for "Instituto Mexicano del Seguro Social" Mexican Institute for Social Security, one of the biggest Public Health Institutions in Mexico.

[7] A reduction of the ability to hear

```
earwax impacted at the bottom of the canal. The same
for L EAC.
Plan: It is required to soften earwax before a
bilateral washing.
Tx. Amalyt Sol. 3 warm drops into e/ear 3 times a day
f. 3 days
Control Monday Nov. 11th 99 10:00 hrs.
```

The words that a Domain-Dependent sub-lexicon should contain can be grouped together into three classes:

Words belonging exclusively to the Domain (Otorhinolaringology), words belonging to its super-Domain (Medicine) and words which can be called "general purpose" words, i.e. words that are not specific of the field but appear in most of the documents[8].

Before the analysis it should be noted that since medical records are written while the patient is talking, one of the following features could be present in the record:

- It is common that a doctor makes mistakes when writing his record, especially when s/he writes it while s/he is making a medical examination.
- The use of abbreviations is a common practice. These abbreviations can be generally accepted ones, like Mr. for Mister. There is another kind of abbreviations that have become commonly accepted, for example, ASAP for As Soon As Possible[9]. A third kind of abbreviations contains some Domain-Dependent ones for instance: Dx. for Diagnostic or Tx. for treatment. We may mention another kind of "personal" abbreviations, or even symbols, that depend on the person who writes the record[10].

In order to solve the first of these problems, we could preprocess the record using a system that corrects the spelling by using certain knowledge of the domain. A system like this is described in [Taghva et al. 95]. To solve the second problem we can include all these abbreviated words in the sublexicon so the system can recognize them.

---

[8] These words are generally used to describe a situation or as links.

[9] We could include here those "abreviations" that have emerged from the use of e-mail for instance :) for expressing happiness.

[10] For instance, a Doctor takes the female sign: ♀ and, writes a number inside. By this he meant a woman and her age, for example: "female patient, 25 years old" would be represented by: ㉕♀

To illustrate how this kind of sub-lexicon would look like, we give an example below. Note that lexicon consists not only of words but also phrases.

- **Domain words:**

Amalyt Sol., bilateral wash, bottom of the canal, earwax impacted, drops, e/ear (for each ear), hypoacousy, le. ear (for left ear), obstructed, R EAC, L EAC (for Right and Left External Auditory Canal respectively), right ear, soften.

- **Super-Domain words:**

appendicitis, control, diabetic, drinks, Dx, got worse, Male, patient, plan, Qx, sensation, smoker, times a day[11], Tx[12].

- **General purpose words:**

and, expressions to denote age, dates, doesn't, due, during, for, had, has been, hrs., important, is noted, is required, numbers, presented, presents, previous, pronouns, same, since, time, warm, time-interval ago[13], yesterday.

- **"Personal words":**

f. (for for), pat. (for patient), yrs (for years).

We present now how we can use this sub-lexicon to make a summary of the medical record above.

The first step is to eliminate all those words not belonging to the sublexicon to obtain a first version of the summarized medical record. Such first version is shown below.

```
November 5th, 1999  9:54.

Male patient, 74 years old, He has been smoker during
last 50 years doesn't drink. He's diabetic since 4
years ago and had Qx due appendicitis ten yrs. ago.
pat. presents sensation obstructed le. ear since; and
sensation got worse. Yesterday he presented same
sensation right ear. Important hypoacousy is noted.
Dx. R EAC earwax impacted bottom of the canal. same
for L EAC.
Plan: It is required soften earwax previous bilateral
wash.
```

[11] This expression is used for indicating a dose, for example "one pill 3 times a day".
[12] Dx, Tx and Qx stand for Diagnostic, Treatment and Surgery respectively.
[13] For instance: two years ago or four months ago.

```
Tx. Amalyt Sol. 3 warm drops e/ear 3 times a day f. 3
days
Control Monday Nov. 10th 99 10:00 hrs.
```

As a second step, those structures considered in the subgrammar will be kept in the record possibly transformed via specific rules. The structures that do not appear in the subgrammar will be deleted.

As an example of one structure contained in the subgrammar, we can mention the following:

```
{patient/he/she}(has been) {smoker/alcoholic }
{during/for} (time-int}) --> {smoker/alcoholic } for
time-int
```

This means: One word belonging to the set {patient, he, she}, followed by the phrase "has been", followed by one word in the set {smoker, alcoholic}, followed by one word in the set {during, for}. Finally a phrase denoting a time interval. Wherever this structure is found, then it must be transformed into:

{smoker/alcoholic} for time-interval.

The resulting summary is shown below[14]:

```
11/05/99  9:54.

Male. age: 74, smoker during last 50 years. doesn't drink.
diabetic 4 years ago. Qx appendicitis ten years ago.
sensation obstructed left ear; sensation got worse.
Yesterday same sensation right ear. Important hypoacousy.
Dx. R EAC earwax impacted bottom of the canal. same for L
EAC.
Plan: soften earwax previous bilateral wash.
Tx. Amalyt Sol. 3 warm drops each ear 3 times a day for 3
days
Control 11/10/99 10:00.
```

The summarized record has 71 words, while the original one has 138 words. Moreover, the relevant information about the domain contained in both records is almost the same.

[14] Here some other rules were used. For instance, to transform the dates to a standard format: mm/yy/dd.

The summary found above can be considered as a general-purpose one because it has no specific purpose. If it had one, some modifications in the sublexicon and the subgrammar should be done.

For example, let's suppose that the summary purpose is to get some information about the patient's symptoms and the diagnostic. Then those words related with sex, age, patient's previous history and habits should be deleted from sublexicon. Some modifications on subgrammar should also be done, such as eliminating the Plan, Tx and Control sections for instance. The resulting summary would look like this:

```
sensation obstructed left ear; sensation got worse.
Yesterday same sensation right ear. Important hypoacousy.
Dx. R EAC earwax impacted bottom of the canal. same for L
EAC.
```

There are only 27 words in this summary.

It is important to note that both of the summaries above, the general and the specific purpose ones, were generated just by using a sublexicon and a subgrammar that depend on the Domain. That is, just by using a Domain-dependent sublanguage.

8 FUTURE WORK

Although the example discussed above is a real world one, the sublanguage that will be used for Othorrinolaringologic records summarization is still under construction. The first thing to do is to finish it.

Next, we need to build a summarizer system, based on the ideas described above, so we can test its performance. It is intended to contrast summaries made by the system against human-made ones as it is described in [Sparck-Jones 98].

The sublexicon and the subgrammar are being constructed manually, we think that this task could be done automatically. That is, to extract the elements of the sublanguage directly from texts. There is already some work in extracting dictionaries or patterns from text, see for example [Chen et al.94; Riloff and Shoen 95; Riloff 96].

As it was stated above, we can think of a hierarchy of domains. For example, the Domain of diseases of coronaries is a subdomain of the artery/venous diseases Domain, which is a subdomain of all diseases domain. The whole domain of all diseases can be seen as a subdomain of a Domain that could be called a fault-

detection domain[15]. This big Domain will have other subdomains such as the car-fault-detection domain, and all those domains that deal with detecting some fault in a machine, communications system, etc..

We think that this idea of a hierarchy of domains could be very important. Besides the usage just described, we believe that it can be used in text classification as follows:

Lets suppose that we have a set of sublanguages, one for each domain in the hierarchy. If it is necessary to classify one text, then this text could be processed with every sublanguage. If the sublanguage structure recognizes the text, then it belongs to the sublanguage's domain. Else, it does not. So, Domain-dependent sublanguages could also be used for text classification tasks. This is another idea we can explore in future work.

9 CONCLUSIONS

Although a text summarizer system based on the ideas proposed in this paper is still under construction, we have made some tests and we have obtained quite good results. That's why we believe that such a system, when finished, will function well.

As can be seen, no syntactic analysis is necessary to generate the summary. The only thing that we need is to make a matching between expressions on the text and those in the sublanguage and to ignore anything that does not belongs to it. We believe that this fact should reduce the time necessary to produce a quality summary[16].

As stated above, the summarized text should contain the same relevant information than the original text. This fact will make easier text processing tasks like Information Retrieval, Information Extraction and Text Classification for instance, since we will have a text equivalent to, but smaller than the original one.

We strongly believe that the ideas developed in this paper will produce quality results. As mentioned above it is necessary to implement a system based on these ideas so we can make tests

[15] We are assuming that the diagnostic of a disease is like to detect a fault in one or more parts of the human body.

[16] This time saving is because there is no need of a deep syntactic analysis but just a pattern matching process between the text and the sublanguage.

REFERENCES

Barker, Ken, Yllias Chali, Terry Copeck, Stan Matwin, and Stan Szoakowicz 1998. *The design of a configurable Text Summarization System*. Technical Report. School of Information Technology and Engineering. University of Ottawa. 1998

Chen, Hsinchun, Bruce Schatz, Joanne Martinez and Tobun Dorbin Ng. 1994. *Generating a Domain-specific Thesaurus Automatically: An Experiment on FlyBase.* 1994

Ciravegna, Fabio. *Understanding Messages in a Diagnostic Domain.* Instituto per la Riecerca Scientifica e Tecnologica. Povo-Trento. Italy

Doyle, Jon, Isaac Kohane, William Long and Peter Szolovits. 1996. *High-Performance Knowledge Base Support for Monitoring, Analysis and Interpretation Tasks.* Massachusets Laboratory for Computer Science Institute of Technology and The Childrens' Hospital (Boston) 1996.

Garcia Menier, Everardo. 1998. *Un Sistema para la Clasificación de Notas Periodísticas.* Memorias del Simposium Internacional de Computacion CIC'98. Mexico 1998

Guzman Arenas, Adolfo. 1998. *Finding the main Themes in a Spanish document.* Journal Expert Systems with Applications. 1998

Juan, Esther Uso and Juan Carlos Palmer Silveira. *A Product-Focused Approach to text Summarization.* Universitat Jaime I-Castello . Spain http://www.aitech.ac.jp/~iteslj/Articles/Juan-TextSummary.html

Lehrberger. 1986. *Sublanguage Analysis*. In R Grishman and Kittredge Editors. *Analyzing Language in Restricted Domains: Sublanguage descriptions and Processing*. Hillsdale NJ. Lawrence Erlbaum Associates Publishers. 1986.

Mallery, John C. 1994 *Beyond Correlation: Bringing Artificial Intelligence to Events Data,* A.I. Laboratory MIT. 1994

Mauldin, Michael L. 1991. *Retrieval Performance in FERRET: A Conceptual Information Retrieval System.* 14th International Conference on Research and Development in Information Retrieval. October 1991.

Mikheev, Andrei. 1996. *Domain Knowledge for Natural Language Processing.* HCRC Language Technology Group. University of Edinburgh, January 17,1996

Riloff, Ellen. 1993. *Automatically Constructing a Dictionary for Information Extraction Tasks.* Proceedings of the eleventh National Conference on Artificial Intelligence 1993 AAAI Press/MIT Press

Riloff, Ellen and Jay Shoen. 1995. *Automatically Acquiring Conceptual Patterns Without an Annotated Corpus.* In Proceedings of the Third Workshop on Very Large Corpora. 1995

Riloff, Ellen. 1996. *Automatically Generating Extraction Patterns from Untagged Text.* Proceedings of the Thirteenth National Conference on Artificial Intelligence. 1996

Sparck Jones, Karen. 1998. *Automatic summarizing: factors and directions.* Advances in automatic text summarization. I. Mani and M. Maybiry Eds. MIT Press 1998.

Taghva, Kazem, Julie Borsack and Allen Condit. 1995. *An expert system for automatically correcting OCR output.* Information Science Research Institute. University of Nevada, Las Vegas. 1995.

Turney, P. 1997. *Extraction of Keyphrases from Text: Evaluation of Four Algorithms.* National Research Council Canada. Institute for Information Technology. 1997

Everardo Garcia-Menier is a researcher at Artificial Intelligence Department from the School of Physics and Artificial Intelligence. University of Veracruz, Mexico. He can be reached at egarcia @mia.uv.mx Phone (52 28) 17-29-57, Fax (52 28) 17-28-55. See http://www.mia.uv.mx/ ~egarcia