Towards the Automatic Summarization of Medical Articles in Spanish: Integration of textual, lexical, discursive and syntactic criteria[1]

Iria da Cunha and **Leo Wanner**
Institute of Applied Linguistics (IULA), Pompeu Fabra University
Las Ramblas, 30-32, 08002 Barcelona, Spain
{iria.dacunha,leo.wanner}@upf.edu

Abstract

Current text summarization strategies often draw upon one specific type of criteria to locate summary relevant text passages. For instance, they are statistical, discourse structure-based, or positional. In this paper, we argue that in order to arrive at an optimal summary, the whole range of linguistic criteria must be taken into account: textual, lexical, discursive, informative, and syntactic. First preliminary experiments carried out with medical articles in Spanish suggest the validity of our argumentation.

1 Introduction

Current "extract"-oriented text summarization strategies are often "mono dimensional" in that they draw upon one specific type of criteria to identify summary relevant text passages. Some of them use statistical criteria and look thus for sentences that contain high frequency terms (cf., e.g., Luhn 1959; Edmunson 1969). Others use positional criteria, selecting text chunks that appear, e.g., at the beginning of the introductory section, that follow specific headings, etc.; see, among others, (Brandow et al. 1995) and (Lin & Hovy 1997). More recent strategies make use of *lexical chains*, i.e., lexical anaphoric link sequences, (as, e.g., Barzilay & Elhadad 1997; Silber & McCoy 2000) or discourse relations (as, e.g., Marcu 2000; Teufel & Moens 2002). Especially the latter attracted particular attention since they naturally ensure the coherence of the summary by selecting specific branches of the discourse structure of the text in question.

Our work on summarization is centred in specialized language texts. More precisely, we focus on articles in medicine. Medical articles are suitable for the development and evaluation of summarization strategies since they reveal, on the one hand, a predefined textual structure, a rather variable discourse structure, and a number of prominent lexical clues – leaving thus room for the use of a variety of different criteria. On the other hand, they obligatorily contain author-written summaries, which can be considered as point of reference for automatically generated summaries. Our choice of medical articles as application domain is thus motivated by the richness of summarization criteria they provide and, therefore, by the potential they offer to develop holistic portable summarization techniques the quality of which can be verified. It is also motivated by the observation that novice authors in general do not abstract their articles well. That is, high quality automatic summarization is highly useful even in a domain with mandatory author summaries.

In our experiments, we started by evaluating the common one-type-of-criteria summarization strategy. We examined a number of medical articles in Spanish, comparing the summaries obtained with the summaries provided by the authors. Our study showed that, e.g., discourse relations as known from the *Rhetorical Structure Theory* (RST) are essential for summarization, but taken on their own, they do not suffice. Rather, in order to arrive at an optimal summary, different types of linguistic information must be considered as being inter-related. In particular, the interrelation between discourse information, information structure, and (generalized) syntactic information must be taken into account. Additionally, textual positional criteria and lexical criteria should be considered. Currently, we are about to develop a rule-based text summarization model for specialized language documents that incorporates all of the criteria mentioned above. For the representation of the discourse structure, we use RST. For the re-presentation of the syntax, we use the (deep)-syntactic structure (DSyntS) of the *Meaning-Text Theory*, MTT (Mel'cuk 1988).[2] As information (= communicative) structure, we

[1]This work is being carried out in the framework of the project "TEXTERM II: Baselines, strategies and tools for the automatic information processing and extraction" led by T. Cabré and financed by the Spanish Ministry of Education and Science. Da Cunha is supported by a doctorate grant from the UPF. Wanner is member of the Institució Catalana de Recerca i Estudis Avançats (ICREA).

[2] We have chosen DSyntS for the representation of syntactic information because (a) it is dependency-based, and dependency relations between linguistic units can be used to judge the relevance of the dependent unit for the summary; (b) it is language-universal and general, reducing the number of distinctions of dependency relations to a minimum.

use the *Communicative Structure* of the MTT (Mel'cuk 2001). As already mentioned, we focus on medical articles in Spanish.

The remainder of the paper is structured as follows. After motivating why several types of linguistic criteria are needed for the summarization of especially medical articles in Section 2, we discuss in Section 3 the specific features of medical articles. Section 4 analyses the various types of linguistic criteria we use to determine the text passages relevant for the summary. Section 5 sketches our approach, and Section 6 discusses a small experiment. In Section 7, finally, some conclusions are made.

2 Summarization Criteria

Whether isolated statistical or positional criteria suffice to provide good summaries depends on the text type. For medical articles they do not suffice. A glance at the first part of a sample medical article that we cite in the Annex confirms this view. In the case of discourse criteria, the question is more complex. Let us consider it in more detail. According to Marcu (2000), his RST-based summarization approach can fail because: (1) the discourse analyzer does not build adequate discourse tree structures, (2) the algorithm for assigning importance scores to the elements of the discourse tree is too simple. Marcu assumes that discourse structure elements that are located higher in the discourse tree are more important than the elements at the bottom. A simple way of quantifying the importance of an element within a tree is to calculate its score on the basis of its distance to a distinguished dominant element (in RST, this would be a *nucleus* of a relation). The higher the score of the element, the more important it will be considered for the summary. This strategy favours relation nuclei over their dependents (i.e., *satellites*). However, as Marcu admits, in certain cases, it is necessary to give more importance to satellites. He offers an example where his program does not take into account two elements that human specialists in his experiment did not hesitate to select for the summary.

Ex. 1. "[Smart cards have two main advantages over magnetic-stripe-card.[3]] [First, they can carry 10 or even 100 times as much information[4]] [–and hold it much more robustly.[5]][Second they can execute complex tasks in conjunction with a terminal.[6]]"

All experts selected units 3, 4 and 6 for the summary, whereas the program selected only unit 3. This is due to the low scores assigned to 4 and 6 because of their satellite status in the elaboration relation with unit 3. We encounter a similar case in our corpus:

Ex. 2. "[El análisis de regresión logística identificó tres variables asociadas, de forma independiente, con una visita apropiada a urgencias:[1]] [acudir a este servicio por indicación de un médico,[2]] [vivir fuera de la región respecto a residir en la ciudad en la que está el hospi-tal[3]] [y pertenecer a los grupos de consultas quirúrgicas y traumatismos respecto a la enfermedad médica y pe-diatría.[4]]" [*The regression of logistic analysis identified three associated variables, in an independent way, with an appropriate visit to emergency services:[1]*] [*to go to this service for indication of a doctor,[2]*] [*to live out of the region with regard to living in the city in which the hospital is located[3]*] [*and to belong to the groups of surgical and traumatism consulting in contrast with medical illnesses and paediatrics.[4]*]"

In other words, as isolated statistical and positional criteria, isolated discourse criteria also do not suffice for an adequate summary.

3 On the Genre *Medical articles*

The texts of the genre "medical journal articles" on which we focus in our work have a predefined fixed structure, which consists of four sections: 1. Introduction, 2. Targeted patients and methods applied, 3. Results, and 4. Discussion. This structure is known from the literature as the *IMRD-structure.*

The *American National Standards Institute* (ANSI) defines the summary in a scientific area as: "an abbreviated, accurate representation of the contents of a document, preferably prepared by its prepared by its author(s) for publication with it" (Bhatia 1993). Therefore, it is not surprising that in the above genre, the summary written by the author of the article is required to reflect the same four sections as encountered in the main article.

For evaluation of automatically produced summaries, we consider the summary written by an experienced author of the original article as "ideal". This is because (a) the author is a specialist on the subject, and (b) the journal in which the article is published gives some guidelines for how the summary should be written. To verify our assumption, we have carried out a small empirical experiment. The experiment is not representative but it gives us, nonetheless, some hints with respect to the correctness of our assumption. In this experiment,

three medical doctors and three linguists were asked to compile summaries of five medical articles (of which they did not see the authors' summaries) by extracting the relevant passages from the article. Restrictions concerning the maximal length of the summary were given.

Figure 1 shows the degree of coincidence (made explicit in quantitative terms using *Multi-dimensional Scaling*) between the summaries of the participants of our experiment and the summaries of the authors. The summaries of the authors and the summaries of other specialists in the field are very similar, while the summaries of the linguists deviate between each other and between the summaries of the experts significantly. We believe that the deviance is due to the lack of expert knowledge by the linguists – which makes them take as main summarization criteria the discourse structure of the text. This suggests that for the summaries of medical articles textual, lexical and other field-specific criteria are of primary importance and cannot be neglected in favour of a predominance of discourse criteria.

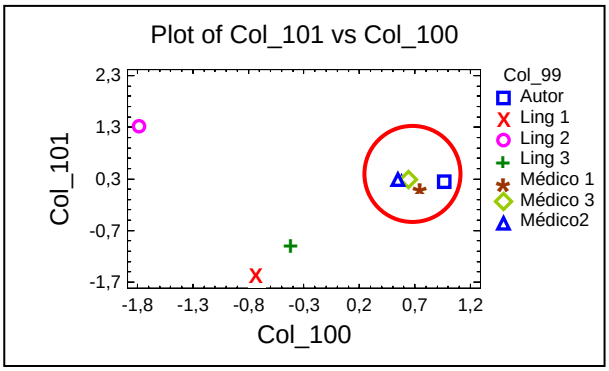

Figure 1: *The similarity of the summaries of the subjects with the summary of the authors*

4 Summarization Clues

Let us discuss now the linguistic clues that we take into account for summarization (textual, lexical, discourse, syntactic, and communicative).

4.1 Textual Clues

In accordance with our statements on the predefined structure of medical articles and their summaries, we require that

- the summary must include passages from each section of the article.

Identification of sections is facilitated by the font and layout of the headings – although it must be taken into account that the headings might differ. Thus, we encountered, among others, the following headings for Section 2: "Patients and Methods", "Subjects and Method", "Material and Method", "Method", "Population and Method".

Furthermore, we require that

- the summary must contain a passage from the last part of each section.

The second criterion is justified by the empirical study carried out on a corpus of medical articles (Da Cunha 2005).

4.2 Lexical Clues

Following the idea that in the given genre certain cue words indicate relevance, our summarization model is also based on a set of lexical criteria. An empirical study we carried out shows that the set of cue words includes the Spanish equivalents of such nouns as

> *objective, object, summary, purpose, intention, result, etc.*

and of such verbs as

> [*to*] *carry out,* [*to*] *associate,* [*to*] *analyze, [to] present, [to] relate,* [*to*] *evaluate,* [*to*] *contribute,* [*to*] *study,* [*to*] *value,* [*to*] *find, etc.*

Detecting one or several of these units in the last part of each section, the model will have a first output of sentences (those that include the units) that will later be contrasted and refined by discourse criteria, syntactic, and communicative criteria discussed immediately below.

4.3 Discourse Clues

Previous work has shown (cf. also Section 2 above) that the discourse structure of a text can be successfully exploited for purposes of summarization. Especially discourse relations as defined by the *Rhetorical Structure Theory*, RST (Mann & Thompson 1988) proved to be useful. An RST-based discourse structure is based on a set of different notions such as predominance of structures with nucleus-satellite patterns, functionality hierarchy, and the communicative role of a given discourse structure element. Among RST-relations used to express these notions are: Circumstance, Elaboration, Motivation, Evidence, Justification, Cause, Purpose, and Condition.

In general, we adopt Marcu's idea that if separated from its nucleus, a satellite of a relation violates the criteria of text coherence and cannot be easily understood. Therefore, nucleus-satellite spans and nuclei are the natural text chunks to be examined for inclusion in a summary.

As is well-known, the automatic detection of RST-relations in general language is a difficult task since by far not all relations in a text are marked by explicit *lexical discourse markers*.

This also applies to medical articles (cf. Da Cunha 2004).

4.4 Syntactic Clues

Another type of clues we use in our model are syntactic dependency relations. As mentioned above, we work with syntactic trees defined by *deep-syntactic dependency relations* (DSyntRels) from the *Meaning-Text Theory* (MTT); cf. (Mel'cuk 1988). The alphabet of DSyntRels consists of numbered actant relations (I, II..., VI), a modifier relation (ATTR), an appenditive relation (APPEND) and a coordinative relation (COORD).

4.5 Communicative Clues

The last type of clues in our model is provided by the Communicative structure of the MTT (Mel'cuk 2001). For the time being, we use only the Theme-Rheme (*Topic-Focus*) opposition known from the works of the *Prague School* (Sgall et al. 1986). For the automatic detection of Topic-Focus in texts, see, e.g., (Hajicova et al. 1995).

4.6 Putting the Clues Together

Each type of the clues discussed above addresses a different level of linguistic description. This ensures a comprehensive coverage of all aspects of text writing, leading to an optimal summary. The contribution of each type of clues to the summarization of a medical article can be resumed as follows:

- *Textual clues*: ensure that the summary contains information on each part of the article.
- *Lexical clues*: ensure that the summary contains information on methods, experiments, etc. described in the article.
- *Discourse clues*: ensure that the summary is coherent and provide a strong hint on the relevance of text chunks that are related to other text chunks by discourse relations.
- *Syntactic clues*: provide a strong hint on the relevance of intra-sentential syntactic units; corroborate or weaken the hypothesis derived from the discourse clues if between the elements of a discourse relation additionally a syntactic dependency relation holds.
- *Communicative clues*: provide a hint on the "thematic progression" of the text and thus also on the length of spans elaborating on the same topic; corroborate or weaken the hypothesis derived from the discourse clues if between the elements of a discourse relation a thematic progression relation holds.

To the best of our knowledge, no work on summarization considered so far the relevance of communicative clues, or the combination of syntactic, communicative and discourse clues.

5 Our Approach

As pointed out above, one of our main research premises is that in order to arrive at an adequate summary, it does not suffice to consider one type of criteria. Rather, textual, lexical, discourse, syntactic and communicative criteria must be considered together.

We envisage a cascaded three level model: first, textual criteria are applied, then, lexical criteria, and, finally, a combination of discourse, syntactic and communicative criteria. The criteria are formulated in terms of rules; cf. three sample syntactic/communicative-discourse rules:

(1)
IF *S* is satellite of ELABORATION E_1
and *S* is ATTR of an element of the nucleus of E_1
THEN ELIMINATE *S*

(2)
IF *S* is satellite of ELABORATION E_1
and S is APPEND of an element of the nucleus of E_1
THEN ELIMINATE *S*

(3)
IF *S* is satellite of ELABORATION E_1
and *S* elaborates on the Rheme of the nucleus *N* of E_1
and the Theme of *S* is equal to the Theme of *N*
THEN KEEP *S*

The compilation of rules is still ongoing work, and experiments are being carried out to validate our theoretical assumptions.

Consider the application of the above first two rules to a fragment of an article from our corpus.

Ex. 3. "[Coincidiendo con ese mismo estudio,] [la visita previa a un médico es un factor asociado a una mayor adecuación del uso del servicio de urgencias,] [lo que estaría en relación con el papel de filtro de la atención primaria.]" *Ex. "[As another research already stated,] [the previous visit to the doctor is a factor associated to a major adequacy of the use of emergency services,] [what would be related to the filter role played by primary care.]"*

By the application of (1) and (2), the sentence is reduced in the following way:

Ex. 4. “La visita previa a un médico es un factor asociado a una mayor adecuación del uso del servicio de urgencias.” *"The previous visit to a doctor is a factor associated to a major adequacy of the use of emergency services."*

The result turns out to be nearly identical to the corresponding part of the original summary provided by the author.

Let us now reconsider Ex1 from Section 2. In Ex1, we encounter a first deep-syntactic actant (**I**) (*They*) and a second deep-syntactic actant (**II**) (*advantages*), followed by further details on **I**. More precisely, the two satellites elaborate on the Rheme of the nucleus, and the Theme of the nucleus is taken up by the Themes of the satellites. That is, the application conditions of rule (3) from above are fulfilled – which leaves us with the correct selection of the information for the summary.

In Ex2, the situation is analogous (*analysis* is here the deep-syntactic actant **I**, and *variables* the deep-syntactic actant **II**). Rule (3) can also be applied to Ex2, rendering the important information for the summary as suggested by human experts.

6 A Small Experiment

To verify the efficiency of our criteria we carried out an additional experiment in which our current lexical, discursive, syntactic and communicative criteria have been applied to the Introduction section of five medical articles selected at random. Since it was a test in which only one section was drawn upon, the application of textual criteria was not considered.

After the application of the syntactic/ communicative-discourse rules, we observed that the most relevant content of the section (always in comparison with the summary written by the author) is being selected correctly in 4 of the 5 summaries. The Annex contains the first section of one single article and the corresponding summary.

The coincidence between lexical and discourse/communicative/syntax criteria has been high because all fragments selected by discourse/ communicative/syntax criteria contained lexical units from the list compiled for the use as lexical clues. On the other hand, some fragments contained lexical items from the list, without being selected by the discourse/syntax/ communicative criteria. Furthermore, we observed that the density of lexical units from the lexical clue list is considerably higher in sentences in which all criteria coincide than in other sentences.

7 Conclusions and Future Work

In this paper, we presented some initial exploratory work that is to be considered as the first attempt to determine the possibilities of integrating several types of summarization criteria (textual, lexical, syntactic, communicative and discourse criteria) in order to arrive at adequate text summaries.

The summarization rules we discussed in the course of this paper also do not constitute a definitive proposal. They must be evaluated with more material. However, we think that they already allow us to observe how the integration of the various types of criteria affects the result of summarization and where further research work is needed.

Once it is clear which clues need to be considered and how the individual clues are inter-related, i.e., once the theoretical linguistic model has been worked out, a number of different approaches can be envisaged to realize the summarization model. One option is to continue with a rule-based implementation pursued so far. Another option is to apply machine learning techniques.

References

BARZILAY, R. & M. ELHADAD (1997): “Using Lexical Chains for Text Summarization”. Proceedings of the Workshop on Intelligent Scalable Text Summarization at the Annual Meeting of the ACL.

BHATIA, V. (1993): *Analyzing genre: Language Use in Professional Settings*. Londres: Longman.

BRANDOW, R., K. MITZE & L. RAU (1995): “Automatic condensation of electronic publications by sentence selection”. Information Processing and Management, 31.

DA CUNHA, I. (2004): “Importancia del marcaje de las relaciones discursivas para la generación automática de resúmenes”. Proceedings of the 6º Congreso de Lingüística General. Universidad de Santiago de Compostela.

DA CUNHA, I. (2005): "Hacia un modelo lingüístico de resumen automático de artículos médicos en español". Thesis Research Project. IULA. Universitat Pompeu Fabra.

EDMUNDSON, H.P. (1969): "New Methods in Automatic Extraction". Journal of the Association for Computing Machinery, 16. California: ACM Press.

HAJICOVA, E., H. SKOUMALOVA & P. SGALL (1995): "An Automatic Procedure for Topic-Focus Identification". Computational Linguistics, 21(1).

LIN, C. & HOVY E. (1997): "Identifying Topics by Position". Proceedings of the Applied Natural Language Processing Conference. Washington: ACL.

LUHN, H. P. (1959): "The Automatic Creation of Literature Abstracts". IBM Journal of Research and Development, 2. New York: IBM Journal.

MANN, W. C. & S.A. THOMPSON (1988): "Rhetorical structure theory: Toward a functional theory of text organization". Text, nº 8, vol.3.

MARCU, D. (2000): *The Theory and Practice of Discourse Parsing Summarization*. Cambridge, MIT Press.

MEL'CUK, I. (1988): *Dependency Syntax*: Theory and Practice. New York, Albany

MEL'CUK, I. (2001): *Communicative Organization in Natural Language*: The semantico-communicative structure of sentences. Amsterdam: John Benjamins.

SGALL, P., E. HAJICOVA & JA. PANEVOVA (1986): *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Dordrecht & Prague: D. Reidel & Academia.

SILBER, H.G. & K. MCCOY (2000): "Efficient Use of Lexical Chains for Text Summarization". Proceedings of the 5th International Conference on Intelligent User Interfaces.

TEUFEL, S. & M. MOENS (2002): "Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status". Computational Linguistics, 28.

ANNEX

The Annex contains the first section of a medical article and the corresponding summary compiled in accordance to our criteria.

1. Estudio seroepidemiológico frente a citomegalovirus en mujeres en edad fértil de la Comunidad de Madrid

Introduction:

El citomegalovirus (CMV) es el más importante agente productor de infección congénita en España, especialmente después de la reducción del número de casos producidos por el virus de la rubéola, como consecuencia de la vacunación.

La infección congénita sintomática por CMV está más relacionada con la infección primaria que con la reinfección o la recurrencia, aunque no todos los niños nacidos de madres con seroconversión al virus durante el embarazo desarrollan la enfermedad congénita. La determinación de anticuerpos en el suero permite establecer el estado inmunitario con respecto al virus, e identificar a las mujeres susceptibles de sufrir una infección primaria.

Los estudios de seroprevalencia realizados en nuestro país no han sido realizados sobre población general, y no se han empleado técnicas sensibles. El objetivo del presente estudio es el conocimiento de la prevalencia de anticuerpos anti-CMV en las mujeres en edad fértil de la Comunidad de Madrid, así como el estudio de diversos factores de riesgo asociados a la presencia de anticuerpos frente al CMV, para lo que se han empleado las muestras obtenidas para el estudio de la seroprevalencia frente a las enfermedades vacunables, hepatitis C y virus varicela zoster en el marco de la II Encuesta Seroepidemiológica de la Comunidad de Madrid.

Abstract:

El objetivo del presente estudio es el conocimiento de la prevalencia de anticuerpos anti-CMV en las mujeres en edad fértil de la Comunidad de Madrid, así como el estudio de diversos factores de riesgo.