Reconocimiento y Clasificación de Entidades Nombradas independiente de la lengua y el dominio mediante perfiles

Language and Domain Independent Named Entity Recognition and Classification through Profiles

Isabel Moreno
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Apdo. de correos, 99 E-03080 Alicante
imoreno@dlsi.ua.es

Resumen: El reconocimiento y la clasificación de entidades nombradas (RCEN) es clave para muchas aplicaciones de procesamiento de lenguaje natural. Sin embargo, la adaptación de un sistema RCEN suele resultar costosa, ya que la mayoría solo funcionan adecuadamente en el escenario para el que fueron desarrollados. Por tanto, el objetivo principal de esta tesis es la investigación, análisis y desarrollo de un sistema adaptable, llamado CARMEN, para el RCEN mediante perfiles y aprendizaje automático supervisado. La atención se centrará en que CARMEN sea independiente del dominio y la lengua para, con el mismo método, conseguir resultados similares sin importar el corpus de entrenamiento utilizado.
Palabras clave: Entidad nombrada, Perfiles, Aprendizaje automático, Dominio independiente, Lengua independiente

Abstract: Named Entity Recognition and Classification (NERC) is a prerequisite to many natural language processing applications. Nevertheless, the adaptation of NERC systems is usually expensive given that most of them only work appropriately on the scenario for which they were created. Therefore, the main purpose of this thesis is to research, analyse and develop an adaptable system, named CARMEN, for NERC through profiles and supervised machine learning. Attention would be focused on CARMEN being domain and language independent so as to achieve similar results, using the same method, regardless of the training corpus utilised.
Keywords: Named entity, Profiles, Machine learning, Domain independent, Language independent

1 Motivación

Desde hace tiempo estamos en la era de la información digital y, aunque esta crece sin descanso, nuestra habilidad para explotarla y procesarla continúa constante (Bendov y Feldman, 2010). Desde esta perspectiva, el Procesamiento del Lenguaje Natural (PLN) investiga y formula mecanismos computacionales para facilitar la interrelación hombre-máquina por medio del lenguaje natural, en lugar de otros lenguajes más formales y restrictivos, sin perder efectividad (Manaris, 1998; Moreno et al., 1999). Más concretamente, las técnicas de Extracción de Información (EI) procesan texto para detectar la información textual explícita de interés y convertirla a un formato fácilmente comprensible por las máquinas, también conocido como estructurado (Ben-dov y Feldman, 2010).

Una de las tareas que lleva a cabo la EI es el Reconocimiento y la Clasificación de Entidades Nombradas (RCEN) (Nadeau y Sekine, 2007; Marrero et al., 2013), que tiene dos objetivos diferenciados. Primero, identificar las menciones de nombres propios en un texto, lo que se conoce como la fase de reconocimiento (REN). Segundo, asignar una categoría, de entre un conjunto predeterminado, a cada una de las entidades previamente reconocidas, llamada fase de clasificación (CEN). Ambos objetivos pueden abordarse de manera conjunta o separada.

Los sistemas RCEN juegan un papel importante en muchas aplicaciones que procesan información textual. La razón es que el RCEN es un prerrequisito para diversas tareas como: la minería de opiniones (Ding, Liu, y Zhang, 2009; Jin, Hay Ho, y Srihari, 2009), la gene-

ración automática de resúmenes (Fuentes y Rodríguez, 2002; Alcón y Lloret, 2015), la generación de lenguaje natural (Vicente y Lloret, 2016), los sistemas de búsqueda de respuestas (Peregrino, Tomás, y Pascual, 2012; Lee, Hwang, y Jang, 2007; Lee et al., 2006) o los sistemas de recuperación de información (Guo et al., 2009; Chen, Ding, y Tsai, 1998), entre otras aplicaciones.

A pesar de que los sistemas RCEN son de uso común, su utilización no siempre es directa. La mayoría de sistemas RCEN fueron desarrollados ad-hoc para un dominio[1] concreto, con requisitos específicos y, a su vez, un conjunto reducido de tipos de entidades de interés en ese dominio. Como resultado, cuando se quiere portar una herramienta RCEN a otro dominio, con otros requisitos y un conjunto diferente de entidades, se requiere un esfuerzo considerable para que funcione adecuadamente (Marrero et al., 2013).

Además, el RCEN está condicionado por la lengua para la que se desarrollan los sistemas. La mayoría de herramientas se construyen para un corpus específico y, como consecuencia, existe una dependencia de la lengua de dicho corpus. La adaptación de un RCEN a un nuevo idioma no siempre es posible por tres razones principales: (i) estos sistemas RCEN suelen necesitar de herramientas de análisis lingüístico que no siempre están disponibles para todos los idiomas (Indurkhya, 2014); (ii) el RCEN depende comúnmente de otros recursos (como diccionarios) que varían entre lenguas (Marrero et al., 2013), si es que existen; y (iii) cada idioma supone retos diferentes que pueden afectar al rendimiento del RCEN, como se observó en (Tjong Kim Sang, 2002; Sang y De Meulder, 2003).

Por ello, el presente proyecto de tesis se centrará en analizar, proponer y desarrollar un sistema RCEN, llamado CARMEN, basado en perfiles que empleará aprendizaje automático supervisado. Se buscará que dicho sistema sea independiente del dominio y de la lengua para, con el mismo método, conseguir resultados similares sin importar el corpus de entrenamiento utilizado.

2 Trabajo relacionado

Hace más de dos décadas que fue acuñado el término Entidad Nombrada (EN) en la sexta conferencia "Message Understanding Conference" (MUC). En ella el objetivo era la RCEN de personas, organizaciones, lugares así como expresiones numéricas de tiempo y cantidad (Grishman y Sundheim, 1996). Desde entonces, diversos foros de PLN han seguido sus pasos, promocionando tareas para evaluar sistemas RCEN (Tjong Kim Sang, 2002; Sang y De Meulder, 2003; Uzuner, Solti, y Cadag, 2010; Segura-Bedmar, Martınez, y Herrero-Zazo, 2013; Ji, Nothman, y Hachey, 2014; Pradhan et al., 2014; Elhadad et al., 2015; Ji, Nothman, y Hachey, 2015).

Se observan dos patrones en el foco de las mismas: (i) un dominio y múltiples idiomas (Tjong Kim Sang, 2002; Sang y De Meulder, 2003; Ji, Nothman, y Hachey, 2014; Ji, Nothman, y Hachey, 2015); o (ii) un dominio restringido y un solo idioma (Uzuner, Solti, y Cadag, 2010; Segura-Bedmar, Martınez, y Herrero-Zazo, 2013; Pradhan et al., 2014; Elhadad et al., 2015).

Un ejemplo del primer caso lo encontramos en la conferencia CoNLL, donde se organizaron dos competiciones (Tjong Kim Sang, 2002; Sang y De Meulder, 2003) para tratar el RCEN en noticias de periódicos en inglés, holandés, castellano y alemán. En ambas ediciones, los sistemas obtuvieron diferentes resultados en cada idioma. Por ejemplo, el mejor sistema de cada edición tuvo una diferencia de al menos 15 puntos en la F1 global, por lo que es discutible que sean completamente independientes del idioma. Más recientemente se han investigado otras aproximaciones RCEN multilingües (Konkol et al., 2015; Agerri y Rigau, 2016), donde también se observan diferentes resultados en cada uno de los idiomas (aproximadamente 15 puntos de F1 global).

Respecto al último caso, un ejemplo de competición centrada en un dominio restringido y un idioma es el DDIExtraction 2013 (Segura-Bedmar, Martınez, y Herrero-Zazo, 2013), organizado dentro del taller internacional SemEval. Uno de sus objetivos principales es el RCEN en dos fuentes médicas de información textual (DrugBank y MedLine). También en este caso los participantes obtuvieron resultados diferentes según la fuente (al menos 20 puntos en la F1 global).

Fuera de estos marcos de evaluación y la multilingualidad, Tkachenko y Simanovsky (2012) diseñan un RCEN y experimentan con varios géneros textuales presentes en el corpus

[1] En esta tesis se entiende por dominio al tópico o área de interés de un corpus, como pueden ser el dominio médico o el educativo.

OntoNotes. Kitoogo y Baryamureeba (2008) definen un RCEN que se probó en dos dominios (general y legislativo): entrenando en el dominio general (Sang y De Meulder, 2003) y evaluando en el legislativo, y viceversa. Ambos trabajos obtienen una diferencia de al menos 20 puntos en la F1 global cuando cambian de dominio o género.

Aunque vemos que se han hecho progresos considerables en el RCEN, los resultados de las investigaciones ponen de manifiesto que los sistemas no han mostrado un rendimiento óptimo cuando cambia el idioma o el dominio, así como la fuente o el género textual.

3 Propuesta de investigación

Dado este panorama general, esta tesis doctoral plantea como objetivo la investigación, análisis y desarrollo de un sistema adaptable para el *RCEN*, llamado CARMEN. La atención se centrará, sobre todo, en que CARMEN proporcione salidas consistentes aun cuando cambie el dominio o la fuente o el género o el idioma del corpus de entrenamiento.

Por tanto, la hipótesis de partida es que el desarrollo de un sistema RCEN basado en perfiles con aprendizaje automático, redunda en herramientas con mínima adaptación, evitando las diferencias observadas en los resultados actuales en relación a dependencias del dominio o de la lengua. En cuanto a la dependencia del dominio, concretamente, nos planteamos el estudio de nuestra aproximación en al menos dos dominios: (i) general, que representa necesidades de información comunes; y (ii) farmacoterapéutico, que representa necesidades de información específicas durante la atención sanitaria. Ambos dominios son altamente representativos en cuanto a géneros textuales, idiomas y entidades nombradas. Por tanto, estos dominios permiten definir un escenario de evaluación apropiado para confirmar nuestra hipótesis y definir objetivos específicos:

- O1 Realizar un estado de la cuestión, sistemático y exhaustivo, para detectar las limitaciones tanto de las aproximaciones para RCEN como de los corpus existentes, al menos, en dos dominios: general y farmacoterapéutico.
- O2 Analizar las entidades nombradas relevantes en el dominio farmacoterapéutico y crear un corpus en español para el RCEN en este dominio, así como evaluar la calidad del recurso generado.
- O3 Diseñar e implementar nuevas técnicas de RCEN que permitan solventar alguna de las limitaciones de las aproximaciones encontradas en el estado de la cuestión:
  - O3.1 diseñar nuevas técnicas de REN,
  - O3.2 diseñar nuevas técnicas de CEN, y
  - O3.3 diseñar nuevas técnicas de desambiguación de entidades con respecto a bases de conocimiento y las relaciones entre las mismas.
- O4 Diseñar y analizar experimentos para evaluar el resultado del objetivo O3, en al menos dos dominios y dos lenguas, realizando para ello una evaluación intrínseca y extrínseca, basada en métodos cuantitativos y cualitativos.

4 Metodología y experimentos

Con el fin de demostrar la hipótesis y los objetivos presentados en la sección anterior, hasta ahora hemos llevado a cabo tres grandes tareas:

Primero, la creación del corpus DrugSemantics para el RCEN en el dominio farmacoterapéutico ha concluido con la creación de un *gold standard*, siguiendo la metodología descrita en (Moreno et al., 2017).

Segundo, la implementación de un RCEN, llamado MaNER, basado en lexicones específicos del dominio médico (Moreno, Moreda, y Romá-Ferri, 2012; Moreno, Moreda, y Romá-Ferri, 2015; Moreno, Moreda, y Romá-Ferri, 2015; Moreno et al., 2017), que sirvió de apoyo en la construcción de un *gold standard* de calidad.

Tercero, el desarrollo del módulo CEN basado en perfiles del sistema CARMEN, que emplea aprendizaje automático supervisado y cuyas características incluyen información local a la entidad (afijos, longitud y la propia entidad) así como información de contexto en una ventana. El contexto se consigue mediante perfiles generados para cada una de las entidades. Los resultados de diversos experimentos nos han permitido estudiar diferentes parámetros en corpus de diferentes dominios (Moreno, Romá-Ferri, y Moreda, 2017d), así como su rendimiento (Moreno, Moreda, y Romá-Ferri, 2016; Moreno, Romá-Ferri, y Moreda, 2017b; Moreno, Romá-Ferri, y Moreda, 2017c). Además, estamos experimentan-

do en varios idiomas (Moreno, Romá-Ferri, y Moreda, 2017a).

5 Elementos específicos para discusión

Siendo el RCEN un tema de gran interés en el PLN, queremos intercambiar experiencias para orientar nuestra investigación.

En concreto, son dos los intereses a debatir:

- nuestra próxima tarea, la REN, así como las diferentes técnicas, características y herramientas que permitirían construir este módulo independiente del dominio y la lengua.
- posibles escenarios y experimentos que nos permitan reforzar nuestra hipótesis.

Agradecimientos

Esta investigación ha sido financiada parcialmente por el Gobierno Español (TIN2015-65100-R y TIN2015-65136-C2-2-R), la Generalitat Valenciana (PROMETEOII/2014/001), la Universidad de Alicante (GRE16-01: Plataforma inteligente para recuperación, análisis y representación de la información generada por usuarios en Internet) y las Ayudas Fundación BBVA a equipos de investigación científica 2016 (ASAP - Análisis de Sentimientos Aplicado a la Prevención del Suicidio en las Redes Sociales).

Bibliografía

Agerri, R. y G. Rigau. 2016. Robust multilingual Named Entity Recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63–82.

Alcón, Ó. y E. Lloret. 2015. Estudio de la influencia de incorporar conocimiento léxico-semántico a la técnica de Análisis de Componentes Principales para la generación de resúmenes multilingües. *Linguamática*, 7(1):53–63, Julio.

Ben-dov, M. y R. Feldman. 2010. Text Mining and Information Extraction. En O. Maimon y L. Rokach, editores, *Data Mining and Knowledge Discovery Handbook*. Springer US, Boston, MA, 2nd edición, capítulo 42, páginas 809–835.

Chen, H., Y. Ding, y S. Tsai. 1998. Named Entity Extraction for Information Retrieval. En *COMPUTER PROCESSING OF ORIENTAL LANGUAGES*, volumen 11.

Ding, X., B. Liu, y L. Zhang. 2009. Entity Discovery and Assignment for Opinion Mining Applications. En *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 1125–1134.

Elhadad, N., S. Pradhan, S. L. Gorman, S. Manandhar, W. W. Chapman, y G. Savova. 2015. SemEval-2015 Task 14 : Analysis of Clinical Text. *Proceedings of the 9th International Workshop on Semantic Evaluation*, páginas 303–310.

Fuentes, M. y H. Rodríguez. 2002. Using cohesive properties of text for automatic summarization. En *Actas de las Jornadas de tratamiento y recuperación de la información (Jotri'2002)*.

Grishman, R. y B. Sundheim. 1996. Message understanding conference-6: A brief history. En *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, páginas 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.

Guo, J., G. Xu, X. Cheng, y H. Li. 2009. Named Entity Recognition in Query. En *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, páginas 267–274, Boston, Massachusetts, USA.

Indurkhya, N. 2014. Natural Language Processing. En T. Gonzalez J. Díaz-Herrera, y A. Tucker, editores, *Computing Handbook, Third Edition: Computer Science and Software Engineering*. CRC Press, capítulo 40, páginas 40:1–17.

Ji, H., J. Nothman, y B. Hachey. 2014. Overview of TAC-KBP2014 Entity Discovery and Linking Tasks. En *Proceedings of Text Analysis Conference*.

Ji, H., J. Nothman, y B. Hachey. 2015. Overview of TAC-KBP2015 Entity Discovery and Linking Tasks. En *Proceedings of Text Analysis Conference 2015*.

Jin, W., H. Hay Ho, y R. K. Srihari. 2009. OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction. En *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 1195–1204, Paris, France.

Kitoogo, F. y V. Baryamureeba. 2008. Towards domain independent named entity recognition. En *Strengthening the Role of ICT in Development*, volumen IV. Fountain publishers, páginas 84 – 95.

Konkol, M., T. Brychcín, Konopí, y M. K. 2015. Latent semantics in Named Entity Recognition. *Expert Systems with Applications*, 42(7):3470–3479.

Lee, C., Y.-G. Hwang, y M.-G. Jang. 2007. Fine-Grained Named Entity Recognition and Relation Extraction for Question Answering. En *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 799–800, Amsterdam, The Netherlands.

Lee, C.-H., Y. G. Hwang, H. J. Oh, S. Lim, J. Heo, C. H. Lee, H. J. Kim, J. H. Wang, y M. G. Jang. 2006. Fine-grained Named Entity Recognition using Conditional Random Fields for Question Answering. En *Information Retrieval Technololgy, Proceedings*, volumen 4182. Springer, Berlin, Heidelberg, páginas 581–587.

Manaris, B. 1998. Natural Language Processing: A Human-Computer Interaction Perspective. En *Advances in Computers*, volumen 47. páginas 1–66.

Marrero, M., J. Urbano, S. Sánchez-Cuadrado, J. Morato, y J. M. Gómez-Berbís. 2013. Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards and Interfaces*, 35(5):482–489.

Moreno, I., E. Boldrini, P. Moreda, y M. T. Romá-Ferri. 2017. Drugsemantics: A corpus for named entity recognition in spanish summaries of product characteristics. *Journal of Biomedical Informatics*, 72:8 – 22.

Moreno, I., P. Moreda, y M. T. Romá-Ferri. 2016. An active ingredients entity recogniser system based on profiles. En *21st International Conference on Applications of Natural Language to Information Systems*, volumen 9612 de *LNCS*, páginas 276–284, Salford. Springer.

Moreno, I., P. Moreda, y M. T. Romá-Ferri. 2015. Estudio de fiabilidad y viabilidad de la Web 2.0 y la Web semántica para enriquecer lexicones en el dominio farmacológico. *Procesamiento del Lenguaje Natural*, 55:65–72.

Moreno, I., P. Moreda, y M. Romá-Ferri. 2012. Reconocimiento de entidades nombradas en dominios restringidos. En *Actas del III Workshop en Tecnologías de la Informática*. páginas 41–57.

Moreno, I., P. Moreda, y M. Romá-Ferri. 2015. MaNER: a MedicAl Named Entity Recogniser for Spanish. En *20th International Conference on Applications of Natural Language to Information Systems*, volumen 9103 de *LNCS*, páginas 418–423, Passau. Springer.

Moreno, I., M. T. Romá-Ferri, y P. Moreda. 2017a. Language independent proposal to profile-based named entity classification. En *The First Workshop on Multi-Language Processing in a Globalising World*, páginas 21–30, Dublin.

Moreno, I., M. T. Romá-Ferri, y P. Moreda. 2017b. Named entity classification based on profiles: A domain independent approach. En *22nd International Conference on Applications of Natural Language to Information Systems*, volumen 10260 de *LNCS*, páginas 142–146, Lieja. Springer.

Moreno, I., M. T. Romá-Ferri, y P. Moreda. 2017c. Propuesta de un sistema de clasificación de entidades basado en perfiles e independiente del dominio. *Procesamiento del Lenguaje Natural*, 59.

Moreno, I., M. Romá-Ferri, y P. Moreda. 2017d. A domain and language independent named entity classification approach based on profiles and local information. En *Recent Advances in Natural Language Processing*, páginas 510–518, Varna (To appear).

Moreno, L., M. Palomar, A. Molina, y A. Fernández. 1999. *Introducción al procesamiento del lenguaje natural*. Publicaciones Universidad de Alicante.

Nadeau, D. y S. Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, jan.

Peregrino, F. S., D. Tomás, y F. L. Pascual. 2012. Question Answering and Multi-search Engines in Geo-Temporal Information Retrieval. En A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part II*. Springer, Berlin, Heidelberg, páginas 342–352.

Pradhan, S., N. Elhadad, W. W. Chapman, S. Manandhar, y G. Savova. 2014. SemEval-2014 Task 7: Analysis of Clinical Text. páginas 54–62.

Sang, E. F. T. K. y F. De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of the 7th Conference on Natural Language Learning*, páginas 142–147.

Segura-Bedmar, I., P. Martınez, y M. Herrero-Zazo. 2013. SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). En *Proceedings of the 7th International Workshop on Semantic Evaluation*, páginas 341–350.

Tjong Kim Sang, E. F. 2002. Introduction to the CoNLL-2002 shared task. En *Proceeding of the 6th Conference on Natural Language Learning*.

Tkachenko, M. y A. Simanovsky. 2012. Selecting Features for Domain-Independent Named Entity Recognition. En *Proceedings of KONVENS 2012*, páginas 248–253.

Uzuner, O., I. Solti, y E. Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–8.

Vicente, M. y E. Lloret. 2016. Exploring Flexibility in Natural Language Generation throughout Discursive Analysis of New Textual Genres. *Proceedings of the 2nd International Workshop Future and Emerging Trends in Language Technologies, Machine Learning and Big Data (FETLT)*.