Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 57-64 recibido 05-05-12 revisado 31-05-12 aceptado 05-06-12

Representación Gráfica de Documentos para Extracción Automática de Relaciones *

Graph-based Document Representation for Relation Extraction

Bernardo Cabaleiro
NLP&IR Research Group
ETSI Informática, UNED, Spain
bcabaleiro@lsi.uned.es

Anselmo Peñas
NLP&IR Research Group
ETSI Informática, UNED, Spain
anselmo@lsi.uned.es

Resumen: Este artículo presenta un sistema de representación de documentos orientado a la compactación, integración y simplificación de información. El sistema genera grafos a nivel de documento a partir de árboles de dependencias sintácticas haciendo explícita la semántica de algunas aristas. El objetivo es crear una representación útil para múltiples tareas de procesamiento de lenguaje natural, entre ellas la extracción automática de relaciones, para la que realizamos una evaluación extrínseca cuantitativa.
Palabras clave: Representación de documentos, grafos semánticos, extracción de relaciones

Abstract: This paper presents a document representation system oriented to compactation, integration and simplification of information. This system generates document-level graphs from syntactic dependency trees making explicit the semantics of some edges. The goal is to create a representation useful for multiple tasks of natural language processing, as relation extraction. For this task we perform a quantitative evaluation.
Keywords: Document representation, semantic graphs, relation extraction

1. Introducción

Este artículo presenta un sistema de representación de documentos orientado a la compactación, integración y simplificación de información con el fin de avanzar el estado del arte en tareas de procesamiento de lenguaje como resolución de correferencias, extracción de información o búsqueda de respuestas.

La hipótesis de partida es que una representación de documentos completos en forma de grafo, simplificando las relaciones morfosintácticas y añadiendo información semántica, beneficia a sistemas orientados a precisión en tareas cómo la extracción automática de relaciones.

Esta representación a nivel de documento se crea a partir del árbol de dependencias, sobre el que se realizan las siguientes operaciones: Colapsado de correferencias, asignación de clases semánticas a entidades nombradas y normalización de determinadas estructuras sintácticas.

La correferencia es una relación lingüística que se establece entre dos o más expresiones que refieren a una misma entidad, sean o no isomorfas. En este artículo, al conjunto de todas estas expresiones para una entidad lo denominamos *referente de discurso* (Karttunen, 1968). Al proceso de crear un referente de discurso agrupando las menciones a medida que aparecen en un discurso lo denominaremos *colapsado*. Este procedimiento es similar a las tarjetas archivo (*file card*) empleado por Heim (1983) o las cestas (*basket*) en Recasens (2010).

Por otra parte, asignaremos clases semánticas a las entidades nombradas mediante reglas que detecten determinadas estructuras con genitivos, compuestos nominales y aposiciones.

Por último, normalizaremos distintas expresiones que no tengan diferencia semántica. Consideramos la voz gramatical, genitivos expresados de diferentes maneras, compuestos nominales, etc.

Las preguntas de investigación que aborda este trabajo son: (1) ¿Qué efecto tiene la representación gráfica a nivel de documento

* This work has been partially supported by the Spanish Ministry of Science and Innovation, through the project Holopedia (TIN2010-21128-C02).

 © 2012 Sociedad Española para el Procesamiento del Lenguaje Natural

en una tarea de extracción automática de relaciones? (2) En esta misma tarea, una vez generados los grafos, ¿supone una mejora enriquecerlos con información semántica?

El objetivo que se persigue es proporcionar una base para las técnicas de extracción de información. Si esta base es apropiada, los clasificadores dispondrán de nuevos rasgos que permitirán la captura de contextos distantes en el texto. El impacto de esta técnica se ha evaluado en las tareas Regular Slot Filling y Temporal Slot Filling correspondientes a la tarea competitiva Knowledge Base Population (Ji, Grishman, y Dang, 2011).

El artículo está estructurado de la siguiente manera: Comenzamos detallando la representación en la sección 2, en la sección 3 evaluamos el efecto de la representación propuesta en la tareas, continuamos en la sección 4 comentando brevemente otros trabajos relacionados y terminamos con las conclusiones en la sección 5 y el trabajo futuro en la sección 6.

2. Propuesta de representación

Diferenciamos dos tipos de representaciones, según la etapa de procesamiento en el que se encuentren, para poder evaluar si la agregación de información es útil para los clasificadores:

- Configuración inicial: En esta configuración están disponibles las anotaciones de los nodos y las aristas con información sintáctica, las correferencias y las relaciones temporales.
- Configuración colapsada: En esta configuración se añaden las aristas con información semántica, se normalizan las aristas sintácticas y se realiza el colapsado, por lo que desaparecen las aristas de correferencia y se emplean los referentes de discurso.

2.1. Configuración inicial

En la configuración inicial, cada documento D está representado por un grafo, G_D, con un conjunto de nodos N_D y un conjunto de aristas A_D. Cada nodo representa una unidad de información, generalmente una palabra, excepto en dos casos: una entidad nombrada de más de una palabra, o un verbo y sus auxiliares.

Los nodos están etiquetados con una serie de atributos, algunos de los cuales son comunes para todos los nodos: palabras que contiene, lemas asociados, etiquetas morfosintácticas y una cadena de caracteres representativa que denominamos *descriptor*. Esta cadena se genera de la siguiente manera: Para los nodos que no son entidades, se compone con los lemas de las palabras. En el caso de las entidades, se escoge la cadena de caracteres tal y como se encuentra en el texto.

Además, hay nodos anotados con más propiedades. Se dividen en tres categorías:

- Eventos: Se corresponden con verbos que describen una acción. Están anotados con el tiempo, el aspecto y la polaridad.
- Expresiones temporales: Identifican palabras o conjuntos de palabras que indican un instante concreto o un periodo de tiempo. Contienen un valor temporal normalizado según el estándar TimeX3.
- Entidades nombradas: Entidades reconocidas en el texto. Se anotan con el tipo de entidad, por ejemplo, organización, persona, lugar, etc. En caso de ser persona también se etiqueta el género y la edad.

Las aristas representan cuatro tipo de relaciones entre los nodos:

- Sintáctica: Indica que existe una relación sintáctica entre dos nodos. Se corresponden con las etiquetas sintácticas del Penn Treebank.
- Coreferencia: Indica que dos nodos son menciones de un mismo referente de discurso.
- Semántica: Indica que existe una relación semántica entre dos nodos. Distinguimos cuatro etiquetas semánticas subespecificadas: is, has, hasClass y hasProperty.
- Temporal: Indica que existe una relación temporal entre un evento y una expresión temporal. Las relaciones pueden ser de los tipos: *before*, *after*, *within*, *throughout*, *beginning* y *ending*.

2.2. Configuración colapsada

El grafo con configuración inicial G_D se transforma para crear el grafo con la configuración colapsada G_C. Cada grupo de nodos relacionados por correferencias $n_0, \ldots, n_i, \ldots, n_k \in N_D$ se agrupan en

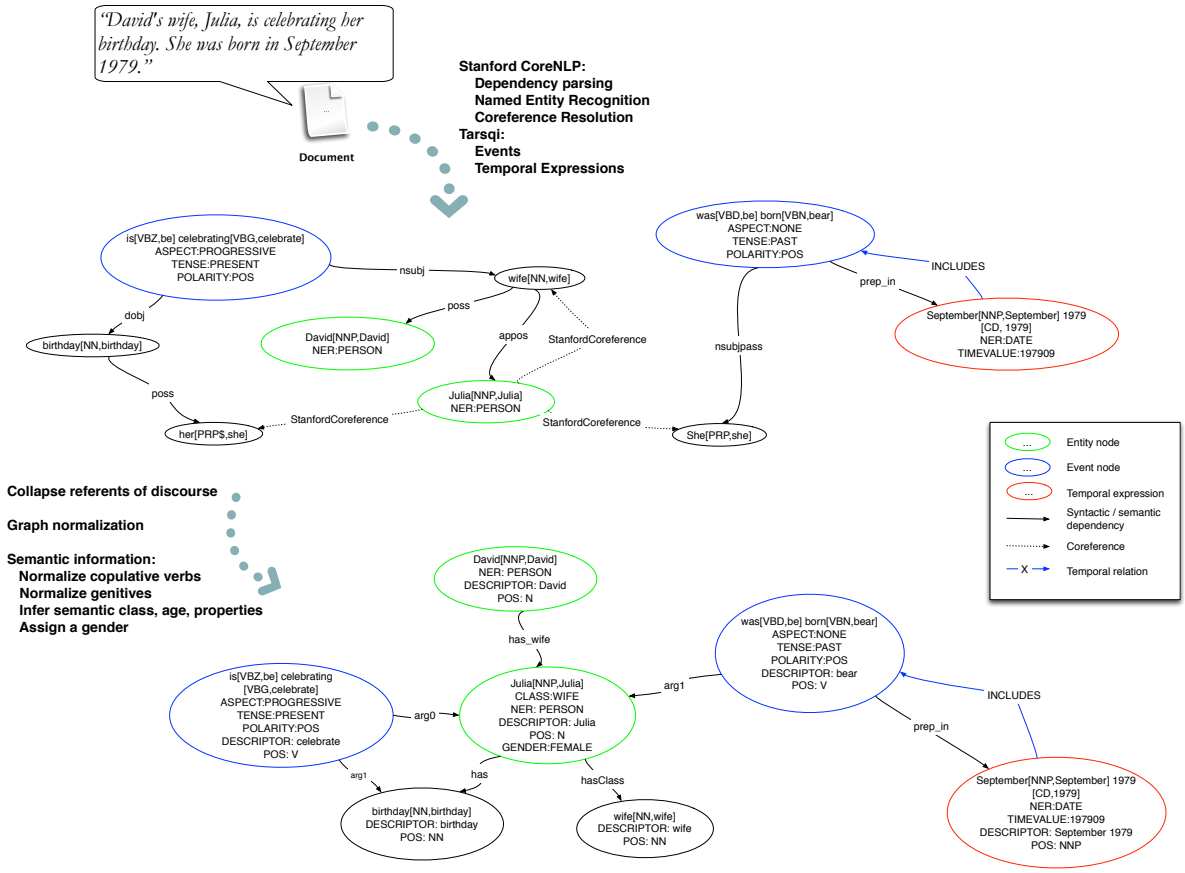

Figura 1: Transformación del grafo inicial, G_D, al colapsado, G_C para el documento de ejemplo: "David's wife, Julia, is celebrating her birthday. She was born in September 1979".

un referente de discurso $r = \cup\, n_0, \ldots, n_k$, creando así un nuevo conjunto de referentes de discurso R_C. Dado otro referente de discurso $r' = \cup\, n'_0, \ldots, n'_j, \ldots, n'_k$, unimos ambos referentes con una arista si algunos de sus nodos estaban unidos, es decir: $\exists\, a(n_i, n'_j) \implies \exists\, a'(r, r')$

Los atributos que consideramos para los referentes de discurso son ligeramente distintos que los de los nodos. Conservan un descriptor y una etiqueta morfosintáctica, pero no los lemas ni las palabras de los nodos origen, así como las categorías de los nodos que los componen (eventos, expresiones temporales y entidades nombradas).

El colapsado provoca que nodos anotados con diferentes atributos se agrupen. Estos atributos pueden reforzar una evidencia o aportar datos complementarios, pero en ocasiones también muestran datos contradictorios. En estos casos, se han tomado decisiones distintas dependiendo del atributo que se esté tratando.

El descriptor debe ser la cadena de caracteres más representativa del referente de discurso. Por ello, como primera aproximación se ha empleado el descriptor más largo de los nodos de origen. En un futuro, podría crearse una base de conocimiento con las distintas concurrencias de descriptores, y a partir de ella escoger el correcto. Esto funcionaría de manera similar a un sistema de desambiguación.

Para la etiqueta morfosintáctica, hemos escogido asignar a las entidades nombradas el valor *N* y a los eventos *V*, mientras que para el resto de palabras se mantiene la correspondiente al nodo de origen. Esto es una aproximación sencilla que busca la normalización de los nodos que suelen contener la información más importante.

El colapsado puede suponer también que algunos referentes de discurso tengan varias etiquetas de una misma categoría. Esto no supone un problema en el caso de los eventos o las expresiones temporales, ya que siempre agregan información complementaria, pero sí en el caso de las entidades nombradas, que pueden tener tipos distintos. En este caso, se escogen aquellos tipos que sean más co-

munes, descartando los demás. Al igual que los descriptores, esta información se puede almacenar en una base de conocimiento y recuperarla para mejorar el proceso de asignación de tipos.

Tanto el proceso de colapsado como el resto de tareas, incluyendo la explicitación de la información semántica y la simplificación y normalización del grafo se han realizado mediante las reglas incluidas en el cuadro 1.

2.3. Procesamiento

Para crear esta representación, se ha seguido un proceso que consta de las siguientes fases: (1)Análisis morfosintáctico, reconocimiento de entidades nombradas y resolución de correferencia. Efectuado con el programa Stanford CoreNLP (Klein y Manning, 2003). (2)Etiquetado de eventos y expresiones temporales, e inclusión de aristas con relaciones temporales mediante el programa Tarsqi Toolkit (Verhagen et al., 2005). (3)Colapsado de nodos en referentes de discurso. (4)Simplificación y normalización del grafo mediante reglas.

En la Figura 1 se pueden observar los grafos con las dos configuraciones. En la configuración inicial (grafo superior) se han realizado los pasos 1 y 2, mientras que en la configuración colapsada (grafo inferior) se han añadido los pasos 3 y 4.

3. Evaluación

Dado que el objetivo de la transformación de documentos en grafos es mejorar el rendimiento de aplicaciones que empleen esta representación, realizaremos la evaluación de su funcionamiento de manera extrínseca. Para evaluar el sistema hemos escogido la tarea Slot Filling (Ji, Grishman, y Dang, 2011) en el marco del Knowledge Base Population en la conferencia Text Analysis Conference.

Para esta tarea se distribuye una base de conocimiento creada a partir de las infoboxes de Wikipedia que contiene las referencias para entrenar los sistemas, más una colección 1.7 millones de documentos de diversas fuentes, como noticias, conversaciones telefónicas y texto web, que será el espacio de búsqueda de soluciones.

Slot Filling se divide en dos subtareas, Regular Slot Filling (RSF) y Temporal Slot Filling (TSF). La evaluación de la representación la hemos realizado sobre esta última.

El objetivo es encontrar el valor de una lista cerrada de atributos para diversas entidades y acotarlo temporalmente. Las entidades pueden ser personas u organizaciones, mientras que los atributos dependen de la entidad, para personas son:*estados de residencia*, *ciudades de residencia*, *títulos*, *miembro de*, *empleado de*, *esposos*; mientras que para organizaciones es unicamente *directivos*.

Para esto se emplean una serie de consultas compuestas por la dupla $\langle entidad, atributo \rangle$, a las que los sistemas deben responder con una tripleta $\langle entidad, atributo, valor \rangle$. Por ejemplo, para la consulta $\langle Barack\ Obama, spouse \rangle$ la respuesta debería ser $\langle Barack\ Obama, spouse, Michelle\ Obama \rangle$.

La respuesta correcta a un atributo puede consistir en una lista de valores. Solo se valoran los valores correctos, y las respuestas redundantes son ignoradas.

En la subtarea TSF se pide además acotar temporalmente las tripletas. Para ello, se define un intervalo temporal impreciso como una tupla de cuatro valores (t_1, t_2, t_3, t_4), que indican que la relación comienza en un punto entre t_1 y t_2 y termina en otro entre t_3 y t_4. Si el valor de t_1 o t_3 está sin rellenar significa que el valor es $-\infty$, mientras que para t_2 o t_4 es $+\infty$.

Para responder a las preguntas de investigación planteamos los siguientes experimentos:

(1) Medir nuestro sistema frente a otros participantes para comprobar si es apropiada la representación en forma de grafo para la tarea.

(2) Entrenar dos clasificadores con representaciones distintas para evaluar la utilidad del colapsado, la normalización y el enriquecimiento semántico. En el primero empleamos la configuración inicial, mientras que en el segundo utilizamos la configuración colapsada.

3.1. Resultados

El cuadro 2 muestra los resultados generales de la tarea Temporal Slot Filling. En ella se puede observar que el funcionamiento del sistema ha sido similar a otros sistemas en el estado del arte, siendo la precisión la más alta de todos los participantes. Además, a pesar de la baja cobertura, el sistema obtiene la tercera mejor medida F_1.

Esto implica que la representación gráfica supone una posibilidad prometedora en la

Colapsado de Referentes de Discurso	
X correference Y	$r = X \cup Y$
X dep1 Y, X dep2 Y, $dep1 = dep\ 2$	X dep1 Y
X nn Y, X amod Y	X nn Y
X dep X	
X class Y, X correference Y	X class Y
Clases Semánticas	
Antecedente	**Consecuente**
NE nn NN	NE hasClass NN
NE appos NN	NE hasClass NN
NE abbrev NN	NE hasClass NN
NN appos NE	NE hasClass NN
NN abbrev NE	NE hasClass NN
NE nsubj NN	NE hasClass NN
NE is NN	NE hasClass NN
Propiedades	
Antecedente	**Consecuente**
JJ nsubj X, JJ cop Y	X has_property JJ
JJ arg0 X, JJ cop Y	X has_property JJ
Genitivos	
Antecedente	**Consecuente**
NN nn NE	NE has_NN NN
NN poss NE	NE has_NN NN
X poss Y	Y has X
NN prep_of NE	NE has_NN NN
X has NE , NE has Class	X has_Class NE, NE has Class
NN nsubj NE	NE has_NN NN
Normalización de Argumentos Verbales	
Antecedente	**Consecuente**
V nsubj X	V arg0 X
V xsubj X	V arg0 X
V csubj X	V arg0 X
V agent X	V arg0 X
V nsubjpass X, V arg1 Y	V arg1 Y, V arg2 X
V nsubjpass X	V arg1 X
V dobj X	V arg1 X
V iobj X	V arg2 X
X partmod V	V arg1 X
V xcomp X	V arg1 X
V ccomp X	V arg1 X
V xcomp X, V arg1 Y	V V arg1 Y, arg2 X
V ccomp X, V arg1 Y	V arg1 Y, V arg2 X
Copulativos	
Antecedente	**Consecuente**
NN nsubj X, NN cop Y	X is NN
Edad	
Antecedente	**Consecuente**
PERSON appos NUMBER	NE hasAge NUMBER
PERSON abbrev NUMBER	NE hasAge NUMBER
Género	
Antecedente	**Consecuente**
NE coreference he	NE hasGender MALE
NE coreference she	NE hasGender FEMALE
he coreference NE	NE hasGender MALE
she coreference NE	NE hasGender FEMALE

Cuadro 1: Reglas para la transformación de los grafos. Cada columna se corresponde con una tripleta $\langle governor,\ dependency,\ dependant \rangle$, donde X e Y son dos nodos diferentes del grafo. Nótese que los antecedentes que no aparecen en el consecuente son borrados.

tarea de extracción automática de relaciones, ya que a pesar de encontrarse en las primeras fases de desarrollo permite competir con los sistemas en el estado del arte.

El sistema está formado por varios componentes que funcionan en cadena: Recuperación de información, representación de documentos, aprendizaje semisupervisado, ex-

System	Precision	Recall	F1
BLENDER2	0.1749	0.3261	0.2277
BLENDER1	0.1749	0.3172	0.2255
BLENDER3	0.1642	0.3099	0.2147
IIRG1	0.2404	0.1299	0.1711
Colapsado	**0.2571**	**0.0656**	**0.1045**
Inicial	**0.2299**	**0.0620**	**0.0977**
Stanford 12	0.0206	0.1724	0.0369
Stanford 11	0.0211	0.1491	0.0370
USFD20112	0.0099	0.0053	0.0069
USFD20113	0.0019	0.0004	0.0006

Cuadro 2: Resultados finales de la tarea Temporal Slot Filling

tracción de relaciones y acotación temporal. Este tipo de sistema se ve afectado por la propagación de errores, por lo que es interesante aislar el impacto de la representación de la influencia de otros factores. Por ello estudiaremos la precisión, cobertura y medida F_1 tras la etapa de extracción de relaciones, ya que la acotación temporal es una etapa que es relativamente independiente de la representación.

En el cuadro 3 se muestra los resultados en la tarea Temporal Slot Filling tras la fase de extracción de relaciones. En ella se muestra la precisión y la cobertura calculadas teniendo en cuenta el número de tripletas $\langle$*entidad*, *atributo*, *valor*$\rangle$ obtenidas correctamente.

Los datos muestran que tanto la precisión como la cobertura obtienen valores muy similares en ambos tipos de representaciones, siendo ligeramente mejores en el caso de los grafos colapsados.

Una inspección manual de los datos muestra que los grafos generados contienen numerosos errores, principalmente en la fase de identificación de descriptores en los grafos colapsados, aunque también otros menores a lo largo del procesamiento. Sin embargo, estos errores parecen compensarse con las mejoras, ya que los resultados no sólo no bajan sino que mejoran.

Con estos datos podemos afirmar que la fase de enriquecimiento es útil para la tarea de extracción automática de relaciones, ya que el balance entre las ganancias resultantes de colapsar los grafos y los errores introducidos por este procesamiento es positivo, y esperamos que un refinamiento posterior haga que mejore todavía más.

Configuración	Inicial	Colapsada
Cobertura	0.08	0.08
Precisión	0.42	0.45
F1	0.14	0.14

Cuadro 3: Resultados de la tarea Temporal Slot Filling en la fase de extracción

4. Trabajo Relacionado

El estudio de la representación de los textos es muy amplio. Los primeros acercamientos están basados en las teorías de dependencia conceptual de Schank (1972), y la de significado-texto (*meaning-text theory*) de Mel'cuk y Polguère (1987). Posteriormente se utilizaron técnicas de representación en grandes colecciones de textos, como por ejemplo en el trabajo de Pradhan et al. (1994) aplicado al dominio médico. Sin embargo cada tarea de procesamiento de lenguaje tiene unas necesidades de representación distintas.

La extracción de información tal y como la entendemos en este trabajo comenzó a estudiarse en la Message Understanding Conference (MUC) (Grishman y Sundheim, 1996; Beth, 1995; Chinchor, 1998) y continuó con el programa Automatic Content Extraction (ACE) (Maynard, Bontcheva, y Cunningham, 2003) hasta 2008. Estas fueron las primeras evaluaciones cuantitativas de sistemas de este tipo y en ellas se profundizó en el procesamiento automático de texto.

En 2009, la Text Analisis Conference (TAC) (McNamee y Dang, 2009) tomó el relevo de ACE en las evaluaciones de sistemas de recuperación de información. En esta conferencia se propuso la tarea Knowledge Base Representation (KBP), que se puede ver como una combinación de extracción de información y búsqueda de respuestas en la que se complicaba la tarea al forzar la adquisición de información en múltiples documentos. En esta tarea se enmarcan las subtareas Regular Slot Filling (SF) y Temporal Slot Filling (TSF).

Como se puede ver en (Ji y Grishman, 2011), SF y TSF permanecen como problemas de investigación abiertos ya que los competidores todavía no son capaces de acercarse a los resultados de la anotación manual. Los participantes en esta edición utilizan sistemas similares de representación, por ejemplo, los sistemas de CUNY (Artiles et al.,

2011) y Stanford (Surdeanu et al., 2011) emplean también tokenización, segmentación, detección de entidades nombradas, resolución de correferencia y análisis de dependencias sintácticas. En el caso del sistema de CUNY también emplean n-gramas de varias longitudes como modelos de bolsa de palabras. Sin embargo ninguno de ellos emplea una representación gráfica a nivel de documento como aquí se propone.

5. Conclusiones

Este trabajo muestra una representación de documentos como grafos morfosintácticos enriquecidos semánticamente. Con ella, disponemos de la base para realizar múltiples tareas de procesamiento de lenguaje natural como resolución de correferencias, extracción de información o búsqueda de respuestas.

Los resultados obtenidos nos permiten comprobar que este tipo de representación resulta prometedora para el funcionamiento de los clasificadores en la tarea de extracción automática de relaciones. El estado de evolución de la representación nos invita a pensar que quedan muchas alternativas por explorar y que una versión más avanzada de la misma nos permitirá mejorar los resultados.

A pesar de los errores del procesamiento, comprobamos cómo los grafos colapsados se comportan mejor que los grafos iniciales. Estas nuevas estructuras nos permitirán realizar procesos como agregación de información, asignación automática de clases semánticas o desambiguación de entidades, que podrían utilizarse como base de conocimiento para enriquecer nuevos grafos.

Además, el proceso de creación de esta representación está basado en herramientas externas intercambiables, lo que permite tener una aplicación modular y flexible en la que encajan fácilmente futuros cambios.

Sin embargo, todavía existe mucho margen de mejora en la representación. Dado que nuestro objetivo nos obligaba a agrupar información muy heterogénea, hemos empleado técnicas sencillas en cada uno de los pasos.

6. Trabajo futuro

La información de un texto que es fácilmente predecible por el lector tiende a omitirse, como se apunta en (Peñas y Ovchinnikova, 2012). Sustituir las aristas genéricas de estructuras que suelan codificar este tipo de información, como genitivos o compuestos nominales, por otras más específicas que indiquen la naturaleza de la relación entre los componentes podría ayudar a mejorar los sistemas de extracción de información.

Otra posibilidad de mejora es incluir un sistema de correferencia de eventos (Humphreys, Gaizauskas, y Azzam, 1997; Hasler, Orasan, y Naumann, 2008), y realizar un proceso de colapsado similar al que se emplea con la correferencia de entidades nombradas. De esta manera se conseguiría una mayor cohesión de la información.

Por último, sería muy interesante generar bases de conocimiento agregando automáticamente la información de distintos documentos de manera similar a (Peñas y Hovy, 2010; Banko et al., 2007; Clark y Harrison, 2009). Por ejemplo, podríamos seleccionar subgrafos para seleccionar estructuras morfosintácticas o semánticas interesantes, o crear ontologías a partir de las clases semánticas inferidas. Esta información podría utilizarse para enriquecer los grafos y mejorar así la cobertura del sistema de extracción de relaciones.

Bibliografía

Artiles, Javier, Qi Li, Taylor Cassidy, Suzanne Tamang, y Heng Ji. 2011. Cuny blender tac-kbp2011 temporal slot filling system description. En *Proceedings of the Text Analysis Conference*.

Banko, Michele, Michael J. Cafarella, Stephen Soderl, Matt Broadhead, y Oren Etzioni. 2007. Open information extraction from the web. En *In IJCAI*, páginas 2670–2676.

Beth, Sundheim. 1995. Proceedings of the sixth message understanding conference. MUC-6, Columbia, MD.

Chinchor, Nancy A. 1998. Overview of proceedings of the seventh message understanding conference. En *Proceedings of the Seventh Message Understanding Conference*, MUC-7, Fairfax, VA.

Clark, Peter y Phil Harrison. 2009. Large-scale extraction and use of knowledge from text. En *Proceedings of the fifth international conference on Knowledge capture*, K-CAP '09, páginas 153–160, New York, NY, USA. ACM.

Grishman, Ralph y Beth Sundheim. 1996. Message understanding conference-6: a

brief history. En *Proceedings of the 16th International Conference on Computational Linguistics*, ICCL, páginas 466–471.

Hasler, Laura, Constantin Orasan, y Karin Naumann. 2008. Nps for events: Experiments in coreference annotation.

Heim, Irene, 1983. *File Change Semantics and the Familiarity Theory of Definiteness*, páginas 164–189. Walter de Gruyter.

Humphreys, Kevin, Robert Gaizauskas, y Saliha Azzam. 1997. Event coreference for information extraction.

Ji, Heng y Ralph Grishman. 2011. Knowledge Base Population: Successful Approaches and Challenges. En *ACL HLT 2011*, páginas 1148–1158.

Ji, Heng, Ralph Grishman, y Hoa Trang Dang. 2011. Overview of the tac2011 knowledge base population track. En *Text Analysis Conference, TAC 2011 Workshop, Notebook Papers*.

Karttunen, Lauri. 1968. *What do referential indices refer to?* Rand Corporation: [Paper]. Rand Corp.

Klein, Dan y Christopher D. Manning. 2003. Accurate unlexicalized parsing. En *ACL 2003*, páginas 423–430.

Maynard, Diana, Kalina Bontcheva, y Hamish Cunningham. 2003. Towards a semantic extraction of named entities. En *In Recent Advances in Natural Language Processing*.

McNamee, Paul y Hoa T. Dang. 2009. Overview of the TAC 2009 knowledge base population track. En *TAC 2009*.

Mel'cuk, Igor y Alain Polguère. 1987. A formal lexicon in the meaning-text theory (or how to do lexica with words). *Computational Linguistics*, 13(3-4):261–275.

Peñas, Anselmo y Eduard Hovy. 2010. Filling knowledge gaps in text for machine reading. En *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, páginas 979–987, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peñas, Anselmo y Ekaterina Ovchinnikova. 2012. Unsupervised acquisition of axioms to paraphrase noun compounds and genitives. En *(Ed.): CICLing 2012, Part I, LNCS 7181*, páginas 388–401, Springer-Verlag.

Pradhan, Malcolm, Gregory Provan, Blackford Middleton, y Max Henrion. 1994. Knowledge engineering for large belief networks. En *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, UAI'94, páginas 484–490, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Recasens, Marta. 2010. Coreference: Theory, annotation, resolution and evaluation.

Schank, Roger C. 1972. Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, 3(4):552–631, October.

Surdeanu, Mihai, Sonal Gupta, John Bauer, David McClosky, Angel X. Chang, Valentin I. Spitkovsky, y Christopher D. Manning. 2011. Stanford's distantly-supervised slot-filling system. En *Proceedings of the Fourth Text Analysis Conference (TAC 2011)*, Gaithersburg, Maryland, USA, November.

Verhagen, Marc, Inderjeet Mani, Roser Sauri, Robert Knippen, Seok Bae Jang, Jessica Littman, Anna Rumshisky, John Phillips, y James Pustejovsky. 2005. Automating temporal annotation with TARSQI. En *ACLdemo'05*.