Revista Pilquen - Sección Ciencias Sociales
ISSN: 1666-0579
revista.pilquen@gmail.com
Universidad Nacional del Comahue
Argentina

Cantamutto, Lucía; Abaitua, Joseba; Buján, David; Díaz Labrador, Jesús Luis; Bermúdez, Josu
RESOLUCIÓN DE CORREFERENCIAS PARA LA CAPTURA DE EVENTOS
Revista Pilquen - Sección Ciencias Sociales, vol. 18, núm. 2, 2015, pp. 40-49
Universidad Nacional del Comahue
Viedma, Argentina

REVISTA Pilquen. Sección Ciencias Sociales • Vol. 18 Nº 2 • 2015
ISSN 1851-3123 • http://www.revistapilquen.com.ar/

RESOLUCIÓN DE CORREFERENCIAS PARA LA CAPTURA DE EVENTOS

Por Lucía Cantamutto, Joseba Abaitua**, David Buján**, Jesús Luis Díaz Labrador**, Josu Bermúdez***
luciacantamutto@uns.edu.ar - joseba.abaitua@deusto.es - david.bujan@deusto.es - josuka@deusto.es - josu.bermudez@deusto.es
*Universidad Nacional del Sur; CONICET. Argentina - **Universidad de Deusto-DeustoTech. España

RESUMEN

El algoritmo *Stanford Multi Sieve Pass* (propuesto por Raghunathan et al. 2010) realiza secuencialmente una serie de pasos de reconocimiento que de manera incremental terminan proponiendo correferencias entre las entidades candidatas identificadas en el texto. En este artículo, presentamos brevemente los trabajos de adaptación de este algoritmo y de otras herramientas de análisis (p. e., OpeNER) a textos en español (Agerri et al. 2013; Bermúdez 2013). A fin de avanzar en el desarrollo de estas herramientas para el Procesamiento del Lenguaje Natural (PLN), aplicamos estas directrices manualmente sobre un corpus experimental extraído de Wikipedia, con los que se pueden configurar textos breves (como por ejemplo tuits u otro tipo de microcontenidos) con sentido pleno. Como resultado, se ofrece un ejemplo de análisis manual que será automatizado en etapas posteriores de la investigación.

Palabras clave: Correferencia; Captura de eventos; Algoritmos de resolución, Simple Event Model; Datos enlazados.

COREFERENCE RESOLUTION FOR CAPTURE OF EVENTS

ABSTRACT

The algorithm (proposed by Raghunathan et al. 2010) sequentially performs a series of pass of recognition, and allows to go incrementally proposing candidates to coreferenced between named entities in the text. The article briefly presents the work of adapting the algorithm Stanford Multi Pass Sieve and other analysis tools (OpeNer) to texts in Spanish (and other Agerri 2013, Bermúdez 2013). The result is a fragmented speech in sentences with full sense, that even being independent of the speech have not lost the discursive framework they belong (to inherit metadata documentaries). This can feed the event-based knowledge systems, or be linked to deposits of open data, or published independently (vg. as tweets). As a result, the article offers an example of manual analysis that, in further research, will be automatic.

Key words: Coreference; Event capture; Resolution algorithm; Simple Event Model; Linked open data.

Recibido: 15|01|15 • **Aceptado:** 01|06|15

ISSN 1851-3123 • http://www.revistapilquen.com.ar/

1. INTRODUCCIÓN[1]

Algoritmos recientes de asignación de referencias plenas (*Named-Entity Recognition* o NER) y resolución de correferencias en el tratamiento automático de textos para la extracción de información están consiguiendo resultados muy satisfactorios aplicados a textos en inglés (Raghunathan et al. 2010; Lee et al. 2011). Sin embargo, para el español, el volumen de datos es relativamente menor y está aún en fase de prueba (véase Gamallo et al. 2014). A partir de los resultados obtenidos sobre el corpus AnCora-CO (utilizado para entrenar la adaptación del algoritmo al español), el conocimiento sobre la correferencia en español se ha enriquecido notablemente (Recasens 2008; Recasens y Hovy 2010; Recasens y Vila 2010) así como también el estudio de la anáfora (Palomar et al., 2001).

La evaluación de los resultados del algoritmo *Stanford Multi Sieve Pass* aplicados al español así como la comprobación de la metodología de anotación manual de correferencias de Recasens y Martí (2010) representan un avance en el conocimiento de las relaciones de correferencia para el análisis del discurso en el plano interoracional. En ese contexto, nuestros objetivos son variados. Primero, definir y caracterizar los fenómenos de correferencia y anáfora en el español en el orden del plano pragmático-discursivo. La correcta identificación de los elementos que producen una cadena de significados sobre la misma *named entity,* "entidad nombrada" (NE) favorece la evaluación de las herramientas de recuperación de información. Por tanto, en textos en español, se reconstruirán de manera manual los procesos de minería de datos a través de instrumentos de análisis tales como OpeNER (Agerri et al. 2013, Bermúdez 2013). Por último, se busca reconocer entidades y resolver correferencias para la captura de eventos en textos culturales (Buján et al. 2013).

2. CLASIFICACIÓN DE ENTIDADES NOMBRADAS (*NAMED ENTITIES*)

Los discursos, a través de los distintos mecanismos de cohesión, utilizan recursos propios de la lengua en su construcción, que resultan en la paráfrasis, las expresiones anafóricas y las cadenas de correferencia. En el marco de las actividades del procesamiento del lenguaje natural (NLP, por su sigla en inglés), el reconocimiento de los referentes directos de estos procesos potencia la consecución satisfactoria de la recuperación y extracción de información (Recasens y Vila 2010).

Antes de comenzar a explicar cómo se resuelven satisfactoriamente las correferencias, una definición del núcleo al que remiten esas cadenas de significado son las *entidades con nombre* (o nombradas u onomásticas, *Named Entities* o NE, es decir, autorreferenciales). Según explica Martínez Rodríguez (2009:110) son "términos que corresponden a nombres propios que identifican una entidad del mundo real y que no se encuentran en los diccionarios como tales". Sin embargo, mayor claridad aporta la definir las NE como aquellos elementos del mundo real a los que se alude en el interior de un texto.

Se han configurado múltiples tipologías de NE a partir de las conferencias MUC[2], que fueron el punto de encuentro para la evaluación y comparación de diferentes herramientas de extracción de información. En 1995 se utilizó por primera vez el término durante la MUC-6:

> En esta conferencia, a parte [sic] de la evaluación de los sistemas habitual, se abarcó la importancia del reconocimiento de la información. Los sustantivos que se refieren a entidades individuales (nombres propios de persona, nombres de organizaciones y nombres locativos entre otros) junto con las expresiones numéricas (fechas, tiempo, porcentajes, dinero)

[1]Algunos fragmentos de este trabajo fueron presentados en las Jornadas TIMM (http://timm.ujaen.es/wp-content/uploads/2014/06/timm2014_submission_4.pdf).
[2]*Message Understanding Conferences* (http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html)

ISSN 1851-3123 • http://www.revistapilquen.com.ar/

> destacaban en esta área. La extracción de dichos elementos (denominados Named Entities) fue entonces reconocida como una de las subtareas más importantes en la identificación de entidades de un texto (ibíd.: 2012).

En este mismo foro se presentó el primer corpus anotado con resolución de correferencias, convirtiéndose, a partir de MUC-6, en una tarea de creciente interés en PLN (Recases y Hovy 2010).

Siguiendo la clasificación propuesta en esa ocasión, las categorías principales que representan las NE se pueden agrupar en (Martínez Rodríguez 2009:12):

1. Antropónimos (nombres de persona)

2. Organizaciones

3. Topónimos (políticos o físicos)

4. Expresiones numéricas fecha-tiempo

5. Otras NE (medidas -porcentajes, monetarias, pesos-, direcciones de correo, direcciones Web, etc.).

Además de estas cinco, Taulé et al. (2008) señalan que los títulos de libros también son NE. Por lo tanto, es posible extrapolar a títulos de otro tipo de artefactos culturales como títulos de películas, canciones. Asimismo, sugerimos que los nombres de acontecimientos también puede constituir NE (i.e. *La batalla de San Lorenzo*). Una revisión de diferentes clasificaciones de NE se puede leer en Nadeau y Sekine (2007).

Sin embargo, los textos no utilizan siempre las mismas formas para referirse a algo. Es, por tanto, una de las tareas de los reconocedores de NE (NER): detectar y desambiguar las entidades presentes en un texto (Nadeau y Sekine, 2007). En este proceso de detección de menciones, es necesario trabajar en diferentes niveles de lengua: morfológico, sintáctico, semántico y textual (véase Raghunathan et al. 2010:492-493); y en distintos ítems: pronombres, diferentes frases nominales, sustantivos comunes y NE (tal como sugiere Recasens 2008:4).

En las páginas siguientes se describirá un proyecto en marcha que busca adaptar la herramienta OpeNER para el español. En primer lugar, el problema se aborda desde la lingüística: es esta la disciplina que puede proveer y alimentar los recursos de las herramientas de PLN[3]. Luego, se muestran los avances en torno a la captura de eventos en textos culturales (textos que narran episodios históricos, por ejemplo), realizados de forma manual. Los eventos permiten explicar la relación entre personas, lugares, acciones y objetos de manera simple (van Hage et al. 2011:128). El objetivo de este trabajo es contribuir con el desarrollo de la web semántica a partir del enriquecimiento de textos en español. De esta forma, los segmentos obtenidos pueden alimentar sistemas de conocimiento basados en eventos, o bien ser enlazados desde depósitos de datos abiertos, o publicados de forma autónoma (v.g. en forma de tuits). Son procesos que de manera experimental han sido abordados en los proyectos TourExp, BiDEI y NeLH (Buján et al. 2013).

3. CORREFERENCIA, ANÁFORA Y PARÁFRASIS: DELIMITAR FENÓMENOS SUPERPUESTOS

Identificar y resolver las correferencias en los textos a través de las herramientas de extracción de información es un paso inexcusable para la detección de unidades lingüísticas que tengan "identidad en la referencia" (Recasens 2002; Recasens y Vila 2010). En caso contrario, "la tarea de extracción proporcionaría entidades diferentes cuando en realidad se trata de la misma entidad" (Muñoz et al. 1999). En el proceso de reconocimiento de NE se distinguen tres fenómenos distintos que presuponen cierta confusión: 1) la anáfora, 2) la paráfrasis y 3) la correfencia.

[3]Raghunathan et al. (2010:492) exponen los problemas típicos de este tipo de tareas computacionales cuando no incluyen conocimiento de tipo lingüístico y refieren a otros trabajos previos. En la introducción muestran un texto en el que el pronombre 'we' y el sintagma nominal 'the israelis' no deben ser puestos en correlación: en el contexto se habla de una entidad geopolítica Israel, que recibe un rasgo de 'inanimado' que evitará su conexión con 'we' 'animado' (íbid.).

ISSN 1851-3123 • http://www.revistapilquen.com.ar/

1) Anáfora

Dentro de los diferentes tipos de deixis que ocurren al interior de un texto, hay palabras que recogen su significado a partir de una parte del discurso que le antecede (cfr. DRAE). De esta manera, en los casos de anáfora la relación de interdependencia es textual. Al requerir necesariamente de una mención anterior, los elementos anafóricos no tienen un significado pleno y dependen siempre de un antecedente en el texto. En general, "los miembros de las clases de los pronombres funcionan como anáforas" (Di Tullio 2010:168). Es así que la resolución de la anáfora se da a partir de la identificación de la cadena nombre-pronombre. En la organización del discurso, la anáfora es la estrategia para referir de manera abreviada a alguna entidad dentro del discurso, con la expectativa de que el lector sea capaz de detectar la referencia[4] (Hirst 1981:4).

2) Paráfrasis

Las expresiones textuales que tienen un significado similar o una equivalencia conceptual relativa pero estructuras distintas son denominadas paráfrasis. Como recurso, la paráfrasis se ubica en el plano del significado y, preferentemente, en el nivel de la oración y la frase, ya que suele estar compuesta por varias unidades que se extralimitan del lexema. La paráfrasis ocurre entre unidades lingüísticas con significado propio; es decir, a diferencia de la anáfora y la correferencia, no sucede entre pronombres.

Sin embargo, dentro de la actual orientación pragmático-discursiva de los estudios lingüísticos, es improbable considerar que dos expresiones tengan el "mismo significado". Es por ello que Bhagat (2009: xv) retoma, en cambio, la noción de *quasi-paraphrases*[5], evitando, así, la difícil identidad entre estructuras distintas. La superposición entre correferencia y paráfrasis ha sido objeto de atención para su comprensión en el PLN (Recasens y Vila 2010; Vila, Martí y Rodríguez 2011). Aunque operen en dos dimensiones distintas, la extracción de paráfrasis suele solaparse con la resolución de correferencias. Como indican Recasens y Vila (2010:641), la divergencia estriba en que dos expresiones son parafrásticas cuando tienen similar significado mientras que cuando dos o más expresiones en un discurso refieren a la misma entidad se identifican como correferentes.

3) Correferencia

Ocurre entre dos unidades lingüísticas (plenas o anafóricas) que se relacionan porque tienen una "identidad en la referencia"; es decir, el mismo referente en el mundo real: esto no solo ocurre en el discurso (Recasens y Vila 2010) sino también en la representación mental que hace el oyente/lector (Recasens 2008:3). Por lo tanto, la relación entre las dos unidades correferentes depende del contexto comunicativo y situacional. A diferencia de la anáfora, la correferencia entre dos entidades establece una relación simétrica y transitiva, en la cual si A tiene una relación de correferencia con B y B la tiene con C, entonces, C es correferente con A (ibíd.: 4). De esta manera, la correferencia se resuelve estableciendo cadenas de elementos que se identifican con el mismo referente. Como se ha señalado, este recurso lingüístico juega un papel fundamental en términos de cohesión textual (Recasens y Vila 2010:643).

En resumen, el conocimiento sobre la correferencia en español se ha enriquecido notablemente gracias al desarrollo de herramientas para el PLN (Muñoz et al. 1999; Recasens 2008; Recasens y Hovy 2010; Recasens y Vila 2010). Los autores mencionados diferencian la correferencia de la anáfora: si bien ambas dan cuenta de relaciones entre elementos dentro del discurso, la anáfora es una relación léxico-gramatical en la que un elemento (generalmente un pronombre) depende de su antecedente, favoreciendo la fluidez del discurso (Recasens 2008). En tal sentido, la correferencia no es una relación unidireccional y asimétrica, sino simétrica y transitiva. La resolución de correferencias es el proceso de "encadenar todas las unidades lingüísticas (menciones) que refieren a la misma entidad en el discurso" [6] (Recasens 2008:4). Estas cadenas de

[4]"Anaphora is the device of making in discourse an abbreviated reference to some entity (or entities) in the expectation that the perceiver of the discourse will be able to disabbreviate the reference and thereby determine the identity of the entity. The reference is called ANAPHOR, and the entity to which refers is the REFERENT or ANTECEDENT. A reference and its referent are said CORREFERENTIAL. The process of determining the referent of an anaphor is called RESOLUTION". (Hirst, 1981:4). Revisado por Recasens (2008:2)

[5]Coinciden, en este sentido, Recasens y Vila (2010) y Vila, Martí y Rodríguez (2011).

[6]"Coreference resolution was thus born as the process of linking in a string all those linguistic units (mentions) that refer to the same entity in the discourse model". La traducción es nuestra.

ISSN 1851-3123 • http://www.revistapilquen.com.ar/

correferencia se dan entre elementos disímiles del discurso (frases nominales, pronombres, nombres comunes) que se refieren a una misma entidad.

4. EL RENDIMIENTO DE *MULTI-PASS SIEVE*

Decidir si dos menciones en el interior de un texto se corresponden a través de una única función dentro de un modelo o algoritmo de análisis es una de las tendencias predominantes aunque resulte problemático, de todas maneras, ya que la información local no es suficiente para poder identificar las cadenas de significado (Raghunathan et al. 2010). Ante ello, el modelo *Multi-Pass Sieve,* superando los algoritmos anteriores, utiliza una sucesión de tamices que van enriqueciéndose en las etapas subsiguientes con información de las anteriores: a través de este modelo, se parte de los reconocimientos con mayor nivel de precisión (cotejo exacto) para ir perfeccionando la agrupación de NE (*clustering*) y las cadenas de significado atribuidas a esa NE. Así, a través de siete pasos (y sin entrenamiento previo de la herramienta), "se garantiza que cada decisión utiliza toda la información disponible en cada momento todo el tiempo"[7] (ibíd.: 493). En general, se utilizan pasos que recorren el texto de izquierda a derecha (Recasens 2008:6; Raghunathan et al. 2010:494).

La metáfora del tamiz (*sieve* en inglés) grafica con claridad el funcionamiento de la herramienta: a partir de una determinada mención, el filtro resuelve si es incapaz de determinar el antecedente, o si la identificación queda pendiente para los siguientes pasos, o, por el contrario, reconoce la NE. En esta herramienta se trabaja la correferencia al interior de un texto. Basado en el analizador sintáctico *Stanford Parser*, son siete los módulos por los cuales pasa el texto para examinar las menciones (Raghunathan et al. 2010):

1. Cotejo exacto: dos construcciones iguales, se resuelven como correferentes. Por ejemplo: *El Imperio Romano* y *El Imperio Romano.* Deben coincidir incluso modificadores y determinantes.

2. Construcciones previsibles: para encadenar dos menciones se requiere que se satisfagan alguna de las siguientes condiciones: a) una de las menciones está en una construcción de aposiciones (i. e. al gobierno del [primer emperador], [César Augusto]); b) las menciones están en una oración copulativa (i. e. [Tiberio] era [hijo de Livia]); c) el candidato tiene un sustantivo que funge como modificador (i. e. [el [emperador] César Augusto])[8]; d) la mención es un pronombre relativo que modifica el núcleo de la frase nominal antecedente (i. e. [Roma], [que] extendió su control en torno al mar Mediterráneo); e) cuando una de las menciones es un acrónimo de la otra (i. e. [Senatus Populusque Romanus]; [SPQR][9]); f) una de las menciones es un gentilicio de la otra (i. e. [Imperio Romano], [los romanos]).

3. Cotejo estricto de núcleos: añade restricciones a núcleos idénticos a partir de cotejar si efectivamente coincide en el sintagma (i.e. [el Imperio Romano] [el Imperio]). Este paso mantiene alta la precisión (91%) mejorando la cobertura (entre 6-8 puntos).

4 y 5. Son variantes de 3.

6. Cotejo laxo de núcleos: Utiliza conjuntos de candidatos a antecedente y solo se aplica a NE. Su incidencia es mínima en la mejora (1 punto).

7. Pronombres: en los seis pasos previos, el modelo ignoró la correferencia pronominal. En esta etapa, al igual que la anterior, la herramienta está preparada por los pases previos que han ido creando listas de candidatos para la resolución de las correferencias pronominales. Se realiza un cotejo de concordancia: género, número, persona, animacidad, etiqueta NER. A través de este cotejo, sube la cobertura un 22% aunque baja la precisión un 8% (véase **tabla 1**).

En la **tabla 1** se observa como en cada paso aumenta acumulativamente la eficacia y cobertura de la herramienta puesta a prueba. La P corresponde a la precisión del modelo en enlazar

[7]La traducción es nuestra. En el original: "This ensures that each decision uses all for the information available at the time".
[8]Veáse Di Tullio (2010:155-157).
[9]Para la constitución de acrónimos, como procedimiento morfológico, véase Casado Velarde (1999: 5082) y para los procesos de lexicalización de las siglas y acrónimos véase De Miguel (2009:80).

ISSN 1851-3123 • http://www.revistapilquen.com.ar/

correctamente las menciones. Con R se mide el *recall* o cobertura a través del cual se determina el porcentaje de elementos con respecto al total que el modelo fue capaz de recuperar. Por último, con F1 se observa en qué medida se combinó la precisión y la cobertura. A medida que el modelo realiza los pasos, disminuye paulatinamente la precisión pero aumenta considerablemente la cobertura y el F1.

Tabla 1: Extraída de Raghunathan (2010:497)

	MUC			B^3			Pairwise		
Passes	P	R	F1	P	R	F1	P	R	F1
{1}	95.9	31.8	47.8	99.1	53.4	69.4	96.9	15.4	26.6
{1,2}	95.4	43.7	59.9	98.5	58.4	73.3	95.7	20.6	33.8
{1,2,3}	92.1	51.3	65.9	96.7	62.9	76.3	91.5	26.8	41.5
{1,2,3,4}	91.7	51.9	66.3	96.5	63.5	76.6	91.4	27.8	42.7
{1,2,3,4,5}	91.1	52.6	66.7	96.1	63.9	76.7	90.3	28.4	43.2
{1,2,3,4,5,6}	89.5	53.6	67.1	95.3	64.5	76.9	88.8	29.2	43.9
{1,2,3,4,5,6,7}	83.7	74.1	78.6	88.1	74.2	80.5	80.1	51.0	62.3

5. Las cadenas de significado en la captura de eventos

De lo que se trata, entonces, es de "agrupar en un único clúster todas las expresiones que hacen referencia a las diferentes entidades" (Martínez Rodríguez 2009:16). A través de una serie de pasos manuales, se detectan las menciones presentes en un texto cultural a fin de ilustrar lo expuesto. Partimos del siguiente texto[10]:

> 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. (Besga Marroquín, 2007).

En una primera etapa de segmentación manual obtenemos las frases plenas que fueron seccionadas intentando que no superasen los 140 caracteres, número máximo aceptado por la red social Twitter. Los tuits se enlistan como referentes de los "microcontenidos" (Lindner, 2006): entendidos como pequeños fragmentos que dan cuenta de un contenido mayor[11]. El texto, segmentado, es el siguiente:

1. Cuando el Imperio Romano de Occidente desapareció en el 476, el reino visigodo era el reino germánico más extenso.

[10]Recomendamos la visita del sitio Besga2007/entities, donde se encuentra la transcripción extensa del artículo con su correspondiente enlazado realizado por el Dr. Joseba Abaitua. Disponible en http://wiki.littera.deusto.es/es/index.php/Besga2007/entities. A partir del texto inicial de Besga Marroquín, fueron extraídas 100 entidades y 260 eventos (véase http://wiki.littera.deusto.es/es/index.php/Besga2007events). Las entidades fueron clasificadas a partir de la ontología provista por el proyecto *Dbpedia* (http://mappings.dbpedia.org/server/ontology/classes/).

[11]En su primera definición, Lindner (2006: 41) indica "This new Web has to be conceptualized not only as a technological and educational infrastructure, but as a complex and dynamic ecosystem based on microcontent: very small pieces, loosely joined, permanently rearranging to form volatile (micro-) knowledge clouds, and making necessary new forms of microlearning". Otros autores (De Juan et al. (2012) sugieren que los microcontenidos son el resumen (¿la condensación?) de los macrocontenidos. Por nuestra parte, consideramos que el microcontenido puede ser una "píldora" independiente.

ISSN 1851-3123 • http://www.revistapilquen.com.ar/

2. En el 476 el reino visigodo se extendía por la 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 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.

El siguiente paso, en este modelo, fue el reconocimiento de menciones candidatas a ser NE, que suele estar encabezada por la detección de frases nominales. A partir de la anotación manual del texto anterior, se encontraron once menciones. En función del enriquecimiento de textos para la web semántica, una manera de considerarlas es como entradas de Wikipedia, que al estar enlazadas con el proyecto Dbpedia[12], pueden ser consideradas fácilmente como NE. Esto ofrece el siguiente panorama (Tabla 2):

Tabla 2: Menciones como entradas de Wikipedia

M_1	Imperio Romano de Occidente, Roma, PRO, su, Occidente, Imperio, Imperio de Occidente, Roma [Wikipedia:es:Imperio romano de Occidente]
M_2	476, 476 [Wikipedia:es:476]
M_3	reino visigodo, reino visigodo, visigodos, pueblo visigodo, reino con capital en Tolosa [Wikipedia:es:reino visigodo]
M_4	reino germánico [Wikipedia:es:reinos germánicos]
M_5	Galia [Wikipedia:es:Galia]
M_6	Hispania [Wikipedia:es:Hispania]
M_7	Pirineos [Wikipedia:es:Pirineos]
M_8	Loira [Wikipedia:es:río Loira]
M_9	península ibérica [Wikipedia:es:península ibérica]
M_{10}	Eurico, sus, Eurico [Wikipedia:es:Eurico]
M_{11}	costa mediterránea francesa [Wikipedia:es:Costa Azul (Francia)]

Rápidamente se observa que las unidades lingüísticas seleccionadas no corresponden al mismo nivel de lengua y que su detección requirió, en algunos casos, la complementación de la correferencia con recursos anafóricos. El resultado siguiente es un texto en el cual las unidades lingüísticas que referencian a los mismos elementos en el interior del discurso fueron reemplazadas por sus equivalentes en las cadenas de correferencia (Tabla 3).

[12]DBpedia es un proyecto para la extracción de datos de Wikipedia al fin de nutrir una versión Web semántica. Es llevado a cabo por la Universidad de Leipzig, Universidad Libre de Berlín y la compañía OpenLink Software. Véase http://es.wikipedia.org/wiki/DBpedia (consulta: julio de 2014).

ISSN 1851-3123 • http://www.revistapilquen.com.ar/

Tabla 3: Texto con cadenas de correferencia

Cuando M_1 desapareció en M_2, M_3 era el M_4 más extenso
En M_2 M_3 se extendía por M_5 e M_6 a ambos lados de M7
Pese a la fama de M3 como aliados de M1, contra ningún otro pueblo luchó M1 tanto en su último siglo de dominio en M1
Nadie había arrebatado a M1 tanto territorio como M3
M3 se extendía desde M8 hasta una zona indeterminada de la mitad meridional de M9
El rey visigodo M10 (466-484) aprovechó la desaparición de M1 para extender sus dominios
M10 completó la ocupación de toda M11, una vieja aspiración que había sido combatida por M1

Como se observa en el texto analizado, en lenguaje natural, y principalmente en los textos escritos, se utilizan diferentes unidades lingüísticas para referir a una misma entidad de la realidad. De esto trata la cohesión textual. En el proceso de recuperación de información, el procedimiento inverso debe llevarse a cabo: todas las expresiones utilizadas para referir a lo mismo se enlazan en una misma mención, en una NE, es decir, a un mismo referente.

6. PALABRAS FINALES

A través de las páginas expuestas, hemos presentado una breve aproximación a las estrategias para la resolución de correferencia para la captura de eventos y el enriquecimiento de la web semántica. Sin intentar hacer un análisis exhaustivo del fenómeno, se verifica con claridad que el aporte de los lingüistas es fundamental para el desarrollo de herramientas de NER y en la correcta identificación de los procesos cohesivos que ocurren al interior de los textos.

Tras la presentación del modelo *Multi-pass Sieve* para la resolución de correferencias (Raghunathan et al. 2010) se evidencia que el trabajo debe ser realizado de manera conjunta a través de todo los niveles de lengua y de manera tal que la herramienta se vaya enriqueciendo a sí misma como es el caso de los módulo del tamiz o a través de modelos tipo tuberías (Agerri, Bermúdez y Rigau 2014). En el futuro nos plantemos mejorar las herramientas de resolución de correferencia aplicadas al español y ponerlas a prueba en textos culturales.

ISSN 1851-3123 • http://www.revistapilquen.com.ar/

BIBILIOGRAFÍA

1. Agerri, Rodrigo, Bermúdez, Josu y Rigau, German. "Multilingual, Efficient and easy NLP processing with IXA pipeline", *Demo Sessions of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014).* Gothenburg: Sweden. 2014.

2. Agerri, Rodrigo et al. "OpeNER: open polarity enhanced named entity recognition." *Procesamiento del Lenguaje Natural,* 51. [Sociedad Española para el procesamiento del lenguaje]. España. 2013, pp. 216-218

3. Bermúdez, Josu. "Reconocimiento conjunto de entidades nombradas y de correferencia para mejorar el acceso a la información multilingüe". Informe de tesis doctoral. Bilbao: Universidad de Deusto. 2013.

4. Besga Marroquín, Armando. "La Batalla de Vouillé", *Historia* 16, Nº 380. 2007, pp. 10-31.

5. Bhagat, Rahul. *Learning parapharases from text*. Doctoral Dissertation. South California: University of Southern California. Disponible en http://digitallibrary.usc.edu/cdm/ref/collection/p15799coll127/id/189848 (consulta: julio de 2014). 2009.

6. Buján, David et al. "Context Management Platform for Tourism Applications". *Multidisciplinary Digital Publishing Institute*, 13, 9. 2013, pp. 8060-8078.

7. Casado Velarde, Manuel. "Otros procesos morfológicos: acortamientos, formación de siglas y acrónimos", en Bosque Muñoz, I. y Demonte Barreto, Violeta. *Gramática descriptiva de la Lengua Española. Vol. III.,* Madrid: Espasa. 1999. pp. 5075/5096.

8. De Miguel, Esther. *Panorama de la lexicología.* Barcelona: Ariel. 2009.

9. Di Tullio, Ángela. *Manual de gramática del español*, Buenos Aires: Waldhuter. 2010.

10. Gamallo, Pedro et al. "Análisis morfosintáctico y clasificación de entidades nombradas en un entorno Big Data", *Procesamiento del Lenguaje Natural*, 53, 2014, pp. 17-24.

11. Hirst, Graeme J. *Anaphora in Natural Language Understanding: A Survey.* Berlin: Springer-Verlag. 1981

12. Lee, Heeyoung et al. "Stanford's multi-pass sieve correference resolution system at the CoNLL-2011 shared task". En *Proceedings of the Fifteenth Conference on Computational Natural Language Learning:* Shared Task. 2011, pp. 28-34.

13. Lee, Heeyoung et al. "Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules". *Computational Linguistics* (January 3): pp. 1–54. doi:10.1162/COLI_a_00152. 2013.

14. Martínez Rodríguez, Jesús. *Sistema de clustering de Named Entities. Tesis de Maestría*, Barcelona: Universidad Politécnica de Catalunya. 2009.

15. Muñoz, Rafael, Martínez-Barco, Patricio & Ferrández, Antonio "Método para la resolución de correferencias de sintagmas nominales definidos incluyendo alias y acrónimos en el sistema de extracción de información EXIT". *Procesamiento del Lenguaje Natural*, 25. 1999, pp. 143-149.

16. Nadeau, David; Sekine, Satoshi. A survey of named entity recognition and classification. *Lingvisticae Investigationes*. 2007, vol. 30, no 1, p. 3-26.

17. Palomar, Manuel, et al. An algorithm for anaphora resolution in Spanish texts. *Computational Linguistics* 27.4. 2001. pp. 545-567.

18. Raghunathan, Karthik et al. "A Multi-pass Sieve for Coreference Resolution". *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. 2010, pp. 492–501.

19. Recasens, Marta. *Towards coreference resolution for Catalan and Spanish*. Diss. Master Thesis. University of Barcelona. 2008.

20. Recasens, Marta y Eduard Hovy. "Coreference Resolution across Corpora: Languages, Coding Schemes, and Preprocessing Information". *Proceedings of ACL* Uppsala, Suecia. 2010, pp. 1423-1432.

ISSN 1851-3123 • http://www.revistapilquen.com.ar/

21. Recasens, Marta y María Antònia Martí, "AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan". *Language Resources and Evaluation*, 44(4). 2010, pp. 315-345.

22. Recasens, Marta y Marta Vila. "On Paraphrase and Coreference". *Computational Linguistics*, 36 (4). 2010, pp. 639-647.

23. Recasens, Marta, Eduard Hovy, and M. Antònia Martí. "Identity, non-identity, and near-identity: Addressing the complexity of coreference". *Lingua*, 121 (6). 2011, pp. 1138-1152.

24. Recasens, Marta et al. "SemEval-2010 Task 1: Coreference Resolution in Multiple Languages". *Proceedings of the ACL Workshop on Semantic Evaluations SemEval.* 2010.

25. Taulé, Mariona, M. Antònia Matí y Marta Recasens. "AnCora: Multilevel Annotated Corpora for Catalan and Spanish", *Proceedings of 6th International Conference on Language Resources and Evaluation*. Marrakesh (Morocco). 2008.

26. Van Hage, Willem et al. "Design and use of the Simple Event Model (SEM)", *Web Semantics: Science, Services and Agents on the World Wide Web*, 9. 2011, pp. 128-136.

27. Vila, Marta; Martí, M. Antònia; Rodríguez, Horacio. "Paraphrase Concept and Typology. A Linguistically Based and Computationally Oriented Approach". *Procesamiento del Lenguaje Natural*, 2011, no 46. 2011, pp. 83-90.

Créditos

28. DRAE: www.rae.es

29. OpenNER (Agerri et al. 2013): http://www.opener-project.org/

NeHL, BiDEI, TourExp (Buján et al. 2013): http://linguamedia.deusto.es/ y http://morelab.deusto.es/

30. Simple Event Model (van Hage et al., 2011) http://www.websemanticsjournal.org/index.php/ps/article/view/190/188

31. Willem van Hage http://wrvh.home.xs4all.nl/wrvhage/

32. Bpedia http://dbpedia.org/About

33. *WordNet: http://wordnetweb.princeton.edu/perl/webwn*