MARÍA CRISTINA TOLEDO BÁEZ
Department of Translation and Interpreting
University of Malaga
Malaga, Spain
toledo@uma.es

Automatic extraction and documentation process in translation

1. Introduction

As Lavid (2005) points out, information has become one of the basic elements in our current society, which may be called the "Third Wave", paraphrasing Alvin Toffler's book (1996). First wave is the society after agrarian revolution; Second wave is industrial. Third Wave represents information and knowledge revolution. New millennium's society is "information society", where Information and Communication(s) Technology (ICT) has a paramount importance. Therefore; the exchange of languages and cultures plays an important role in this information society. Consequently, translators and interpreters may become fundamental mediators at global levels.

In this context, Internet appears as an essential tool since it offers new ways of communication and scientific knowledge spreading. In addition, it facilitates and improves the documentation process. The translator, as an information user and an information producer, considers Internet to be a valuable documentation source and a useful communication system.

According to Pinto Molina (2002: 2), the "informational revolution" makes possible to compile more information in less time and, consequently, to improve translator's efficiency. With the mushrooming of the quantity of on-line text information, triggered in part by the growth of the World Wide Web, it is especially useful to have tools which can help users digest information content.

Nevertheless, translators have to be really skillful during documentation process since they need to be able to distinguish and choose only reliable information resources. This is because Internet, although it is a valuable and really useful tool, contains much unreliable information.

In that regard, an abstract may be quite useful for translators since it helps to choose the correct information in the documentation process. Given that translators normally have to meet tight deadlines, abstracting articles or electronic resources is an advantageous solution and heavily facilitates the translation process.

In this paper[1], we approximate to automatic text extraction in two languages (English and Spanish) by means of two elements:

1. Terms and conditions of package tours –specifically cruises- extracted from multilingual macrocorpus *Turicor*, which belongs to the project *TURICOR: A multilingual corpus of tourism contracts (German, Spanish, English, Italian) for automatic text generation and legal translation* [*Turicor: compilación de un corpus de contratos turísticos (alemán, español, inglés, italiano) para la generación textual multilingüe y la traducción jurídica*] (Ref. no. BBF2003-04616 (2003-2006, Spanish Ministry of Science and Technology).
2. Six multilingual and free automatic extraction systems: *Copernic Summarizer, Extractor, GistSumm,* Microsoft Word Summarization, *SweSum*, and *WebSumm*.

Once the texts are automatically extracted, advantages and disadvantages of automatic extraction were analyzed in order to check their impact and contribution to translators' documentation process.

2. AG: between NLG and AG

Initially, Natural Language Generation (NLG) and Abstracts Generation (AG), both belonging to the field of Natural Language Processing (NLP), could be considered to have the same characteristics since both are related to automatic language generation. In fact, some researchers think that AG is part of NLG (Lavid, 2005: 216). However, other researchers have a different point of view. Regarding NLG, Reiter and Dale (1997:57) define it as:

> Natural language generation is the subfield of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce understandable texts in […] human languages from some underlying non-linguistic representation of information.

Concerning AG, Acero et al. (2001: 282) give the following definition:

> Por generación automática de resúmenes de texto entendemos el proceso por el cual se identifica la información sustancial proveniente de una fuente (o varias) para producir una versión abreviada destinada a un usuario particular (o grupo de usuarios) y una tarea (o tareas).

Therefore, NLG and AG share the same aims, i.e., text generation, but they use different information resources because NLG is based on non-linguistic representation of information and AG is based on a linguistic representation. However, we consider AG to be closely related to NLG and, consequently, they both belong to the Translation Technologies (TT) field. AG is at its very peak due to the huge amount of information on Internet and the necessity of filtering the electronic resources and information.

---

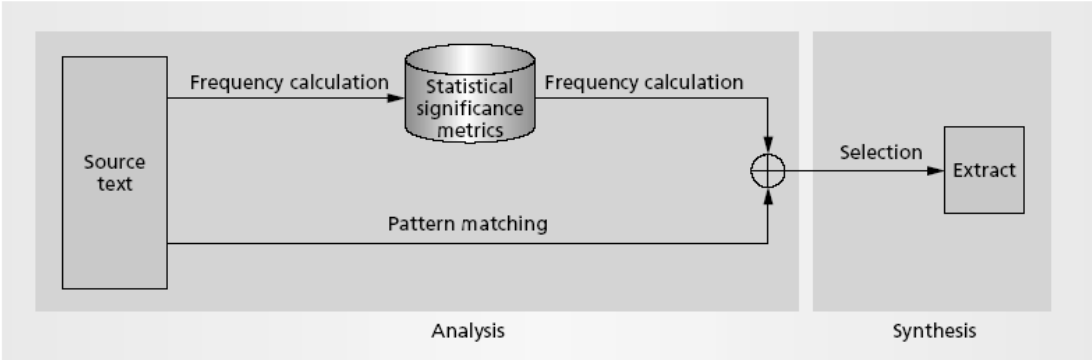
[1] The research reported in this paper has been partially carried out in the framework of R & D projects Ref. nos. BBF2003-04616 (Funding source: Spanish Ministry of Science and Technology, National R&D Program, and European Union, European Regional Development Fund- ERDF. 2003-2006) and HUM-892 (Andalusian Ministry for Education, Spain. 2006-2009)

In Spain the most representative researching on text automatic text summarization are being conducted in two Universities: on the one hand, University of Alicante, with Acero et al.'s researchs on automatic and personalized summarization within the HERMES project[2]; on the other hand, the Institut Universitari de Linguistica Aplicada-Universitat Pompeu Fabra, where da Cunha (2005) has researched on an automatic text summarization system for medical articles using the *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1988).

3. Automatic summarization

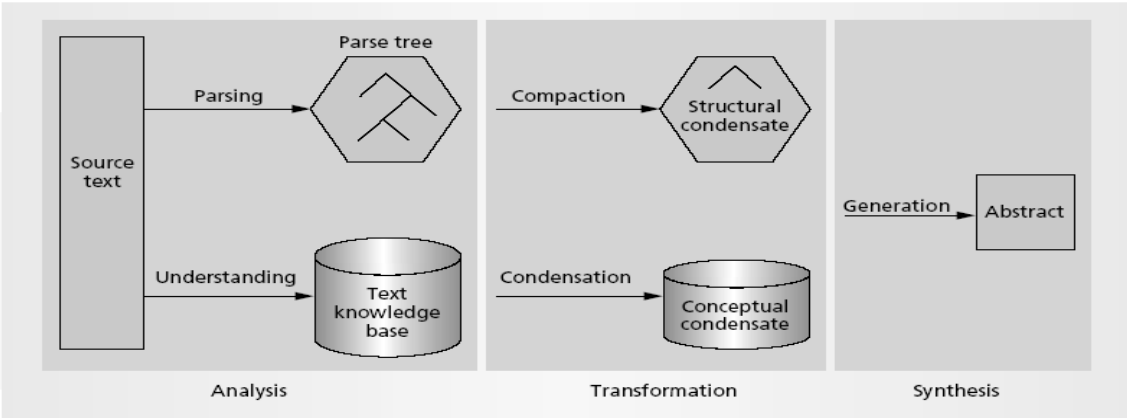
Brandow, Mitze, and Rau (*apud* Mani and Maybury, 2001: 239) point out that automatic summarization of text is a worthy goal of NLP; however, it challenges the field in ways that related areas (e.g. machine translation, extraction of information from text, etc.) do not. This is because in order to achieve readable, appropriate summaries, a system must 1) understand the content of a text at a fairly deep level, 2) be able to ascertain the relative importance of the material, and 3) generate coherent output. Consequently, automatic summarization is an interesting challenge for NLG.

As many researchers underline[3], automatic summarization systems are divided into two categories: *extract* and *abstract*. Teufel and Moens (*apud* Mani and Maybury, 2001: 156) an *extract* is a collection of sentences selected verbatim from a text. Most of the automatic text summarization approaches (Luhn, 1959; Edmundson, 1969) make use of this category since it is really useful for information retrieval (IR) and for an approximation to the content of a text, even though the results do not have a high quality (Mani et al., 1999).



Graph 1. ***Architecture for extraction*** (Hahn and Mani, 2000: 30)

On the other hand, an abstract is a brief digest that summarizes the essential information of an article (Mani et al.,1999; Marcu, 2000). Abstracts often help readers decide whether to read the article itself. According to Hahn and Mani (2002: 32), abstraction has two basic approaches. The first one uses a traditional linguistic method that parses sentences syntactically. The second abstraction has its roots in artificial intelligence and focuses on natural language understanding.



Graph 2. ***Architecture for abstraction*** (Hahn and Mani, 2000: 30)

In this paper we will analyze automatic extraction because it is the only available option in automatic summarization systems on Internet.

[2] HERMES project can be reached at http://nlp.uned.es/hermes/index.html [15-08-06]

[3] Mani and Maybury, 2001; Mani, 2001; Endres-Niggemeyer, 1998; Radev, McKeown and Hovy, 2002; Hahn and Mani, 2000; De Cunha Fanego, 2005.

4. Evaluation of automatic extraction systems

Next, we will analyze and evaluate a group of automatic extraction systems which can be found on Internet for free. Most of these systems merely extract a text, but we have to bear in mind that there are some commercial systems[4] that sum up real abstracts — not extracts — such as *Summarist* (Hovy and Lin, 1999), *DimSum* (Aone et al., 1999) and SCISOR (Rau, Jacobs and Zernik, 1989).

Categorization of input

The input of this analysis will be the terms and conditions of package tours –specifically cruises- taken out from multilingual macrocorpus *Turicor* (Ref. no. BBF2003-04616 (2003-2006, Spanish Ministry of Science and Technology). The multidisciplinary, applied R&D TURICOR project belongs to the field of Information and Communication(s) Technology (ICT) applied to tourist and legal translation.

Given the input, we will work with legal texts. However, it should be added that legal discourse is a sublanguage that, in order to extremely precise, gives some problems to NLP. According to Somers (2003: 286), "an extreme example of a distinctive sublanguage which has numerous features which are quite MT- *un*friendly is 'legalese'". Therefore, it is pertinent to wonder if these automatic systems will be more efficient when working with controlled languages (Rico and Torrejón, 2002; Mitamura, 1999). However, this subject will not be studied in this paper.

Automatic extraction systems available on Internet

In order to analyze how automatic extraction systems work, we needed to find free on-line systems that yield summaries in English and Spanish (and, if possible, in other languages). Most of the systems we found were based upon statistical methods, such as sentence location (Edmundson, 1969), word and sentence frequency (Luhn, 1958) or key words or sentences (e.g., "it is important to consider") (Edmundson, 1969). Furthermore, these systems frequently use the tf.idf weighting ("*Term-frequency times inverse document frequency*") (Salton and McGill, 1983; Hovy and Lin, 1999; Mani and Bloedorn, 1999), which is usually used in information retrieval.

Once presented the common features of these systems, we must specify their names:

1. *Copernic Summarizer*
2. *Extractor*
3. *GistSumm*
4. Resumidor de Microsoft Word
5. *SweSum*
6. *WebSumm*

Analysis and evaluation of outcomes

For the analysis and assessment of the outcomes, we focused on the demo version of an automatic extraction system: *Copernic Summarizer*. Next, we show a part of the extract in English:

Concepts:

reserve, members, travel, booking, contract, charge, request, suppliers, payment, departure, amounts, holiday, costs, arrangements, agents.

Summary:

- B "You, your" and related expressions refer to the lead member and all other members of the party named in the booking and those on whose behalf the booking is made.

- A Where we are supplying you with two or more significant services e.g. flights and accommodation our contract incorporates all of the information on Our Site relevant to the services you wish to purchase, any written materials we or any of our suppliers provide to you, our confirmation invoice and the copy of these booking conditions which we send to you following your telephone reservation or which you accept online.

Then, the extract of terms and conditions in Spanish is the following:

Concepts:

viaje, contrato, consumidor, organizador, viaje combinado, precio, detallista, oferta, condiciones, documentación del viaje, fecha, transporte, Condiciones Generales, folleto, Agencia.

Summary:

- A los efectos de las presentes Condiciones Generales, el programa / folleto es el documento informativo al que éstas se incorporan.

- El programa / oferta es la descripción del combinado contenida en el programa / folleto que constituye el objeto del contrato de viaje combinado.

- La información sobre el programa / oferta contenida en el programa / folleto es vinculante para el organizador o detallista, salvo que concurra alguna de las siguientes circunstancias: Que los cambios en dicha información se hayan comunicado claramente por escrito al consumidor antes de la celebración del contrato y tal posibilidad haya sido objeto de expresa mención en el programa oferta.

[4]On Internet there are many comercial automatic summarization systems, such as: *Brevity Document Summarizer* (http://www.lextek.com/brevity/), *Intelligent Miner for Text* (http://www-306.ibm.com/software/data/iminer/), *Inxight Summarizer* (http://www.inxight.com), *Sinope Summarizer* (http://www.sinope.info/en/Download) and *TextAnalyst* (http://www.megaputer.com/products/ta/index.php3).

Regarding the extracts yielded by *Copernic Summarizer* both in Spanish and English, it is worth mentioning that we find the following advantages:

- It gives a list of concepts, which really helps to know in advance the content of the document.
- Sentences are complete, coherent and consistent.
- Most of the basic elements in a package tour are on the summary, such as “booking”, “reservation”, “compensation”, etc.

Nevertheless, the following shortcomings must be added:

- Some words are often repeated.
- There is some irrelevant information that should not be on a summary.

Utility of automatic extraction for translation process

It is worth pointing out the following remarks concerning the utility of automatic extraction in translation process, specifically for translating terms and conditions of package tours:

- It gives a list of key concepts in two languages and, consequently, one can know beforehand the content of a document. Then, if the translator needs to know the most used terminology in English contracts as well as in terms and conditions, he/she will find ““viaje, contrato, consumidor, organizador, viaje combinado, precio, detallista, oferta, condiciones, documentación del viaje, fecha, transporte, Condiciones Generales, folleto, Agencia”. In English, the most remarkable terms are “reserve, members, travel, booking, contract, charge, request, suppliers, payment, departure, amounts, holiday, costs, arrangements, agents”, among others.
- It is really useful for knowing the main topic of the text. In addition, the most important thing is that abstracts often help translators decide whether to use the text in their documentation process.
- In legal translation, the extract gives indications of the most used structure, for instance, the English structure may + infinitive without to (Borja Albí, 2000: 15) (“Our Site and to create binding legal obligations for any liability you may incur as a result of such use”).
- Human postedition may be required in order to improve the outcomes because the yielding abstracts are not very coherent and, consequently, their effectiveness decreases.

In short, in spite of some shortcomings, automatic extraction systems contribute to the documentation process. Thus, using automatic extraction in the documentation process may be really interesting since translators familiarize themselves with the terminology and the syntax of both parallel texts and other information sources.

Proposal for improving automatic abstraction: MUSI project

We propose a multilingual and automatic summarization system that would be based upon a project sponsored by the EU: MLIS-MUSI project (*Multilingual Summarization for the Internet*) (Lenci et al., 2002). Although a small scale research project, MUSI has tried to tackle the challenges set by multilingual summarization by adopting an original approach based on the definition of a shared ontology and representation language, and on the reuse of existing linguistic resources. MUSI combines a statistic-based module for relevant sentence extraction and a concept-based component to generate multilingual summaries. Thus, our system will be based on this structure and it will be used for professional and educational purposes both in professional translation and in Translation Studies.

5. Conclusions

Multilingual automatic summarization highly facilitates the use of Internet as an information resource, especially for translators, since it makes possible to compile more information in less time and, consequently, to improve translator’s efficiency. Thus, the implementation of the future multilingual automatic summarization program will constitute an interesting tool for both translation teaching and professional translation. Therefore, it tends to be an important and new application of AG and NLG to the Translation Studies.

6. References

ACERO, I. ALCOJOR, M. DÍAZ, A. GÓMEZ, J. M. y MAÑA, M. J. 2001. “Generación automática de resúmenes personalizados”, *XVII Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN'2001)*, 12-14 septiembre 2001, Jaén (España). Publicado en *Procesamiento del Lenguaje Natural*, nº 27, septiembre 2001, 281-287. <http://www.sepln.org/revistaSEPLN/revista/27/> [05-02-06]

AONE, C. M., OKUROWSKI, E., GORLINSKY, J. y LARSEN, B. 1999. “A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques”. En MANI, I. y MAYBURY, M. 1999. *Advances in Automatic Text Summarization*. Cambridge: Massachusetts Institute of Technology.

CORPAS PASTOR, G. (ed.) 2003. *Recursos documentales y tecnológicos para la traducción del discurso jurídico (español, alemán, inglés, italiano, árabe)*. Granada: Editorial Comares.

DA CUNHA, I. 2005. "Hacia un modelo lingüístico de resumen automático de artículos médicos en español". Universidad Pompeu Fabra, Instituto Universitario de Lingüística Aplicada, Doctorado en Ciencias del Lenguaje y Lingüística Aplicada. <http://www.upf.edu/pdi/iula/iria.dacunha/#0202> [07-04-06]

Dpto. de Lenguajes y Sistemas Informáticos, UNED. *Proyecto* Hermes. Disponible en <http://nlp.uned.es/hermes/> [15-04-06]

EDMUNDSON, H.P. 1969. “New Methods in Automatic Abstracting”. En MANI, I. y MAYBURY, M. 1999. *Advances in Automatic Text Summarization*. Cambridge: Massachusetts Institute of Technology.

ENDRES-NIGGEMEYER, B. 1998. *Summarizing Information.* Hannover: Springer- Verlag Berlin Heidelberg.

HAHN, U. y MANI, I. 2000. “The Challenges of Automatic Summarization”. *IEEE Computer*, 33, 2000 (11), 29-36.

HOVY, E. y LIN, C. 1999. “Automated Text Summarization in SUMMARIST”. En MANI, I. y MAYBURY, M. 1999. *Advances in Automatic Text Summarization*. Cambridge: Massachusetts Institute of Technology.

LAVID, J. 2005. *Lenguajes y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI.* Madrid: Cátedra.

LENCI, A., BARTOLINI, R., CALZOLARI, N., AGUA, A., BUSEMANN, S., CARTIER, E., CHEVREAU, K., y COCH, J. 2002. “Multilingual summarization by integrating linguistic resources in the MLIS-MUSI project”. *Proceedings of the 3rd international conference on language resources and evaluation (LREC'02)*, Las Palmas, España. <http://www.coli.uni-saarland.de/publikationen/softcopies/Lenci:2002:MSI.pdf> [03-03-06]

LUHN, H.P. 1969. “The Automatic Creation of literatura Abstracts”. En MANI, I. y MAYBURY, M. 1999. *Advances in Automatic Text Summarization*. Cambridge: Massachusetts Institute of Technology.

MANI, I. y MAYBURY, M. 1999. *Advances in Automatic Text Summarization*. Cambridge: Massachusetts Institute of Technology.

MANN, W.C., y THOMPSON, S.A. 1988. “Rhetorical Structure Theory: Toward a functional theory of text organization”. *Text*, 8 (3), 243-281.

MARCU, D. 2000. *The theory and practice of discourse parsing and summarization.* Cambridge: Massachusetts Institute of Technology.

MITAMURA, T. 1999. “Controlled Language for Multilingual Machine Translation”. *Machine Translation Summit VII*, Singapur. <http://www.lti.cs.cmu.edu/Research/Kant/PDF/MTSummit99.pdf> [10-02-06]

PINTO MOLINA, M. 2000. “Documentación para la Traducción en la sociedad de la información”. *XV Coloquio Association Internationale de Bibliologie***.** Salamanca, AIB, 2000. <http://www.mariapinto.es/web/mainframe.htm> [05-04-06]

RADEV, D. HOVY, E. y MCKEOWN, K. 2002. “Introduction to the Special Issue on Summarization”. En *Computational Linguistics*, 28 (4), 339-408.

RAU, P., JACOBS, S. y ZERNIK, U. 1989. “Information Extraction and Text Summarization using Linguistic Knowledge Acquisition”. *Information Processing and Management*, vol. 25, no. 4, pp.419-428.

REITER, E. y DALE, R. 1997. “Building applied natural-language generation systems”. *Journal of Natural-Language Engineering*, 3:57- 87.

RICO, C. y Enrique TORREJON. 2002. “Controlled Translation: A New Teaching Scenario Tailor-made for the Translation Industry”. *Proceedings of the 6th European Association for Machine Translation Workshop.* UMIST, Manchester.

SALTON, G. y MCGILL, M. 1983. *Introduction to Modern Information retrieval.* New York: McGraw Hill.

SOMERS, H. (ed.) 2003. *Computers and Translation. A translator's guide.* Philadelphia: John Benjamins Publishing Company.