Correferencias: resolución, discursos fragmentados y captura de eventos

Lucía Cantamutto
Universidad Nacional del Sur
Bahía Blanca, Argentina
luciacantamutto@gmail.com

Josu Bermúdez
DeustoTech-INTERNET
Deusto Institute of Technology
Universidad de Deusto
josu.bermudez@deusto.es

Joseba Abaitua
DELi - LinguaMedia
Universidad de Deusto
joseba.abaitua@deusto.es

David Buján
DeustoTech-INTERNET
Deusto Institute of Technology
Universidad de Deusto
david.bujan@deusto.es

JosuKa Díaz-Labrador
DELi - LinguaMedia
Universidad de Deusto
josuka@deusto.es

http://www.deli.deusto.es/
http://linguamedia.deusto.es/
http://www.morelab.deusto.es/labman/

Resumen

Se presenta el doble trabajo doctoral en marcha relacionado con la resolución de correferencias: uno es la adaptación de los algoritmos conocidos al español, y otro la aplicación a un corpus experimental de textos breves (mensajería y tuits). Además de ello, se presenta una hipotética alimentación de la resolución de correferencias a recursos semánticos conocidos como la DBpedia, *Linked Open Data*, o *Simple Event Model*. El objetivo sería capturar "eventos" a partir de textos: estos eventos podrían convertirse en noticias en proyectos turísticos, o en recursos específicos de índole cultural, histórico, antropológico…

1. Introducción

Este resumen gira alrededor de la *correferencia*, concepto que motiva los proyectos doctorales de los dos primeros autores. Los objetivos son variados:

- Adaptar y mejorar algoritmos como la *multi-pass sieve* [Raghunathan et al., 2010] y otras herramientas de análisis al español, para obtener herramientas de recuperación de información o minería de datos como el proyecto OpeNER de Agerri et al. [2013]; se trata del proyecto doctoral de Bermúdez [2013].
- Profundizar en el conocimiento de la correferencia en español, analizando además de corpus conocidos, un corpus experimental de textos breves (de mensajería y tuits) anotado manualmente; es el proyecto doctoral de Cantamutto [Cantamutto et al., 2014].

In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org

- Aplicar y evaluar herramientas que incluyen la resolución de correferencias como OpeNER a la captura de "eventos", tal como son definidos en el *Simple Event Model* de van Hage et al. [2011]. Procesos similares ya se han abordado de manera experimental en proyectos como *tourExp* [Buján et al., 2013], sobre aplicaciones turísticas.

- La información resultante sería relevante para organizaciones de índole cultural, histórico, como EuskoMedia, Wikipedia, topHistoria, etc.

- Finalmente, pero no menos importante, contribuimos a proyectos como DBpedia y *Linked Open Data*.

2. Resolución de correferencias

En el reconocimiento de entidades con nombre (antropónimos, organizaciones, topónimos políticos o físicos, títulos, expresiones numéricas fecha-tiempo, y otras como medidas, direcciones de correo, direcciones web, etc.) la correferencia y la anáfora son problemas conocidos [Hirst, 1981]. Sin embargo, hay ciertas diferencias.

En la anáfora, los elementos anafóricos siempre dependen de un antecedente en el texto: su significado no es pleno, requiere necesariamente de una mención anterior. La correferencia ocurre en el plano pragmático: la relación depende del contexto comunicativo y situacional, ocurre entre dos unidades lingüísticas (plenas o anafóricas) que se relacionan porque tienen una "identidad en la referencia" [Recasens and Vila, 2010], es decir, el mismo referente en el discurso.

Por tanto, la correferencia, a diferencia de la anáfora, no es una relación unidireccional y asimétrica, sino simétrica y transitiva [Recasens, 2008]. La resolución de la anáfora es nombre-pronombre, mientras que la resolución de la correferencia ha de obtener cadenas de elementos que tienen idéntico referente.

3. Captura de eventos

La aplicación propuesta de la resolución de correferencias a la captura de eventos puede mostrarse con el siguiente ejemplo desarrollado de forma manual a partir del pasaje de Besga Marroquín [2007] recogido en la fig. 1.

> "Cuando el Imperio Romano de Occidente desapareció en el 476, el reino visigodo, que se extendía a los dos lados de los Pirineos, era el reino germánico más grande. Pese a la fama que se ha dado a los visigodos como aliados de Roma, con ningún otro pueblo luchó tanto tiempo el imperio en su último siglo de existencia en Occidente, ni ningún otro le arrebató tanto territorio. Así, en el 476 el reino visigodo, con capital en Tolosa, se extendía desde el Loira hasta una zona indeterminada de la mitad meridional de la península ibérica (no se puede precisar más porque se desconoce la cronología de la ocupación visigoda de gran parte de la Península). Nadie tenía entonces más territorios en Francia y en la península ibérica. Además, uno de los grandes reyes visigodos, Eurico (466-484) aprovechó la desaparición del Imperio Romano de Occidente para extender aún más sus dominios. Efectivamente el reino visigodo completó entonces la ocupación de toda la costa mediterránea francesa, una vieja aspiración que había sido combatida por los romanos."

Figura 1: Texto original de Besga Marroquín [2007]

El objetivo intermedio sería la identificación de entidades de la DBpedia, por ejemplo `http://es.dbpedia.org/page/Imperio_Romano_de_Occidente` en el caso de la primera de las entidades de la fig. 4, y lo mismo con todas las demás.

1. Cuando el Imperio Romano de Occidente desapareció en el 476, el reino visigodo era el reino germánico más grande.
2. En el 476 el reino visigodo se extendía por Galia e Hispania a ambos lados de los Pirineos.
3. Pese a la fama de los visigodos como aliados de Roma, contra ningún otro pueblo luchó PRO tanto en su último siglo de dominio en Occidente.
4. Nadie había arrebatado al Imperio tanto territorio como el pueblo visigodo.
5. El reino con capital en Tolosa se extendía desde el Loira hasta una zona indeterminada de la mitad meridional de la península ibérica.
6. El rey visigodo Eurico (466-484) aprovechó la desaparición del Imperio de Occidente para extender sus dominios.
7. Eurico completó la ocupación de toda la costa mediterránea francesa, una vieja aspiración que había sido combatida por Roma.

Figura 2: Identificación de entidades y correferencias

Con ello, podría enriquecerse el texto de la fig. 2 en la forma que se ve en la fig. 5, de manera que pudiera integrarse en proyectos como *Linked Open Data* o *Simple Event Model*, entre otros, y pudiera aprovecharse la información semántica en aplicaciones como las mencionadas.

M1 Imperio Romano de Occidente, Roma, PRO, su, Occidente, Imperio, Imperio de Occidente, Roma
M2 476, 476
M3 reino visigodo, reino visigodo, visigodos, pueblo visigodo, reino con capital en Tolosa
M4 reino germánico
M5 Galia
M6 Hispania
M7 Pirineos
M8 Loira
M9 península ibérica
M10 Eurico, sus, Eurico
M11 costa mediterránea francesa

Figura 3: Listado de entidades y correferencias

M1 wiki-es:Imperio_Romano_de_Occidente
M2 wiki-es:476
M3 wiki-es:Reino_visigodo
M4 wiki-es:Reinos_germánicos
M5 wiki-es:Galia
M6 wiki-es:Hispania
M7 wiki-es:Pirineos
M8 wiki-es:Loira
M9 wiki-es:Península_Ibérica
M10 wiki-es:Eurico
M11 wiki-es:Costa_Azul_(Francia)

Figura 4: Identificación de entidades de DBpedia

1. Cuando el M1 desapareció en el M2, el M3 era el M4 más grande.
2. En el M2 el M3 se extendía por M5 e M6 a ambos lados de los M7.
3. Pese a la fama de los M3 como aliados de M1, contra ningún otro pueblo luchó M1 tanto en su último siglo de dominio en M1.
4. Nadie había arrebatado al M1 tanto territorio como el M3.
5. El M3 se extendía desde el M8 hasta una zona indeterminada de la mitad meridional de la M9.
6. El rey visigodo M10 (466-484) aprovechó la desaparición del M1 para extender sus dominios.
7. M10 completó la ocupación de toda la M11, una vieja aspiración que había sido combatida por M1.

Figura 5: Texto enriquecido con entidades

Referencias

Rodrigo Agerri, Montse Cuadros, Sean Gaines, and German Rigau. OpeNER: Open Polarity Enhanced Named Entity Recognition. *Procesamiento del Lenguaje Natural*, 51:215–218, 2013. ISSN 1135-5948. URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4891.

Josu Bermúdez. Reconocimiento conjunto de entidades nombradas y de correferencia para mejorar el acceso a la información multilingüe. Informe de seguimiento de tesis doctoral, 2013.

Armando Besga Marroquín. La batalla de Vouillé. *Historia 16*, (380):10–31, 2007.

David Buján, David Martín, Ortzi Torices, Diego López-de Ipiña, Carlos Lamsfus, Joseba Abaitua, and Aurkene Alzua-Sorzabal. Context Management Platform for Tourism Applications. *Sensors*, 13(7):8060–8078, June 2013. ISSN 1424-8220. doi: 10.3390/s130708060. URL http://www.mdpi.com/1424-8220/13/7/8060.

Lucía Cantamutto, Josu Bermúdez, Joseba Abaitua, Rodrigo Agerri, David Buján, and Josuka Díaz-Labrador. Resolución de correferencias en discursos fragmentados para la captura de eventos. In *XLIII Simposio Internacional de la Sociedad Española de Lingüística. Resúmenes de las comunicaciones*, pages 154–155. Sociedad Española de Lingüística, 2014. URL http://www.sel.edu.es/sites/default/files/Libro%20de%20res%C3%BAmenes%20definitvo%20%2810%20enero%29.pdf.

Grahame Hirst. *Anaphora in Natural Language Understanding*. Springer Verlag, 1981.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A Multi-Pass Sieve for Coreference Resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics, 2010.

Marta Recasens. Towards Coreference Resolution for Catalan and Spanish. Trabajo presentado como requisito parcial para la obtención del DEA, 2008. URL http://clic.ub.edu/sites/default/files/users/dea-recasens.pdf.

Marta Recasens and Marta Vila. On Paraphrase and Coreference. *Computational Linguistics*, 36(4):639–647, 2010. URL http://www.mitpressjournals.org/doi/pdfplus/10.1162/coli_a_00014.

Willem Robert van Hage, Véronique Malaisé, Roxane H Segers, Laura Hollink, and Guus Schreiber. Design and Use of the Simple Event Model (SEM). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2), 2011. ISSN 1570-8268. URL http://www.websemanticsjournal.org/index.php/ps/article/view/190.