Resumen automático de artículos médicos en castellano: integración de técnicas de análisis textual, léxico, discursivo y sintáctico-comunicativo

Iria da Cunha Fanego*, Leo Wanner**

* Instituto Universitario de Lingüística Aplicada (IULA), Universidad Pompeu Fabra
C/ Ramblas, 30-32, Barcelona, 08002
** ICREA y Departamento de Tecnología de la Universidad Pompeu Fabra
Passeig de Circunvalació, Barcelona, 08003
iria.dacunha@upf.edu, leo.wanner@upf.edu

Resumen

Uno de los principales problemas en la investigación sobre resumen automático es la falta de utilización de conocimiento lingüístico, que se refleja en múltiples carencias del resumen resultante. En este trabajo se propone explotar conocimiento lingüístico de varios tipos (en concreto, textual, léxico, discursivo y sintáctico-comunicativo) para llevar a cabo un modelo de resumidor automático de artículos médicos en castellano. Así, explotaremos la información textual, la información discursiva mediante las relaciones de la *Teoría de la Estructura Retórica* (Mann & Thompson 1988), la información sintáctico-comunicativa a partir de la *Teoría Sentido-Texto* (Mel'cuk 1998) y, finalmente, la información léxica presente en estos textos.

Palabras clave: Resumen automático, artículo médico, estructura textual, léxico, estructura discursiva, sintaxis de dependencias, estructura comunicativa.

Abstract

One of the main problems in the research on automatic summarization is the scarce use of linguistic knowledge, which is reflected in multiple deficiencies of the resulting summaries. In this paper, we suggest to exploit linguistic knowledge of several types (textual, lexical, discourse and syntactic-communicative knowledge) in a model of automatic summarization of medical articles in Spanish. The discourse information is used in terms of the relations of the *Rhetorical Structure Theory* (Mann & Thompson 1988) and the syntactic-communicative information following the dependency structures proposed in the *Meaning-Text Theory* (Mel'cuk 1998). As lexical information, we use key words determined in an empirical study of the domain.

Key words: Automatic summarization, medical article, textual structure, lexicon, discursive structure, dependency syntax, communicative structure.

Resum

Un dels principals problemes en la investigació sobre resum automàtic és el poc ús de coneixement lingüístic, que es reflecteix en múltiples mancances del resum resultant. El propòsit d'aquest treball és explotar coneixement lingüístic de diversos tipus (en concret, textual, lèxic, discursiu i sintàctic-comunicatiu) per dur a terme un model de resumidor automàtic d' articles mèdics en castellà. Així, explotarem la informació textual, la informació discursiva mitjançant les relacions de la *Teoria de l' Estructura Retòrica* (Mann & Thompson 1988), la informació sintàctico-comunicativa a partir de la *Teoria Sentit-Text* (Mel'cuk 1998) i, finalment, la informació lèxica present en aquests textos.

Paraules clau: Resum automàtic, article mèdic, estructura textual, lèxic, estructura discursiva, sintaxi de dependències, estructura comunicativa.

Tabla de contenidos

1. Introducción[1]

La sociedad actual está inmersa en un devenir de información que en una gran cantidad de ocasiones nos sobrepasa. No disponemos del tiempo suficiente para asimilar toda esta información que nos llega y debemos ser conscientes de que necesitamos discriminar entre lo que realmente nos interesa y nos beneficia de cara a nuestros propósitos (ya sean intelectuales o profesionales), y lo que no. En este sentido sería útil acceder a los resúmenes de los documentos de nuestro interés, entendiendo por resumen de textos en general una condensación de los conceptos principales del contenido del texto al que hace referencia (Burgos et al. 1994) y por resumen en el ámbito científico (*abstract*): "an abbreviated, accurate representation of the contents of a document, preferably prepared by its author(s) for publication with it" (American National Standards Institute; Bhatia 1993:78). Por esto, desde hace años se está investigando sobre resumen automático, como un recurso que nos sirve de ayuda a la hora de seleccionar los documentos (textuales o no) que realmente nos interesan. Esta idea de economía está muy ligada al concepto de sociedad actual, donde no debe desperdiciarse tiempo ni esfuerzo analizando contenidos innecesarios. El resumen automático es uno de los múltiples recursos que se utilizan hoy en día para actuar de este modo y la investigación en este ámbito sigue vigente, ya que hay muchas carencias que deben ser solventadas.

Las estrategias actuales aplicadas al resumen automático suelen ser monodimensionales, es decir, que utilizan un solo tipo de criterio para llevar a cabo el resumen. Algunas de estas estrategias se centran en la búsqueda de oraciones con términos de alta frecuencia (Luhn 1959; Edmunson 1969); otras se basan en la posición de determinados fragmentos del texto (Brandow et al. 1995; Lin & Hovy 1997); algunas explotan secuencias de palabras vinculadas con relaciones léxico-semánticas, es decir, cadenas léxicas (Barzilay & Elhadad 1997; Silber & McCoy 2000); muchas se basan en modelos estadísticos (como por ejemplo, los modelos bayesianos (Kupiek et al. 1995)); otras explotan la estructura discursiva de los textos (Marcu 1998; Teufel & Moens 2002; Alonso 2005), etc. Algunas estrategias combinan dos tipos de técnicas, como por ejemplo, las cadenas léxicas y la estructura discursiva de los textos (Alonso & Fuentes 2003); pero la combinación de diversas técnicas de cara al resumen automático aún no ha sido explotada con detalle.

En nuestro estudio se trabajará sobre documentos textuales, en concreto, artículos médicos en castellano. Hemos decidido restringir el género al artículo y el ámbito a la medicina porque, como demuestran diversos experimentos, desarrollar un programa de resumen automático de textos en general resulta muy costoso y es difícil llegar a buenos resúmenes sin restringir el ámbito de los documentos que se desee resumir. Además los artículos médicos se publican con sus correspondientes resúmenes redactados por su mismo autor, lo cual nos servirá de ayuda a la hora de evaluar los resultados. Consideramos adecuado el resumen del propio autor de un artículo médico de investigación porque es él mismo quien lo escribe y por tanto sabrá qué contenidos incluir, porque dichos artículos están orientados a especialistas en la materia, porque el autor es uno de estos especialistas y porque la revista en donde se publicará el texto

[1] El presente artículo se enmarca en una tesis doctoral sobre resumen automático que se está llevando a cabo en el Instituto Universitario de Lingüística Aplicada (IULA) de la Universidad Pompeu Fabra de Barcelona. Se integra a su vez en un proyecto de investigación más amplio denominado "TEXTERM II: Fundamentos, estrategias y herramientas para el procesamiento y extracción automáticos de información" financiado por el Ministerio de Educación y Cultura Español y dirigido por Mª Teresa Cabré en el marco del grupo Iulaterm.

marca unas pautas especificando qué información debe incluirse en cada apartado.[2] Los textos sobre los que trabajaremos pertenecen a la revista Medicina Clínica y a su vez forman parte del Corpus Técnico[3] del Instituto Universitario de Lingüística Aplicada de la Universidad Pompeu Fabra de Barcelona (en concreto del subcorpus de medicina).

A continuación ofreceremos un segundo apartado que se refiere al artículo médico, como tipo de texto que se desea resumir. Después, en el tercero, se tendrá en cuenta la problemática del resumen automático, en cuanto que el resumen integra diversas informaciones lingüísticas que deberán ser tenidas en cuenta. En el cuarto apartado aportaremos el marco teórico del trabajo y realizaremos un acercamiento, por un lado, a la estructura discursiva en términos de la *Rhetorical Structure Theory* (Mann & Thompson 1988) y, por otro, a la noción de estructura sintáctica de dependencias y a la estructura comunicativa (básicamente la contraposición entre Tema y Rema), ambas integradas en la *Meaning-Text Theory* (Mel'cuk 1998). En el quinto apartado explicaremos los criterios específicos que tenemos en cuenta para alcanzar un resumen justificado lingüísticamente (textuales, léxicos, discursivos y sintáctico-comunicativos) y llevaremos a cabo la formalización de dichos criterios. En el sexto, los aplicaremos sobre textos concretos, extrayendo algunas conclusiones. Finalmente dedicaremos el séptimo apartado a ofrecer conclusiones generales y a establecer vías de trabajo futuras.

2. El artículo médico

Debido a nuestra intención de resumir artículos médicos en castellano creemos conveniente empezar este estudio aclarando su estructura y contenidos habituales. Además, como ya hemos mencionado, partimos de la base de que el resumen que acompaña al texto, redactado por el mismo autor, refleja de manera adecuada los contenidos más relevantes del artículo original, lo cual nos será de ayuda a la hora de validar los resultados de nuestro modelo.

El ámbito de la medicina está cubierto por una gran variedad de géneros discursivos pero en este estudio nos centraremos en el Artículo Original. Este tipo de textos y sus respectivos resúmenes siguen un patrón habitual solicitado por las revistas para su publicación que consiste en cuatro apartados que se caracterizan por seguir el orden lógico del pensamiento científico: Introducción, Pacientes y métodos, Resultados y Discusión. Esta estructura se denomina habitualmente en la bibliografía como *estructura IMRD*. Los cuatro apartados principales mencionados siguen unas directrices marcadas por las revistas en las que se publican. A continuación concretamos qué información debe contener cada uno de los apartados.

El apartado correspondiente a la Introducción deberá ser breve y aportar sólo la explicación necesaria para que el lector pueda comprender el texto que sigue a continuación, no debe contener tablas ni figuras, y debe exponer de forma clara el/los objetivo/s del trabajo.

El apartado de Pacientes y métodos debe indicar el centro donde se ha realizado el experimento o la investigación, el período de duración, las características de la serie

[2] Para corroborar esta hipótesis de que el resumen del autor es adecuado hemos realizado un experimento con una técnica estadística denominda *Multidimensional Scaling* que puede encontrarse detallado en Da Cunha & Wanner (2005). Mediante esta validación empírica se corrobora que los resúmenes de los autores de artículos médicos son adecuados y, por tanto, podremos compararlos con los ofrecidos por nuestro modelo de resumidor.

[3] Puede accederse al programa de explotación de este corpus desde la siguiente dirección electrónica: http://bwananet.iula.upf.edu/

estudiada, el criterio de selección empleado y las técnicas utilizadas, proporcionando los detalles suficientes para que una experiencia determinada pueda repetirse sobre la base de esta información. Además deben describirse con detalle los métodos estadísticos.

El apartado de Resultados debe resaltar, que no interpretar, las observaciones efectuadas con el método empleado.

En el apartado de Discusión los autores tienen que exponer sus propias opiniones sobre el tema e interpretar los resultados.

3. El problema del resumen automático

Para ilustrar la problemática del resumen automático, observaremos los apartados de Introducción y de Resultados de un artículo médico titulado "Estudio seroepidemiológico en mujeres en edad fértil de la Comunidad de Madrid" (Texto 1) y el resumen que de ellos ha redactado su autor (Texto 2).

Texto 1

Introducción

1[El citomegalovirus (CMV) es el más importante agente productor de infección congénita en España,] 2[especialmente después de la reducción del número de casos producidos por el virus de la rubéola como consecuencia de la vacunación.] 3[La infección congénita sintomática por CMV está más relacionada con la infección primaria que con la reinfección ola recurrencia,] 4[aunque no todos los niños nacidos de madres con seroconversión al virus durante el embarazo desarrollan la enfermedad congénita.] 5[La determinación de anticuerpos en el suero permite establecer el estado inmunitario con respecto al virus,] 6[e identificar a las mujeres susceptibles de sufrir una infección primaria.] 7[Los estudios de seroprevalencia realizados en nuestro país no han sido realizados sobre población general] 8[y no se han empleado técnicas sensibles.] 9[El objetivo del presente estudio es el conocimiento de la prevalencia de anticuerpos anti-CMV en las mujeres de edad fértil de la Comunidad de Madrid,] 10[así como el estudio de diversos factores de riesgo asociados a la presencia de anticuerpos frente al CMV,] 11[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.]

Resultados

12[La prevalencia de anticuerpos en el total de la muestra estudiada ha sido del 75,6%.] 13[Por grupos de edad,] 14[la seroprevalencia ha sido del 60,6% en el grupo de 15-24 años 15[(IC del 95%, 66, 4-54,9),] del 84,4% en el grupo de 25-35 años 16[(IC del 95%, 89, 0-79,8)] y del 94,6% en el grupo de 36-45 años 17[(IC del 95%, 99, 0-90,2)]]. 18[Las diferencias en la seroprevalencia entre los tres grupos de edad han sido estadísticamente significativas.]

19[Al estudiar los factores de riesgo,] 20[se ha encontrado que haber tenido hijos, 21[independientemente del número], aumenta la probabilidad de tener anticuerpos 3,3 veces.] 22[Igualmente, el vivir en condiciones de hacinamiento 23[(< 20 m^2/persona)] aumenta la probabilidad 9,74 veces,] 24[mientras que el tener estudios universitarios aparece como factor de protección,] 25[siendo la probabilidad de tener anticuerpos en las mujeres que han realizado estudios universitarios 2,5 veces menor que en el resto.] 26[Por otra parte, en este estudio no aparece relación entre el estado inmunitario y el haber recibido transfusiones.]

Texto 2
Introducción Determinar la prevalencia de anticuerpos frente al citomegalovirus en mujeres de la Comunidad de Madrid, y estudiar diversos factores de riesgo. **Resultados** La seroprevalencia ha sido del 75,6% y se ha incrementado desde el 60,6% (15-24 años) hasta el 94,6% (36-45), siendo significativamente mayor en las mujeres con hijos y en las que viven en condiciones de hacinamiento, y menor en las universitarias.

La información más relevante del apartado de Introducción, es decir, la que indica cuál es el objetivo del presente artículo, se incluye en las unidades 9 y 10, mientras que la más relevante del apartado de Resultados, es decir, la que aporta los datos que se derivan de la investigación, está incluida en las unidades 12, 14, 20, 22 y 24. Estos contenidos coinciden con aquellos ofrecidos en el resumen del autor del texto. El problema reside en encontrar las informaciones lingüísticas que nos llevan a estas inferencias para, posteriormente, poder sistematizarlas de cara a un modelo de resumidor automático.

Veamos ahora observaciones lingüísticas de diversos tipos que señalan fragmentos importantes para el resumen.

En primer lugar, en relación con la estructura textual del Texto 1, habrá que seleccionar contenidos de cada uno de los dos apartados para mantener la coherencia, tal y como hace el autor en el resumen del Texto 2.

En segundo lugar, si nos fijamos en la estructura del discurso, tendremos que tener en cuenta que en el Texto 1 se dan una serie de relaciones discursivas que aportan informaciones que pueden ser relevantes a la hora de resumirlo. Por ejemplo, el elemento 1 aporta una información previa sobre los elementos 9 y 10 (es decir, ofrece los antecedentes del tema sobre el que se habla) y el elemento 11 aporta una información adicional a ellos (es decir, los elabora). También los elementos 13, 15, 16 y 17 elaboran al 14, ya que aportan informaciones adicionales de las que podría prescindirse en el resumen.

En tercer lugar, la estructura sintáctico-comunicativa también puede ofrecernos informaciones de interés. El primer argumento (sintáctico) del elemento 1, por ejemplo, introduce el tema (comunicativo) del que se hablará en el artículo (*citomegalovirus* o *CMV*), que se repite más adelante en los elementos 9 y 10, que indican el objetivo de la investigación. Así mismo, el primer argumento del elemento 12 introduce el tema sobre el que versará el apartado de Resultados (*prevalencia de anticuerpos*).

Finalmente, las unidades léxicas del texto también ofrecen pistas acerca de cuáles son los fragmentos relevantes. Por ejemplo, el sustantivo *objetivo* en el elemento 9 indica que en este fragmento se especifica la intención del trabajo que se presenta; y la forma verbal *se ha encontrado* señala que a continuación se ofrecerá algún tipo de resultado.

Apoyándonos en estos criterios se observa, por ejemplo, que las informaciones adicionales (elaboraciones) y las informaciones previas (antecedentes) de algún elemento pueden no ser significativas para el resumen, mientras que sí lo serán los elementos que indiquen el tema principal del discurso. Además ciertas unidades léxicas, como hemos visto, serán indicadoras de relevancia. Si aplicamos todos estos criterios llegaremos a un resumen adecuado, que

coincidirá con el del autor del artículo. Así, como ilustra este ejemplo (Texto 1) de manera simplificada, es esencial disponer de información lingüística de varios tipos para obtener un buen resumen. Por esto, nuestro trabajo consiste en el desarrollo de un conjunto de reglas en las se formalizan criterios lingüísticos de varios tipos para utilizarse en un sistema de resumen automático.

En el apartado siguiente introducimos los marcos teóricos que modelizan estos criterios.

4. Marco teórico del trabajo

Para poder entender la visión que se propone del resumen creemos conveniente situarnos de manera general en el marco teórico de donde partimos. Por tanto, haremos una breve presentación de las ideas principales de cada uno de los marcos con la intención de que, una vez asimiladas éstas, pueda entenderse la perspectiva de integración que se propone.

4.1 Estructura discursiva

Para el análisis de los textos desde la perspectiva discursiva tomaremos como base teórica la *Rhetorical Structure Theory* (RST) de Mann & Thompson (1988). La RST se creó en el marco de estudios de generación automática de textos y tiene validez en la actualidad como una teoría descriptiva de organización del texto muy útil para describirlo caracterizando su estructura a partir de las relaciones que mantienen entre sí los elementos discursivos del mismo (*Circunstancia, Elaboración, Motivación, Evidencia, Justificación, Causa, Propósito, Antítesis, Condición,* entre otras). Estas relaciones pueden ser asimétricas o simétricas: en las primeras el elemento principal se denomina "núcleo" y el secundario "satélite", mientras que en las segundas todos los elementos son "núcleos". Los satélites no serían comprensibles separados del núcleo y podrían ser fácilmente sustituibles, lo que nos lleva a enlazar esta teoría con el ámbito de la generación de resúmenes automáticos.

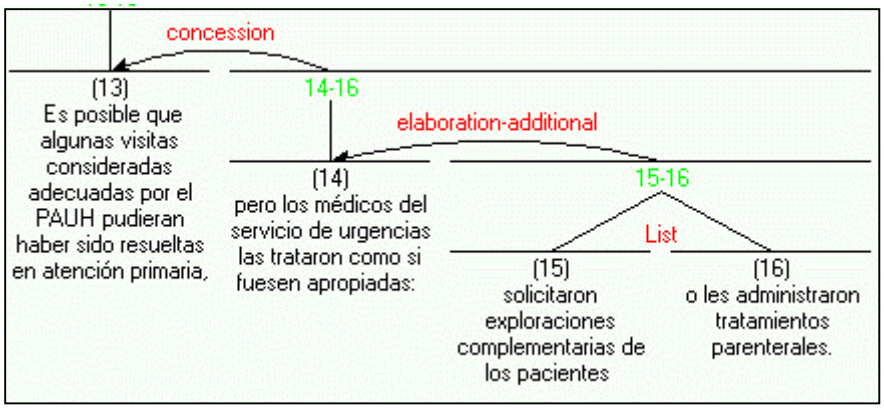

Gráfico 1. Fragmento de estructura arbórea con relaciones de la RST[4]

En el Gráfico 1 observamos un ejemplo de un fragmento de estructura arbórea con relaciones de la RST, en donde se ofrece una relación de *Concesión*, otra de *Elaboración* y una última

[4] Para la representación gráfica de las estructuras discursivas utilizaremos la RSTtool desarrollada por Marcu en la que se proponen varias relaciones de Elaboración. La más general, que se corresponde con la Elaboración genérica de la RST, se denomina *elaboration-additional.*

relación *Multinuclear* de *Lista*[5]: "Es posible que algunas visitas consideradas adecuadas por el PAUH pudieran haber sido resueltas en atención primaria, pero los médicos del servicio de urgencias las trataron como si fuesen apropiadas: solicitaron exploraciones complementarias o les administraron tratamientos parenterales."

4.2 Estructura sintáctica

Para el análisis de los textos desde la perspectiva sintáctica utilizaremos la sintaxis de dependencias que se integra en la *Meaning-Text Theory* (MTT) de Mel'cuk (1988). En concreto, en este estudio llevaremos a cabo un análisis de dependencias profundo, un subtipo de sintaxis de dependencias. Ésta se concibe como un árbol cuyos nodos son etiquetados por unidades léxicas y cuyos arcos son etiquetados por relaciones actanciales (I, II, ..., VI) y circunstanciales en un sentido amplio: atributiva (ATTRIB), apenditiva (APPEND) y coordinativa (COORD).

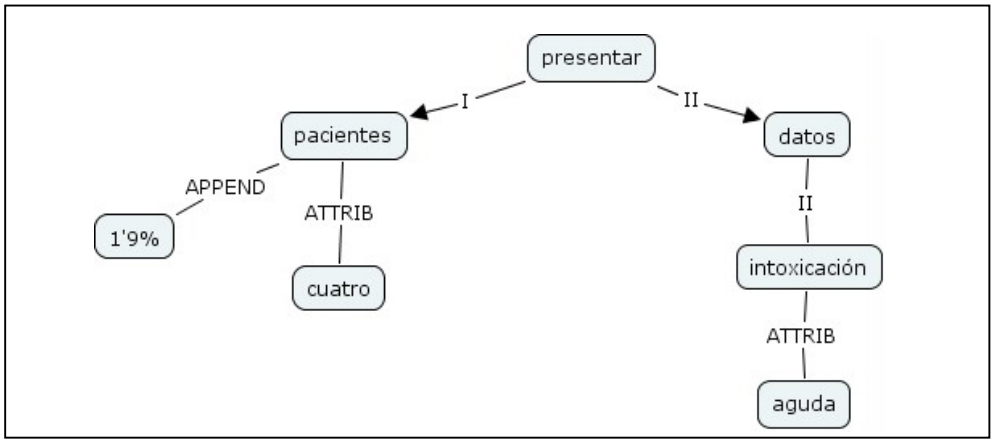

Gráfico 2. Fragmento de estructura sintáctica profunda de dependencias

En el Gráfico 2 observamos un ejemplo de estructura arbórea de dependencias con un verbo (*presentar*) con dos actantes, I (*pacientes*) y II (*datos*), de los que a su vez se desprenden otros elementos (apenditivo y atributivo del primero; actante II con atributivo del segundo): "Cuatro pacientes (1'9%) presentaron datos de intoxicación aguda".[6]

4.3 Estructura comunicativa

Para el análisis de la estructura comunicativa nos basaremos también en la *Meaning-Text Theory*. Mel'cuk (2001) afirma que para obtener una completa representación de la estructura semántico-comunicativa se requieren al menos ocho oposiciones comunicativas: *Thematicity, Giveness, Focalization, Perspective, Emphasis, Presupposedness, Unitariness* y *Locucionallity*. Para nuestros propósitos de cara al desarrollo de un modelo de resumidor automático de momento será pertinente la primera de las ocho oposiciones. *Thematicity* abarca la contraposición entre Tema y Rema[7] elementos que Mel'cuk (2001:18) define como:

[5] En las relaciones asimétricas las flechas se orientan desde el satélite hacia el núcleo.
[6] Para no complicar la representación, el árbol no incluye rasgos morfosintácticos como tiempo, modo, persona, número, etc.
[7] En la oposición comunicativa de *Thematicity* también habría que considerar un tercer elemento, el *Sem-Com-Specifier,* que de momento no tendremos en cuenta para este trabajo.

"The Communicative Predicate is that part of the meaning of an utterance which is presented (by the Speaker) as being communicated. It is also called the Rheme, or Comment. The Communicative Subject is what the Rheme applies to and communicates about. It is called the Theme, or Topic."

Esta elección se debe a que dicha oposición será más relevante de cara al resumen que las otras siete, porque nos indicará de qué se habla en cada oración y qué se dice acerca de ello, lo cual nos resultará de gran ayuda para discernir cuál será la información indispensable para el resumen automático.

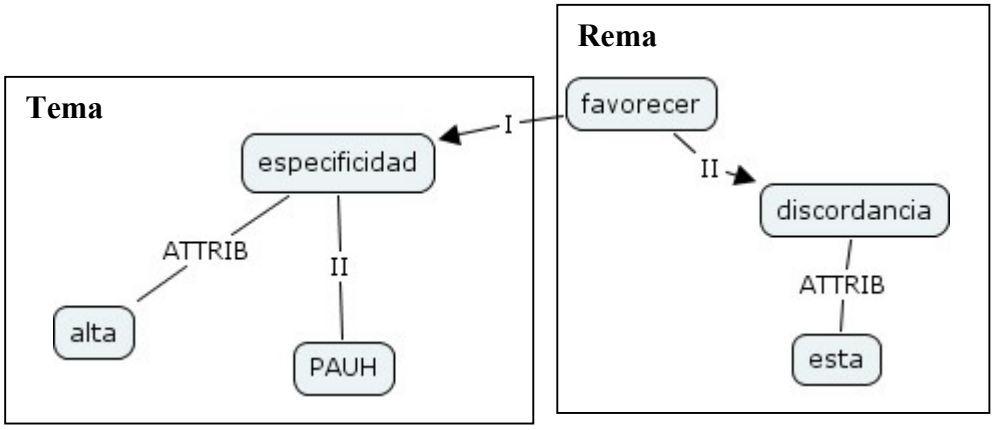

Gráfico 3. Fragmento de estructura sintáctica profunda de dependencias con especificación del Tema y el Rema

En el Gráfico 3 observamos un fragmento de estructura sintáctica profunda de dependencias en donde se especifica además la estructura comunicativa (Tema y Rema) de la oración: “La alta especificidad del PAUH favorece esta discordancia”.

5. Hacia un resumen justificado lingüísticamente

Como ya hemos mencionado, para llevar a cabo nuestro modelo de resumidor automático de artículos médicos en castellano, integraremos varias informaciones lingüísticas.

Partimos de la idea de que un texto es multidimensional, es decir, que si uno desea formarse una impresión global del contenido del mismo, éste debe ser observado desde varias perspectivas: textual, léxica, discursiva y sintáctico-comunicativa. Varios aspectos específicos de estas perspectivas ofrecen datos que nos guían a la hora de seleccionar o descartar informaciones para el resumen. A partir de estos supuestos previos, se llevará a cabo el resumen, en cuanto texto también, de la misma manera; es decir, se realizará un análisis de los textos desde dichas perspectivas para posteriormente alcanzar un resumen que tenga en cuenta todas ellas.

Los criterios lingüísticos empleados y su formalización serán explicados en detalle a continuación.

5.1 Criterios lingüísticos

5.1.1 Criterios textuales

La primera información lingüística que integrará nuestro modelo de resumidor automático será la textual, basada en el reconocimiento de los apartados del texto: Introducción, Pacientes y métodos, Resultados y Discusión. Las revistas médicas de investigación solicitan, además del artículo que se presenta, un resumen adjunto dividido en los cuatro apartados del texto inicial. Así, a la hora de escribir un resumen de este tipo deben seleccionarse contenidos de cada uno de estos apartados para mantener el orden lógico del pensamiento científico-médico que se sigue en el artículo.

Para el reconocimiento de los apartados nos hemos basado en sus títulos y, para ello, hemos realizado un estudio de los mismos sobre nuestro corpus. Observamos que, aunque los títulos de los apartados están en principio fijados, hay ligeras variaciones entre ellos que deben ser tenidas en cuenta para poder llevar a cabo un reconocimiento efectivo. Por ejemplo, hemos encontrado las siguientes variantes del segundo apartado del artículo médico: *Pacientes y métodos, Sujetos y método, Material y método, Método, Población y método.*

Además habrá que tener en cuenta cómo se distribuye la información dentro del texto. Normalmente los contenidos relevantes suelen encontrarse al final de cada uno de los apartados, con lo cual se considerarán más importantes aquellos fragmentos que se encuentren en el último párrafo, siendo candidatos para el resumen.

5.1.2 Criterios discursivos y sintáctico-comunicativos

La aplicación de criterios discursivos y sintáctico-comunicativos se refleja mediante la integración de las relaciones discursivas de la RST con las relaciones de sintaxis profunda de dependencias y la oposición entre Tema y Rema de la MTT.

Hay algunos estudios (Marcu 2000) que consideran que la sola utilización de la estructura discursiva de los textos es válida para llegar al resumen. En ellos se muestran resultados positivos, pero se admite también que existen carencias. Como ya se ha mostrado anteriormente (Da Cunha & Wanner 2005), el solo análisis de la estructura discursiva es insuficiente para llegar a un resumen adecuado ya que en determinadas ocasiones se necesita un análisis más profundo. Así, creemos necesaria la integración de dicha estructura discursiva con la estructura sintáctica (de dependencias) y la estructura comunicativa, la cual llevaremos a cabo mediante el desarrollo de una serie de reglas.

La metodología de trabajo para la creación de estas reglas discursivo-sintáctico-comunicativas (desde ahora y para abreviar "Reglas DiSiCo") consiste en analizar las estructuras discursivas, sintácticas y comunicativas de nuestro "corpus de análisis" formado por artículos médicos de investigación en castellano. Las reglas desarrolladas recogen información sobre los contenidos que deben mantenerse en el texto y sobre aquellos que deben eliminarse, comparándose posteriormente con los resúmenes de los autores (que aportan el conocimiento especializado en la materia) para su validación.

5.1.3 Criterios léxicos

La aplicación de criterios léxicos sigue la idea de que en este tipo de textos hay ciertas unidades que son indicadoras de relevancia. Se trataría básicamente de sustantivos del tipo *objetivo, objeto, resumen, propósito, intención, resultado,...,* y de verbos frecuentes en este tipo de textos que no sean auxiliares ni posean un significado muy general (*realizar, asociar, analizar, presentar, relacionar, evaluar, aportar, estudiar, valorar, incluir, observar, llevar a cabo, obtener, alcanzar, encontrar*).

A continuación mostramos fragmentos extraídos de nuestro corpus (en los que se incluyen algunas de estas unidades) cuyos contenidos coinciden con aquellos ofrecidos en el resumen del autor:

Ej. I "Los *resultados* indican que la infección por el CMV es frecuente entre los 15 y 45 años, pudiendo afectar en este período hasta el 30% de las mujeres."

Ej. II "*Se han estudiado* un total de 692 muestras de suero, que representan a las mujeres en edad fértil de la Comunidad de Madrid."

Ej. III "El *objetivo* de nuestro trabajo fue analizar las características del FRA en el TaloPH y sus variaciones según las etiologías implicadas."

Ej. IV "Posteriormente, *hemos analizado* las características clínicas de los 92 pacientes que presentaron FRA."

5.2 La formalización de los criterios

En los subapartados siguientes indicamos cómo pueden formalizarse los criterios empleados, teniendo en cuenta que estos no son excluyentes.

5.2.1 Formalización de criterios textuales

Partimos de dos criterios textuales principales:

1) El modelo dividirá el texto en las cuatro secciones mencionadas, para posteriormente seleccionar algunas oraciones de cada una de ellas.

2) Serán candidatos para el resumen aquellos contenidos que se incluyan en las dos últimas oraciones del párrafo final de cada sección, lo cual se formalizará de la siguiente manera:

Given $O_1 \ldots O_n$ in last Paragraph of section Se
Keep O_{n-1} and O_n

Como hemos visto, esto ocurre en el Texto 1, ya que algunas de las informaciones relevantes se encuentran en estas posiciones.

5.2.2 Formalización de criterios discursivos y sintáctico-comunicativos

Como ya hemos mencionado, la parte central del modelo de resumidor consta de una serie de reglas lingüísticas que se basan en la integración de las relaciones discursivas de la RST con

las relaciones de sintaxis profunda de dependencias y la oposición entre Tema y Rema de la MTT.

Observemos a continuación algunas de las Reglas DiSiCo acompañadas de ejemplos extraídos de nuestro corpus de análisis:

3) **IF** *S* is satellite of ELABORATION *E*
and *S* is ATTRIB of an element of the nucleus of *E*
THEN ELIMINATE *S*

Esta primera regla nos indica que si encontramos un satélite de una relación discursiva de *Elaboración* que además sea un ATTRIB (sintáctico), éste puede ser eliminado. En el siguiente ejemplo, por tanto, podría eliminarse para el resumen el fragmento que aparece subrayado.

Ej. V "El objetivo de este estudio fue analizar la evolución de la prevalencia de infección por el VIH en las madres de los nacidos entre 1996 y 1999 en siete comunidades autónomas (CCAA), que representan el 26,5% de la población y el 25% de los nacimientos en España."

4) **IF** CONTRAST between *N1* and *N2*
and *N1* and *N2* possess different Themes and different Rhemes
THEN KEEP *N1* and *N2*

Esta segunda regla indica que si encontramos una relación discursiva *Multinuclear* de *Contraste* en la que los dos núcleos de la relación posean diferentes Tema y Rema, debemos mantener información de ambos. En los siguientes ejemplos los Remas aparecen en cursiva y los Temas en negrita.

Ej. VI "La prevalencia del **VIH-1** encontrada se corresponde *con una situación endémica*, mientras que el **VIH-2** presenta *un patrón de casos esporádicos*."

5) **IF** *S* is satellite of INTERPRETATION of the nucleus *N*
and *S* is ATTRIB of an element of N (with connector "por/con lo que/cual")
THEN ELIMINATE *N* and the connector

Esta tercera regla indica que si encontramos una relación discursiva de *Interpretación* cuyo satélite sea además un ATTRIB (sintáctico) con un conector del tipo "por /con lo que/cual", el núcleo de la relación y el conector que los une, pueden ser eliminados para el resumen. Por tanto, en el siguiente ejemplo podría descartarse el fragmento que aparece subrayado.

Ej. VII "En la región de París se identificaron prevalencias del VIH dos veces superiores en mujeres que interrumpían voluntariamente el embarazo que en madres de nacidos vivos, por lo que la seroprevalencia del VIH entre las mujeres que inician el embarazo será probablemente mayor que la encontrada en madres de recién nacidos."

El conjunto de Reglas DiSiCo se aplicará sobre un corpus de contraste (seleccionado al azar pero también formado por artículos médicos) para observar los resúmenes resultantes y compararlos posteriormente con los de sus autores para comprobar la coincidencia de los contenidos que ambos ofrecen.

5.2.3 Formalización de criterios léxicos

La tercera parte del modelo se basa en la aplicación de una serie de criterios léxicos. Si el resultado de la aplicación de los criterios discursivos y sintáctico-comunicativos ofreciese un resumen más largo de lo esperado al seleccionar una cantidad de contenidos más elevada de la necesaria, habría que afinar mediante estos criterios léxicos. Esto se llevará a cabo dando una mayor importancia a los fragmentos que contengan las unidades léxicas de nuestra lista: se ofrecerá a cada uno de estos fragmentos un punto más de relevancia por cada unidad léxica encontrada en su interior. Así se establece una gradación entre los fragmentos seleccionados por los criterios discursivos y sintáctico-comunicativos. Esto se formaliza de la siguiente manera:

6) Given the list of key words L
IF a sentence S contains a lexeme l ∈ L
THEN keep S

En el apartado 5.1.3 pueden observarse ejemplos en los que esta regla ofrece buenos resultados.

6. Aplicación de criterios lingüísticos para el resumen

Para comprobar la efectividad de nuestros criterios y observar si las reglas correspondientes seleccionan los contenidos más relevantes de los artículos, en comparación con los resúmenes redactados por los autores, hemos realizado un experimento: hemos aplicado las reglas de las que disponemos hasta el momento (textuales, léxicas, discursivas y sintáctico-comunicativas) sobre un corpus de contraste seleccionado al azar consistente en 3 artículos médicos de investigación.

Se observa que los resultados son positivos y prometedores, ya que la información seleccionada es adecuada en los tres resúmenes realizados a partir de la aplicación de nuestros criterios. De todas maneras habría que realizar algunas observaciones lingüísticas de cara a afinar los resultados.

El procedimiento que se ha seguido en la aplicación de criterios se explica a continuación.

En primer lugar, se ha dividido cada texto en cuatro partes (los cuatro apartados del artículo) para seleccionar información relevante de cada uno de ellos (Regla 1).

En segundo lugar se han aplicado las Reglas DiSiCo sobre cada una de las partes y se han seleccionado fragmentos relevantes (más adelante observaremos algunos ejemplos).

En tercer lugar, se ha observado si en el interior de estos fragmentos aparecen unidades léxicas de nuestra lista de elementos relevantes (sustantivos y verbos) (Regla 6).

Veamos un ejemplo. Después de la aplicación de las Reglas DiSiCo sobre el apartado de Discusión de uno de los textos del corpus de contraste se han eliminado varios fragmentos, mientras que otros han sido seleccionados. De entre estos, habrá que escoger los que serán incluidos en el resumen ya que, en esta ocasión, después de la aplicación de las reglas se han obtenido demasiados fragmentos con la misma importancia. Para solucionar este tipo de situaciones, como ya hemos comentado, se ofrecerá un punto por cada unidad léxica

encontrada en estos fragmentos; finalmente el que tenga más puntos se incluirá en el resumen, repitiéndose el procedimiento hasta alcanzar la longitud deseada. Por ejemplo, de entre estas dos oraciones, se seleccionará la segunda de ellas por contener *ser valorado*, ya que el verbo *valorar* es una de las unidades léxicas de nuestra lista.

Ej. VIII "La utilización de técnicas como el lavado gástrico, la endoscopia, la extracción manual transanal o el uso de laxantes por vía rectal para intentar extraer los paquetes aumenta el riesgo de rotura de los mismos, por lo que se desaconseja su uso."

Ej. XIX "Si aparecen signos de obstrucción intestinal, el enfermo debe *ser valorado* por el servicio de cirugía para ser intervenido quirúrgicamente."

La información incluida en el Ej. XIX coincide con la ofrecida en el resumen del autor del artículo, con lo cual se confirma su validez de cara a la inclusión en el resumen.

Volviendo a los contenidos del texto seleccionados por las Reglas DiSiCo, hay varias cuestiones que deben ser matizadas.

Una de las carencias del modelo se refleja en el hecho de que de momento no puede reconocer situaciones de correferencia. Esto es un problema ya que, por un lado, en ocasiones una oración es correctamente seleccionada para el resumen pero carece de referente, con lo cual leída por sí sola no tiene todo el sentido que debería; y por otro, hay ocasiones en que la longitud del resumen podría ser menor si se solucionasen ciertos casos de correferencia. Veamos un ejemplo.

Ej. X "[La radiografía simple de abdomen demostraba cuerpos extraños en más del 90% de los casos,] [con lo que consideramos esta técnica el método radiológico de elección tanto en el diagnóstico como en el seguimiento.]"

Después aplicar sobre el Ej. X la Regla 5, observamos que la primera parte entre corchetes de la oración puede ser eliminada para el resumen.

Así, se mantendría la segunda parte de la oración, que incluye la información relevante, eliminando el conector *con lo que*. El problema de correfencia en este caso lo encontramos en la secuencia *esta técnica*, que se refiere a *radiografía simple de abdomen*.

Debemos investigar en este sentido para solucionar casos de este tipo y lograr generar resúmenes de mayor calidad.

7. Conclusiones y trabajo futuro

En este artículo se ha presentado un trabajo exploratorio que puede ser considerado como un intento de integración de varias perspectivas lingüísticas (textual, léxica, discursiva y sintáctico-comunicativa) de cara a la creación de un modelo de resumidor automático, en concreto, de artículos médicos en castellano. Hemos explicado qué informaciones aporta cada una de las perspectivas y cómo podemos aunarlas de manera efectiva para poder aplicarlas a esta tarea. Consideramos especialmente novedoso el desarrollo de las Reglas DiSiCo, que integran discurso (con relaciones de la RST), sintaxis (en concreto la sintaxis profunda de dependencias integrada en la MTT) y estructura comunicativa (en términos de Tema-Rema, también integrada en la MTT).

Como ha podido observarse, los resultados de los primeros experimentos son prometedores, pero queda aún mucho trabajo por delante después de esta exploración inicial, que no es un análisis exhaustivo, sino más bien un intento de configurar una primera visión que nos permita determinar las posibilidades de integración de dichas perspectivas lingüísticas.

Queda por delante el desarrollo del conjunto definitivo de reglas a partir del análisis de más textos, la validación del mismo y el desarrollo de diferentes métodos de evaluación de la propuesta, teniendo en cuenta que se considera que el resumen del autor del artículo es adecuado en este tipo de textos y, por tanto, podemos tomarlo como modelo.

8. Referencias bibliográficas

Alonso, L. (2005). *Representing discourse for automatic text summarization via shallow NLP*. Tesis. Universidad de Barcelona.

Alonso, L. & M. Fuentes (2003). *Integrating cohesion and coherence for Automatic Summarization*. Actas de la EACL'03 Student Session. Budapest: ACL. 1-8.

Barzilay, R.& M. Elhahad (1997). "Using lexical chains for text summarization". Actas del ACL/EACL Workshop on Intelligent Scalable Text Summarization. Madrid: ACL. 10-17.

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

Brandow, R.; Mitze, K.; Rau, L. (1995). "Automatic condensation of electronic publications by sentence selection". *Information Processing and Management* 31. 675-685.

Burgos, R., J. A. Chicharro & M. Bobenrieth (1994). *Metodología de investigación y escritura científica en clínica*. Escuela andaluza de salud pública. Granada.

Da Cunha, I. & L. Wanner (2005). "Towards the Automatic Summarization of Medical Articles in Spanish: Integration of textual, lexical, discursive and syntactic criteria". Actas del Workshop "Crossing Barriers in Text Summarization Research". RANLP-2005 (*Recent Advances in Natural Language Processing*). Borovets (Bulgaria).

Edmundson, H. P. (1969). "New Methods in Automatic Extraction". *Journal of the Association for Computing Machinery* 16. 264-285.

Kupiec, J.; Pedersen, J. O. & Chen, F. (1995). "A trainable document summarizer". Actas de la 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR-95). Nueva York: ACM. 68-73.

Lin, C. & Hovy E. (1997). "Identifying Topics by Position". Actas de la ACL Applied Natural Language Processing Conference. Washington: ACL. 283-290.

Luhn, H. P. (1959). "The Automatic Creation of Literature Abstracts". *IBM Journal of Research and Development* 2. Nueva York: IBM Journal. 159-165.

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

Marcu, D. (1998). *The rhetorical parsing, summarization, and generation of natural language texts*. Thesis. Department of Computer Science, University of Toronto.

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

Mel'cuk, I. (1988). *Dependency Syntax*: *Theory and Practice*. Nueva York: Albany.

Mel'cuk, I. (2001). *Communicative Organization in Natural Language. The semantic-communicative structure of sentences*. Amsterdam: John Benjamins.

Silber H. G. & K. F. McCoy (2000). "Efficient text summarization using lexical chains". Actas de la *Conference on Intelligent User Interfaces (IUI'2000)*. Nueva York: ACM. 252-255.

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