Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 151-154 recibido 21-03-2018 revisado 19-04-2018 aceptado 03-05-2018

Proyecto TAGFACT: Del texto al conocimiento. Factualidad y grados de certeza en español

TAGFACT Project: From text to knowledge. Factuality and degrees of certainty in Spanish

Laura Alonso[1], Irene Castellón[2], Hortensia Curell[3], Ana Fernández-Montraveta[3], Sonia Oliver[3], Gloria Vázquez[4]

[1]Universidad de la República, [2]Universitat de Barcelona,
[3]Universitat Autònoma de Barcelona, [4]Universitat de Lleida
[1]lauraalonsoalemany@gmail.com, [2]icastellon@ub.edu,
[3]{hortensia.curell, ana.fernandez, sonia.oliver}@uab.cat, [4]gvazquez@dal.udl.cat

Resumen: El objetivo general de este proyecto es crear una herramienta para la anotación de la factualidad expresada en textos en español a través del procesamiento automático. Pretendemos que dicha representación sea muy rica, por lo que se llevará a cabo desde tres ejes distintos: multinivel, multidimensional y multitextual. El análisis multinivel da cuenta de las distintas marcas lingüísticas que expresan el grado de certeza de un evento a nivel morfológico y sintáctico, pero también discursivo; el análisis multidimensional, de un número variado de las voces que evalúan dicho evento; y el análisis multitextual, de distintos textos sobre un mismo evento, siendo este último uno de los aspectos más innovadores de la propuesta.
Palabras clave: factualidad, procesamiento semántico, implicación textual, presuposición, modalidad, polaridad, certeza, asertividad.

Abstract: The main aim of this project is to create a tool for the automatic annotation of factuality-expressed in Spanish texts. We intend the representation to be very rich, so it will be carried out from three different perspectives: multilevel, multidimensional and multitextual. The multilevel analysis gives account of the various linguistic markers that express the degree of certainty of an event at the morphological and syntactic, as well as discourse, levels. The multidimensional analysis accounts for the varied number of voices that assess an event. Finally, the multitextual analysis will take into account the various texts about the same event, which is one of the most innovative aspects of our proposal.
Keywords: factuality, semantic processing, textual implication, presupposition, modality, polarity, certainty, assertivity.

1 Introducción

En los textos se predica sobre eventos y situaciones (a partir de ahora, eventos). La asignación de valores de factualidad (certeza) a los eventos es un campo que tiene un interés muy destacado en el ámbito de la lingüística de corpus y en el PLN. La factualidad de un evento no es una propiedad intrínseca de este sino que siempre está en relación al modo en que es presentado dicho evento, teniendo en cuenta que puede ser evaluado por una o más personas y en el mismo momento o en momentos distintos.

El objetivo general del proyecto que presentamos (TAGFACT), que está en la fase inicial de desarrollo, es crear una herramienta para anotar automáticamente la factualidad de los eventos expresados en textos en español. Pretendemos desarrollar un sistema de anotación en el que se tengan en cuenta tres ejes distintos: multinivel, multidimensional y multitextual.

En este sentido, nuestra propuesta tiene de innovadora que es globalizadora, ya que pretendemos construir un sistema automático capaz de aportar las diversas interpretaciones factuales de los eventos a partir de las distintas marcas de factualidad que se pueden identificar en los textos, abarcando desde la morfología, la sintaxis y el discurso (análisis multinivel), pero también un abanico de fuentes muy

ISSN 1135-5948. DOI 10.26342/2018-61-20

diversificado (análisis multidimensional), que van más allá de las que se encuentran en los límites de la frase y del propio texto, ya que se tienen en cuenta también distintos textos que tratan sobre el mismo evento global (análisis multitextual).

Además, queremos investigar las ventajas que puede aportar crear una herramienta que use exclusivamente conocimiento lingüístico en el caso del español, ya que para esta lengua se ha trabajado muy poco en este ámbito y básicamente usando conocimiento estadístico.

2 *La factualidad en la lingüística de corpus y el PLN*

Uno de los sistemas automáticos pioneros en la anotación de la factualidad son el de Light et al. (2004) y el de Medlock y Briscoe (2007). FactBank (Sauri, 2008) supuso una propuesta innovadora para la representación de información factual para el inglés. Es un referente en este ámbito, ya que el tratamiento que hace de los diferentes fenómenos lingüísticos relacionados con la factualidad es muy amplio. Contiene 8.837 eventos (208 documentos) etiquetados manualmente en función de distintas fuentes. Cabe decir que un aspecto que no se considera en el modelo de anotación propuesto es la interpretación factual del evento según el momento temporal. Tampoco distingue los eventos futuros y condicionales del resto y no queda claro cómo se evalúan en términos de factualidad los estados atemporales ni los eventos habituales.

En el marco de su tesis la autora también implementa un anotador de factualidad (De Facto). Aunque esta herramienta contiene un algoritmo para etiquetar automáticamente la factualidad, algunos de los módulos de conocimiento utilizados han sido creados manualmente, por lo que no se puede considerar que la etiquetación sea un proceso automático. Los resultados de este sistema respecto a cobertura y precisión, así como de la F-measure son buenos aunque no tiene en cuenta algunas construcciones que aportan información de factualidad, como las condicionales o las temporales, entre otras.

Aproximadamente a partir de la constitución de FactBank y De Facto, la comunidad de PLN consolida su interés sobre la factualidad. Son diversas las contribuciones que se han ido realizando en la última década en este ámbito. Asimismo, también aparecen trabajos en los que se avanza en el campo de la anotación de la modalidad (Hendrickx et al., 2012; Rupenhofer y Rehbein, 2012; Morante y Daleemans, 2012).

Velupillai (2011) trabaja con corpus del dominio de la medicina clínica y aporta una herramienta de etiquetación automática de eventos respecto a la polaridad y la certeza. El campo de la biomedicina es uno en los que más se ha trabajado en relación a la anotación de la factualidad, donde destaca Vincze et al. (2008) y Nawaz et al. (2010). Otros sistemas basados en el aprendizaje supervisado en este ámbito son los de Farkas et al. (2010), Tang et al. (2010) y Morante et al. (2010). Como herramientas construidas con conocimiento lingüístico mencionamos Harkema et al. (2009) y Wu et al. (2009).

Otros trabajos bastante recientes basados en Sauri (2008), Van Son et al. (2014) y Prabhakaran et al. (2015). Respecto a este último, el corpus anotado manualmente es de aproximadamente un millón de palabras (es el más grande que existe para el inglés), del cual un 10% aproximadamente se usa para evaluar distintas herramientas automáticas. En otros trabajos, contrariamente al marco usado en FactBank y De Facto, se opta por considerar la factualidad ligada al conocimiento del mundo (Marneffe et al., 2012).

Para el español existen algunos corpus etiquetados con información temporal. Pero en el campo de la factualidad solo conocemos el trabajo de Wonsever et al. (2016). Estos autores adoptan el modelo de Sauri (2008), con algunos cambios, y crean un corpus anotado con información sobre factualidad y una herramienta de anotación automática basada en aprendizaje automático supervisado.

Si se pretende llevar a cabo un tratamiento global de la factualidad, es necesario también identificar las equivalencias entre eventos dentro del mismo texto y entre textos distintos. Este campo está muy poco explorado, pero existe una línea en la que se trabaja la similitud textual (Agirre et al., 2014) que puede aportar luz en el tema. También los sistemas de identificación de correferencias anafóricas son importantes para tratar las correferencias de eventos. Para el español destacamos Björkelund y Kuhn (2014) y Durret y Klein (2013).

3 *Metodología*

Para llevar a cabo el proyecto TAGFACT el primer paso es estudiar la expresión de la factualidad en español a diferentes niveles lingüísticos (morfológico, sintáctico y

discursivo). Para ello, en este momento, estamos confeccionando un corpus sobre noticias relacionadas con la política extraído de diarios españoles de distintas tendencias. Dicho corpus se anotará manualmente por diversos miembros del proyecto y se evaluará el *agreement* con las medidas oportunas. Así obtendremos un Gold Standard que servirá como referente para evaluar la anotación automática.

En lenguas como el español, hay diversos marcadores lingüísticos de la factualidad como algunos elementos y el tiempo verbal, ya que, inicialmente el futuro y el condicional no expresan hechos. Además hay determinadas construcciones que afectan a la interpretación de la factualidad y un subconjunto de predicados que subcategorizan eventos y que proyectan información factual sobre estos (Palmer, 1986; Quirk et al., 1985). Por ejemplo, en "María *ha conseguido* pasar de curso" el verbo *conseguir* claramente proyecta la certeza del aprobado de María.

Otro elemento clave en la expresión de la factualidad es la modalidad, que ha sido categorizada en diversos tipos, entre los cuales la que tiene un efecto claro respecto a la factualidad es la epistémica. Las tres categorías que se suelen utilizar para describir la modalidad epistémica asociada a las proposiciones son: certeza, probabilidad y posibilidad. La polaridad negativa completaría la escala del continuum con los no hechos (Sauri, 2008; Repiso, 2015).

Los distintos marcadores de factualidad de las categorías mencionadas pueden interactuar entre sí en la descripción de un mismo evento. Uno de los retos del proyecto en esta fase de estudio es analizar estas combinaciones y dar cuenta del resultado que se obtiene en relación con los valores factuales.

Los desafíos más destacables en la elaboración de la herramienta giran en torno a tres ejes: en primer lugar, la desambiguación de los posibles valores de algunos elementos léxicos y construcciones; en segundo lugar, el establecimiento de correferencias entre eventos, más allá de la frase, dentro de una misma noticia y entre noticias distintas; y, en tercer lugar, la resolución automática de las distintas interpretaciones que se van construyendo en las oraciones que incorporan distintas fuentes (discurso directo e indirecto) y también en los casos de sintaxis altamente compleja a nivel de subordinación.

4 Conclusiones

En TAGFACT pretendemos aportar una herramienta que represente la información diferenciando los hechos reales de aquellas expresiones que presentan creencias, opiniones o posibilidades y constituir un corpus de referencia en la anotación de la factualidad.

En este sentido, la detección del grado de certeza de los eventos sobre los que se predica en los textos y su representación en un lenguaje formal se considera vital para poder realizar inferencias con el objetivo de adquirir nuevo conocimiento y actualizar y crear ontologías y bases de conocimiento, basadas en hechos, de diferentes ámbitos. Hay que tener en cuenta que muchas aplicaciones de diversos ámbitos utilizan ontologías, y el mantenimiento de estas ontologías es un proceso muy costoso. La automatización de esta tarea permite ahorrar esfuerzos en el desarrollo del conocimiento necesario para cualquier aplicación automática que trabaje con lenguaje. Así, el recurso creado tendrá gran utilidad en otras aplicaciones como el análisis de opiniones (*sentiment analysis*), la extracción de información textual, los sistemas de pregunta-respuesta, en las que la interpretación del grado de factualidad es una tarea primordial.

Finalmente, otro valor añadido del proyecto es que el anotador automático que se creará, al estar basado en conocimiento lingüístico, será más estable entre diferentes dominios que los sistemas basados en aprendizaje automático, que son más dependientes del dominio de desarrollo. Esta característica dotará a la herramienta de más potencialidad para procesar de forma automática grandes volúmenes de textos y de gran diversidad.

Agradecimientos

Este proyecto está siendo desarrollado por miembros del grupo GRIAL (http://grial.uab.es) y está financiado por el Ministerio de Economía, Industria y Competitividad - FFI2017-84008-P.

Bibliografía

Agirre, E., C. Baneab, C. Cardiec, D. Cerd, M. Diabe, A. González-Agirre, W. Guof, R. Mihalceab, G. Rigau y J. Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. *Proceedings of the 8th International Workshop on Semantic Evaluation*, 81-91.

Björkelund, A. y J. Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-

local features. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 47-57.

Durret, G. y D. Klein. 2013. Easy victories and uphill battles in coreference resolution. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1971-1982.

Farka, R., V. Vincze, G. Móra, J. Csirik y G. Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. *Proceedings of the 14th CoNLL*, 1-12.

Harkema, H., J. N. Dowling, T. Thornblade y W. W. Chapman. 2009. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42: 839-851.

Hendrickx, I., A. Mendes y S. Mencarelli. 2012. Modality in text: a proposal for corpus annotation. *LREC 2012, Eighth International Conference on Language Resources and Evaluation*, 1805-1812.

Light, M., X. Ying Qiu y P. Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. En L. Hirschman y J. Pustejovsky (Eds.), *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases, 17-24.*

Marneffe, M. C., C. D. Manning y C. Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics* 38-2:301-333.

Medlock, B. y T. Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. *Proceedings of the 45th ACL*, 992-999.

Morante, R., V. Van Asch and W. Daelemans. 2010. Extraction of biomedical events. *LOT Occasional Series*, 16. 91-105.

Morante, R. y W. Daelemans. 2012. Annotating Modality and Negation for a Machine Reading Evaluation. *CLEF 2012 Evaluation Labs and Workshop Online Working Notes.* Disponible en: http://www.clef-initiative.eu/documents/71612/463956e9-2b22-4e68-aa39-b711302c97b1

Nawaz, R., P. Thompson y S. Ananiadou. 2010. Evaluating a meta-knowledge annotation scheme for bio-events. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, 69-77.

Palmer, F. R. 2004. *Mood and Modality*. Cambridge: Cambridge University Press.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Repiso, I. 2015. Talking about counterfactual worlds. A comparative study of French and Spanish. *Journal of Romance Studies* 15-1, 52-72.

Rupenhofer, J. e I. Rehbein. 2012. Annotating the senses of the English modal verbs. *LREC 2012, Eighth International Conference on Language Resources and Evaluation*, 1538-1545.

Sauri, R. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. Thesis. Brandeis University.

Tang, B., X. Wang, X. Wang, B. Yuan y S. Fan. 2010. A cascade method for detecting hedges and their scope in natural language text. In *Proceedings 14th CoNLL*, 13-17.

Van Son, C., M. van Erp, A. Fokkens y P. Vossen. 2014. Hope and Fear: Interpreting Perspectives by Integrating Sentiment and Event Factuality. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, 26-31.

Velupillai, S. 2011. Automatic classification of factuality levels. A case study on Swedish diagnoses and the impact of local context. En A. Moen, S. K. Andersen, J. Aarts y P. Hurlen (Eds.) *Proceedings of the XXIII International Conference of the European Federation for Medical Informatics. User Centred Networked Health Care*, Oslo: IOS Press, 559-563.

Vincze, V., G. Szarvas, R. Farkas, G. Móra y J. Csirik. 2008. The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9, 38-45.

Wonsever, D., A. Rosá y M. Malcuori. 2016. Factuality annotation and learning in Spanish texts. *Proceedings of Language Resources and Evaluation*, 2076-2080.

Wu, A. S., B. H. Do, J. Kim y D. Rubin. 2009. Evaluation of negation and uncertainty detection and its impact on precision and recall in search. *Journal of Digital Imaging*, 24(2): 234-242.