IRIA DA CUNHA*,,*****
JUAN-MANUEL TORRES-MORENO**
GERARDO SIERRA*
Instituto de Ingeniería (Universidad Nacional Autónoma de México, México D.F., México)*
Laboratoire Informatique d'Avignon (Université d'Avignon et des Pays de Vaucluse, Avignon, Francia)**
Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra, Barcelona, España)***
iria.dacunha@upf.edu, juan-manuel.torres@univ-avignon.fr, GSierraM@iingen.unam.mx

Aplicaciones lingüísticas del análisis discursivo automático

1. Introducción

El análisis del discurso es un tema de investigación ampliamente tratado desde hace años. Una de las teorías discursivas más utilizadas hoy en día en procesamiento del lenguaje natural escrito es la *Rhetorical Structure Theory* (RST) de Mann y Thompson (1988). La RST se ha empleado en diversas aplicaciones, como generación de texto, resumen automático, traducción automática, extracción de información, etc. Es importante notar que, en la mayor parte de estos trabajos, las lenguas empleadas fueron el inglés, el japonés y el portugués, ya que estas son las tres únicas lenguas que disponen de analizadores discursivos automáticos basados en la RST. Para el inglés encontramos el analizador de Marcu (2000a, 2000b), para el japonés el analizador de Sumita et al. (1992) y para el portugués de Brasil el analizador de Pardo y Nunes (2008). El desarrollo de analizadores discursivos en otras lenguas es necesario para poder emplear la RST en aplicaciones de lingüística computacional. Por este motivo, actualmente tenemos un proyecto de investigación en curso para desarrollar el primer analizador discursivo automático para el español (ADAe). El objetivo de este trabajo es presentar el estado actual de este proyecto y, sobre todo, mostrar cuatro de sus aplicaciones que nos interesan especialmente y en las que estamos trabajando.

En este artículo, exponemos nuestro marco teórico en el apartado 2. En el apartado 3 mencionamos el estado actual del proyecto ADAe. En el apartado 4 detallamos las principales posibles aplicaciones del análisis discursivo automático. Finalmente, en el apartado 5, subrayamos algunas conclusiones del trabajo.

2. Marco teórico

La *Rhetorical Structure Theory* es una teoría organizativa del texto que describe su estructura a partir de las relaciones que se establecen entre sus diferentes elementos discursivos. Así, establece un listado de relaciones internas del texto, en las que, por lo general, uno de los elementos es el gobernante (*núcleo*), mientras que el otro aporta cierta información acerca de él (*satélite*). El esquema estructural más frecuente es el de dos unidades de texto (casi siempre adyacentes, aunque hay excepciones) relacionadas de tal manera que una de ellas tiene un papel específico con respecto a la otra: se denominan relaciones *Núcleo-Satélite*. Un ejemplo es el de una afirmación que aporta una información básica acerca de alguna cuestión, seguida de una información adicional sobre la misma. La RST establece en este caso una relación de *Elaboración* entre las dos unidades. La relación también expresa que la afirmación es más relevante en el texto que la información adicional, de tal manera que la afirmación se convierte en el núcleo de la relación y la información adicional en su satélite. Otras relaciones de este tipo serían: *Circunstancia*, *Elaboración*, *Motivación*, *Evidencia*, *Justificación*, *Causa*, *Propósito*, *Antítesis*, *Condición*, etc. En el caso de relaciones que no presentan una unidad central con respecto a los propósitos del autor, la relación se denomina *Multinuclear*. Un ejemplo lo constituye la relación *Multinuclear* de *Lista*, en la cual se realiza una enumeración de varios elementos que tienen la misma importancia. Otras relaciones de este tipo son: *Contraste*, *Secuencia* o *Unión*. En la Figura 1 mostramos un ejemplo de un fragmento de estructura arbórea con relaciones de la RST, en donde se ofrece una relación *Núcleo-Satélite* de *Concesión*, una relación *Núcleo-Satélite* de *Elaboración* y una relación *Multinuclear* de *Lista*. El fragmento analizado es el siguiente: "Es posible que algunas visitas consideradas adecuadas por el PAUH pudieran haber sido resueltas en atención primaria, aunque los médicos del servicio de urgencias las trataron como si fuesen apropiadas: solicitaron exploraciones complementarias de los pacientes o les administraron tratamientos parenterales."

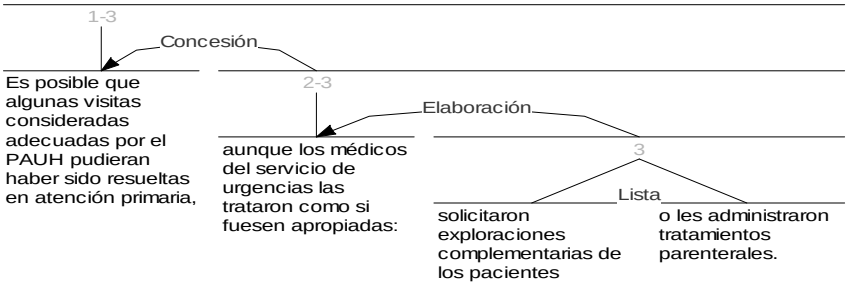

Figura 1. Fragmento de estructura arbórea con relaciones de la RST

3. Estado actual del proyecto ADAe

La idea del desarrollo de un analizador discursivo automático para el español surge al detectar su gran potencial de uso para diversas aplicaciones relacionadas con la lingüística computacional. Como ya hemos comentado, existe un analizador de este tipo para el portugués de Brasil (Pardo y Nunes, 2008). Al ser el portugués y el español lenguas muy cercanas, decidimos emplear una metodología similar a la seguida para la realización de

dicho analizador. Así comenzó el proyecto ADAe, que supone una colaboración entre grupos de investigación de varias universidades: el equipo TALNE (Laboratoire Informatique d'Avignon), el grupo NILC (Universidade de São Paulo), el grupo GIL (Instituto de Ingeniería de la Universidad Nacional Autónoma de México) y el grupo Iulaterm (Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra de Barcelona).

En análisis discursivo automático consta de tres etapas: I) segmentación discursiva, II) detección de relaciones discursivas y III) desarrollo de árboles discursivos. Con respecto a la etapa (I), hemos desarrollado un segmentador discursivo (da Cunha et al., en prensa) que puede consultarse en línea: http://daniel.iut.univ-metz.fr/~iula/WebDiSeg/. El sistema emplea reglas léxicas y sintácticas basadas en marcas como conectores discursivos, conjunciones, adverbios, verbos, signos de puntuación, etc. Nuestra concepción de segmentación discursiva similar a la de Tofiloski et al. (2009: 77):

> "Discourse segmentation is the process of decomposing discourse into elementary discourse units (EDUs), which may be simple sentences or clauses in a complex sentence, and from which discourse trees are constructed".

Así, consideramos que una EDU debe incluir un verbo y reflejar claramente una relación discursiva. Por ejemplo, la oración 1a se dividiría en dos EDUs, mientras que la oración 1b constituiría una única EDU:

> 1a. [El hospital es conocido por sus tratamientos innovadores para enfermedades infecciosas raras,]$_{EDU1}$ [pero también tiene éxito en la cura de pacientes con enfermedades genéticas.]$_{EDU2}$
>
> 1b. [El hospital es conocido por sus tratamientos innovadores para enfermedades raras, además de las enfermedades genéticas.]$_{EDU1}$

Para la consecución de la etapa (II) llevamos a cabo el análisis discursivo de un corpus de cara a obtener un listado de patrones léxicos que reflejen relaciones discursivas. Por ejemplo, el marcador "aunque" indicaría la existencia de una relación de *Concesión*; una secuencia formada por un determinante, alguna palabra indicadora de propósito (ejs. "objetivo", "objeto", "propósito", etc.), algún adjetivo opcional (y demás variantes) indicaría la existencia de una relación de *Propósito*.

Una vez realizada la segmentación discursiva y la detección de las relaciones discursivas mediante patrones, en la etapa (III), aplicamos el algoritmo probabilístico de Marcu (2000a) para el desarrollo de árboles discursivos.

Toda esta información se vuelca en una plataforma multilingüe de trabajo colaborativo denominada DiZer 2.0., disponible en línea: http://www.nilc.icmc.usp.br/dizer2/.

4. Posibles aplicaciones del análisis discursivo automático

Como ya hemos comentado, el análisis discursivo automático puede ser muy útil para el desarrollo de diversas aplicaciones. En este trabajo detallamos las aplicaciones en las que nosotros trabajamos. Para un análisis detallado de los autores que han empleado la RST en el desarrollo de aplicaciones lingüísticas, remitimos al trabajo de Taboada y Mann (2005). Nuestro trabajo toma como base este artículo para seleccionar algunas referencias y además lo complementa, aportando nuevas investigaciones y abriendo nuevas líneas de trabajo y aplicaciones.

4.1. Resumen automático

Las técnicas de resumen automático pueden dividirse en tres tipos: técnicas de nivel superficial, medio y profundo. Las técnicas de nivel superficial son aquellas que no explotan en profundidad la estructura lingüística de los textos, pero sí emplean ciertos aspectos lingüísticos de los mismos para detectar los fragmentos más relevantes. Las técnicas de nivel medio utilizan algún tipo de información lingüística más elaborada que las técnicas de nivel superficial, pero no tanto como las técnicas de nivel profundo, que emplean, por ejemplo, la estructura discursiva (véanse, entre otros, los trabajos de Corston-Oliver, 1998; Marcu, 2000a, 2000b; Teufel y Moens, 2002). Algunas de estas técnicas de nivel profundo toman como base teórica la RST. Estos sistemas de resumen automático se basan en la idea de que un texto viene definido por su estructura interna y las relaciones discursivas que la forman, dando más importancia a los componentes nucleares de dichas relaciones. En concreto, Marcu (2000a, 2000b) parte de la segmentación del texto en unidades discursivas mínimas y del conjunto de relaciones que pueden mantener entre ellas para proporcionar una formalización de la estructura retórica arbórea, con una orientación hacia el resumen automático. Además, como ya hemos mencionado en la introducción, desarrolla un analizador discursivo automático para el inglés, basado en gran medida en el uso de marcadores discursivos. Una vez creada automáticamente la estructura discursiva de un texto en términos de la RST, aplica un algoritmo que proporciona un peso y un orden a cada elemento discursivo de la estructura (cuanto más alto esté el elemento en la estructura, más peso tendrá, y a la inversa), seleccionando para el resumen los elementos con mayor peso y eliminando aquellos que tengan el peso más bajo. Dependiendo de la longitud que se desee para el resumen se escogerán más o menos elementos, pero siempre siguiendo el orden fijado por el algoritmo.

La mayor parte de los resumidores basados en la RST se han desarrollado para el inglés, aunque también podemos encontrar trabajos para el japonés (Ono et al., 1994; Sumita et al., 1992) y para el portugués de Brasil (Pardo y Rino, 2001). Esto se debe a que, como ya hemos comentado, existen analizadores discursivos para estas lenguas.

Sin embargo, también existen trabajos que emplean la RST para el resumen automático en español. En da Cunha (2008) se toma la RST como base para el desarrollo de un modelo de resumen automático de textos médicos en español. La principal estrategia es la eliminación o selección para el resumen de determinados elementos (núcleos o satélites) de la estructura discursiva de textos médicos, derivada de un análisis previo de un corpus de artículos médicos y de sus correspondientes resúmenes escritos por sus autores. Sin embargo, este modelo de resumen no pudo implementarse computacionalmente en su totalidad debido a la carencia actual

de analizadores discursivos para el español. Otros trabajos en esta línea son los de da Cunha et al. (2009), donde se presenta un resumidor híbrido que aúna técnicas estadísticas (concretamente el Modelo del Espacio Vectorial y estrategias de física estadística) y técnicas lingüísticas (basadas en la RST). El resumidor obtiene mejores resultados que cualquiera de las dos técnicas por separado, confirmando la adecuación de la combinación de la lingüística y la estadística en aplicaciones de lingüística computacional. No obstante, el resumidor no pudo implementarse por la misma razón.

4.2. Extracción de información

Existen algunos trabajos relacionados con el resumen automático que podrían considerarse también desde la óptica de la extracción de información. Por ejemplo, Shinmori et al. (2002) extraen la idea más importante de aplicaciones de patentes en japonés a partir de su análisis con la RST; Haouam y Marir (2003) usan la RST con fines de indexación para extraer una información más completa que las tradicionales palabras clave. Sin embargo, consideramos que la RST está poco explotada en este sentido y que podrían existir nuevas aplicaciones en donde ser empleada. Por ejemplo, un posible contexto de uso del análisis discursivo automático en el ámbito de la extracción de información sería el dominio médico. Un hipotético escenario sería el de un médico que necesitase determinadas informaciones sobre un tema en particular, bien para realizar una investigación, para recopilar información sobre la enfermedad algún paciente, etc. En este caso, un analizador discursivo automático le proporcionaría la posibilidad de detectar el tipo de información deseada en diversos textos de su interés, como, por ejemplo, los resultados de diversos estudios sobre el cáncer de mama. El sistema de extracción podría tomar como entrada la salida de un analizador discursivo, para extraer los satélites de *Resultado* que proporciona la RST y conformar así un compendio de resultados sobre un mismo tema, sin necesidad de leer todos los textos y extraer la información de manera manual. En la actualidad estamos trabajando en esta aplicación y en otras relacionadas también con la extracción de información.

4.3. Evaluación automática de resumen automático

Hoy en día, la evaluación de resúmenes automáticos es un tema que suscita gran interés. Inicialmente, los métodos de evaluación de resúmenes fueron manuales (véanse los trabajos de Morris et al., 1992; Maybury, 1995; Mani et al., 1998; Saggion y Lapalme, 2000, entre otros). Posteriormente, se desarrollaron sistemas semiautomáticos, como ROUGE (Lin, 2004), *Pyramid Method* (Nenkova y Passonneau, 2004) o *Basic Elements* (Hovy et al., 2005). Sin embargo, estos sistemas semiautomáticos siguen requiriendo una gran inversión, ya que todos ellos necesitan resúmenes humanos para emplear como modelo en la comparación con los resúmenes automáticos. Actualmente, se está abriendo una nueva línea de investigación que aboga por sistemas de evaluación completamente automáticos (Louis y Nenkova, 2009; Torres-Moreno et al., 2010).

En este sentido, consideramos que el análisis discursivo automático podría ser útil para desarrollar un sistema de evaluación automática de resúmenes. La idea principal podría explicarse por medio de la "metáfora del bonsái": un bonsái de pino se parece a un pino en miniatura; un bonsái de fresno se parece a un fresno. El objetivo es crear un bonsái por medio de un proceso de poda. Esto puede lograrse si podamos un árbol (texto fuente), preservando las ramas principales que dan forma a dicho árbol, sin las ramas innecesarias (resumen). Así, se podría presuponer que un texto y su resumen deberían tener una estructura discursiva similar, aunque uno de los dos textos sea más reducido que el otro. En la actualidad estamos realizando experimentos para corroborar esta hipótesis y diseñar un sistema de evaluación automática de resúmenes automáticos que explote esta idea.

4.4. Traducción automática

Hoy en día, la mayor parte de los trabajos sobre traducción automática siguen estrategias estadísticas (ejs. Koehn et al., 2007; Zens y Ney, 2007; Koehn, 2009). Aunque las traducciones suelen ser aceptables (siempre con una corrección manual posterior), distan mucho de traducciones humanas. Para optimizar los resultados de estos sistemas de traducción, una vía posible podría ser la utilización de la estructura discursiva, en concreto mediante la RST. Una traducción ideal (texto meta) debería reflejar la misma estructura discursiva que su texto de origen (texto fuente). No obstante, esto no ocurre en una gran cantidad de ocasiones. El hecho de contar con un analizador discursivo que proporcione automáticamente la estructura discursiva del texto fuente será de gran utilidad para realizar la traducción en base a esa estructura, permitiendo que el texto meta refleje la misma estructura discursiva y, por lo tanto, comunique más adecuadamente su contenido, haciéndolo más claro para el lector. Actualmente estamos llevando a cabo experimentos en este sentido.

En esta línea existen muy pocos trabajos. Ghorbel et al. (2001), por ejemplo, emplean la RST para alinear fragmentos de textos en distintas lenguas, mientras que Marcu et al. (2000) realizan traducciones japonés-inglés a partir de árboles discursivos basados en la RST.

4. Conclusiones

Mediante este artículo creemos haber demostrado la pertinencia del análisis discursivo automático, ya sea como una aplicación en sí misma o como una herramienta para el desarrollo de otras aplicaciones lingüísticas, como resumen automático, traducción automática, extracción de información o evaluación automática de resúmenes. Hemos mostrado algunos trabajos relevantes que tratan sobre estos temas, además de haber presentado algunas líneas de trabajo futuro novedosas y con gran potencial.

Así pues, consideramos necesaria la finalización de nuestro proyecto interuniversitario para el desarrollo de un sistema de análisis discursivo para el español (ADAe). Además, para poder desarrollar las aplicaciones mencionadas en otras lenguas, es indispensable realizar la adaptación lingüística de este analizador. La plataforma multilingüe DiZer está diseñada precisamente para incluir analizadores de lenguas diversas y

consideramos que el desarrollo de analizadores discursivos automáticos para lenguas de características similares es factible mediante la utilización de la metodología empleada en el proyecto ADAe.

Referencias

Brown, P.F.; Cocke, J.; Pietra, S.A.D.; Pietra, V.J.D.; Jelinek, F.; Lafferty, J.D.; Mercer, R.L.; Roossin, P.S. (1990). "A statistical approach to machine translation". *Computational linguistics* 16(2). 79-85.

Corston-Oliver, S. (1998). "Beyond string matching and cue phrases: Improving efficiency and coverage in discourse analysis". En *Proceedings of AAAI 1998 Spring Symposium Series, Intelligent Text Summarization.* 9-15. Madison.

da Cunha, I.; SanJuan, E.; Torres-Moreno, J-M.; Lloberas, M.; Castellón, I. (en prensa). "DiSeg: un analizador discursivo automático para el español". *Procesamiento del Lenguaje Natural*.

da Cunha, I.; Torres-Moreno, J-M.; Velázquez, P.; Vivaldi, J. (2009). "Un algoritmo lingüístico-estadístico para resumen automático de textos especializados". *Linguamática* 2. 67-79.

da Cunha, I. (2008). *Hacia un modelo lingüístico de resumen automático de artículos médicos en español*. Barcelona: Institut Universitari de Lingüística Aplicada. [CD-ROM] (Sèrie Tesis; 23)

Ghorbel, H.; Ballim, A.; Coray, G. (2001). "ROSETTA: Rhetorical and Semantic Environment for Text Alignment". En Rayson, P.; Wilson, A.; McEnery, A.; Hardie, A.; Khoja, S. (eds). *Proceedings of Corpus Linguistics*. 224-233. Lancaster.

Haouam, K.; Marir, F. (2003). "SEMIR: Semantic indexing and retrieving web document using Rhetorical Structure Theory". *Intelligent Data Engineering and Automated Learning* 2690. 596-604.

Hovy, E.; Lin, C. Y.; Zhou, L. (2005). "Evaluating DUC 2005 using basic elements". En *Proceedings of the Document Understanding Conferences*. Vancouver. 1-6.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; Dyer, C.; Bojar, O.; Constantin, A.; Herbst, E. (2007). "Moses: Open Source Toolkit for Statistical Machine Translation". En *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Praga.

Lin, C. Y. (2004). "Rouge: A Package for Automatic Evaluation of Summaries". En *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*. 25-26.

Louis, A.; Nenkova, A. (2009). "Automatically Evaluating Content Selection in Summarization without Human Models". En *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapur. 306-314.

Mani, I.; House, D.; Klein, G.; Hirschman, L.; Obrst, L.; Firmin, T.; Chrzanowski, M.; Sundheim, B. (1998). *The Tipster Summac Text Summarization Evaluation: Final report*. Technical report. DARPA.

Mann, W. C.; Thompson, S. A. (1988). "Rhetorical structure theory: Toward a functional theory of text organization". *Text* 8(3). 243-281.

Marcu, D. (2000a). *The Theory and Practice of Discourse Parsing Summarization*. Massachusetts: Institute of Technology.

Marcu, D. (2000b). "The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach". *Computational Linguistics* 26(3). 395-448.

Marcu, M.; Carlson, L.; Watanabe, M. (2000). "The automatic translation of discourse structures". En *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*. Vol. 1. 9-17. Seattle.

Maybury, M. (1995). "Generating Summaries from event Data". *Information Processing and Management* 31(5). 735-751.

Morris, A.H.; Kasper, G.M.; Adams, D.A. (1992). "The effects and limitations of automated text condensing on reading compression performance". *Information Systems Research* 2(1). 17-35.

Nenkova, A.; Passonneau, R. (2004). "Evaluating content selection in summarization: The pyramid method". En *Proceedings of the HLT-NAACL Conference*. Boston. 145-152.

Ono, K.; Sumita, K.; Miike, S. (1994). "Abstract generation based on rhetorical structure extraction". En *Proceedings of the International Conference on Computational Linguistics*. Kyoto. 344-348.

Pardo, T.A.S.; Rino, L.H.M. (2001). "A summary planner based on a three-level discourse model". En *Proceedings of Natural Language Processing Pacific Rim Symposium.* Tokyo. 533-538.

Pardo, T.A.S.; Nunes, M.G.V.; Rino, L.H.M. (2008). "DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese". *Lecture Notes in Artificial Intelligence* 3171. 224-234.

Saggion, H.; Lapalme, G. (2000). "Concept identification and presentation in the context of technical text summarization". En *Proceedings of the ANLP/NAACL Workshop on Automatic Summarization*. Seattle.

Shinmori, A.; Okumura, M.; Marukawa, Y.; Iwayama, M. (2002). "Rhetorical structure analysis of Japanese patent claims using cue phrases". En *Proceedings of the 3rd NTCIR Workshop*. Tokyo.

Sumita, K.; Ono, K.; Chino, T.; Ukita, T.; Amano, S. (1992). “A discourse structure analyzer for Japanese text”. En *Proceedings of the International Conference on Fifth Generation Computer Systems.* Tokyo. 1133-1140.

Taboada, T.; Mann, W.C. (2005). “Applications of rhetorical structure theory”. *Discourse Studies* 8(4). 567-588.

Teufel, S.; Moens, M. (2002). “Summarizing scientific articles: Experiments with relevance and rhetorical structure”. *Computational Linguistics* 28(4). 409-445.

Tofiloski, M; Brooke, J; Taboada, M. (2009). “A Syntactic and Lexical-Based Discourse Segmenter”. En *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*. Singapur.

Torres-Moreno, J-M; Saggion, H.; da Cunha, I.; Velázquez-Morales, P.; SanJuan, E. (2010). “Évaluation automatique de résumés avec et sans référence”. En *Actes de la 17e Conférence sur le TALN*. Université de Montréal et École Polytechnique de Montréal.

Zens, R.; Ney, H. (2007). “Efficient Phrase-table Representation for Machine Translation with Applications to Online MT and Speech Translation”. En *Proceedings of the NACLCL*. Rochester.