Procesamiento del Lenguaje Natural, Revista nº 52 marzo de 2014, pp 87-90 recibido 27-09-13 revisado 06-02-14 aceptado 07-02-14

Diseño y generación semi-automática de patrones adaptables para el Reconocimiento de Entidades

Design and semi-automatic generation of adaptable patterns for Entity Recognition

Mónica Marrero
Universidad Carlos III de Madrid
Avenida de la Universidad 30, Leganés
mmarrero@inf.uc3m.es

Resumen: Tesis doctoral en Ciencia y Tecnología Informática realizada por Mónica Marrero Llinares en la Universidad Carlos III de Madrid bajo la dirección de la Dra. Sonia Sánchez Cuadrado y el Dr. Jorge Morato Lara. El acto de defensa tuvo lugar el martes 21 de mayo de 2013 ante el tribunal formado por los doctores Juan Lloréns Morillo (Universidad Carlos III de Madrid), Rafael Valencia García (Universidad de Murcia) y Roberto Carniel (Universidad de Udine). Obtuvo mención internacional y la calificación de Sobresaliente Cum Laude por unanimidad.
Palabras clave: Extracción de Información, Reconocimiento de Entidades Nombradas, generación automática de patrones

Abstract: PhD Thesis in Computer Science written by Mónica Marrero Llinares at the University Carlos III of Madrid under the supervision of Dra. Sonia Sánchez Cuadrado and Dr. Jorge Morato Lara. The author was examined on 21st May 2013 by a committee formed by the doctors Juan Lloréns Morillo (University Carlos III of Madrid), Rafael Valencia García (University of Murcia) and Roberto Carniel (University of Udine). It obtained the grade of Excellent Cum Laude unanimously and received the international mention.
Keywords: Information Extraction, Named Entity Recognition, Automatic Pattern Generation

1 Introducción

La información digital es uno de los principales activos de cualquier organización, y su buena gestión puede determinar el éxito en muchos sectores. Esta gestión supone no sólo el tratamiento de la información generada en la organización, sino también la posibilidad de hacer uso de la información que nos rodea, y que puede suponer ventajas competitivas: estudios de mercado, análisis de opiniones, seguridad, vigilancia tecnológica, etc. En este entorno cobra especial importancia la Extracción de Información, que tiene por objetivo recuperar directamente aquellos elementos de información de nuestro interés en lugar de recuperar documentos completos como ocurre en Recuperación de Información, reduciendo así la cantidad de información que necesitamos leer.

Una de las áreas de la Extracción de Información es la llamada Reconocimiento de Entidades Nombradas (Named Entity Recognition –NER–), que tiene por objetivo la identificación de semánticas relevantes en un texto (e.g. personas, localizaciones y fechas en textos periodísticos). Se trata de un área de investigación especialmente activa en ámbitos como el de la Biomedicina (Marrero et al., 2010a) y con aplicación en otras áreas relacionadas con la gestión de información y el procesamiento del lenguaje natural, como la anotación semántica, los sistemas de búsqueda de respuesta, la población de ontologías y el análisis de opinión.

El núcleo de los sistemas NER son los patrones capaces de reconocer las entidades de interés. La generación manual de estos patrones es compleja, por lo que es habitual el uso de métodos de generación automática que usan corpus anotados para el aprendizaje. Sin embargo, la efectividad de estos métodos está frecuentemente reñida con el esfuerzo que supone para el usuario anotar los corpus.

Tras una introducción del área y sus antecedentes en los capítulos 1 y 2, la tesis ana-

ISSN 1135-5948

liza las herramientas, objetivos y evaluación del área en los capítulos 3 y 4, y muestra que NER está lejos de ser un área resuelta como se ha llegado a afirmar en la literatura. Presenta entonces un método capaz de generar patrones adaptables a los nuevos tipos de entidades y dominios exigidos actualmente, y más consecuente con el problema de no disponer de corpus anotados. Para ello en el capítulo 5 se diseña un modelo de representación que soporta patrones:

- Flexibles: permiten incorporar diferentes tipos de atributos del texto capaces de describir las entidades o su contexto.
- Potentes: capaces de reconocer diferentes estructuras del lenguaje.
- Editables: legibles y basados en estándares para facilitar su edición ante pequeños cambios, evitando así generarlos de nuevo con los costes que ello conlleva.

En el capítulo 6 se describe un método de generación semi-automática de estos patrones que no requiere corpus anotados de partida, sino que guía al usuario en el proceso de anotación. La relación efectividad/coste de anotación lograda es evaluada en el capítulo 7, donde se muestra que se obtienen mejores tasas en relación al estado del arte. Finalmente los capítulos 8 y 9 exponen las conclusiones y trabajos futuros de la tesis.

2 Principales Contribuciones

El Reconocimiento de Entidades Nombradas ha sido evaluado en grandes foros a nivel internacional. Las elevadas tasas de efectividad observadas en dos de estos foros, Message Understanding Conference (MUC) (Grishman y Sundheim, 1996) y Conference on Natural Learning (CoNLL) (Sang y Meulder, 2003), entre finales de los 90 y principios de los 2000 podrían justificar que la tarea de reconocimiento de entidades en texto esté resuelta, y así ha llegado a afirmarse en la literatura (Cunningham, 2005).

La tesis profundiza en la evaluación del área y muestra que se ha estancado en el reconocimiento de entidades típicas, para las que habitualmente existen recursos anotados. Las contribuciones de esta tesis comienzan por mostrar que la validez de estos foros es limitada ante las necesidades actuales. Es necesario avanzar en NER hacia foros de evaluación más dinámicos y sistemas más prácticos, capaces de adaptarse a las nuevas semánticas exigidas, y que además faciliten la anotación de los recursos necesarios para la evaluación (Marrero et al., 2009; Marrero et al., 2013).

La contribución principal de este trabajo consiste en presentar una alternativa ante los métodos de generación del estado del arte que resulta ser más adaptable ante el reconocimiento de nuevas entidades, y más práctica al no requerir la anotación de corpus previos. Esta alternativa consta de un modelo de representación de patrones y un método de generación de estos patrones. Los resultados obtenidos muestran que reduce los costes necesarios para alcanzar los mismos niveles de efectividad frente a métodos similares.

2.1 Modelo de Representación

Para poder incorporar cualquier atributo del lenguaje que pueda ser útil en la identificación de entidades (e.g. categoría gramatical, ortografía, etc.) es necesario contar con un modelo de representación de patrones que lo soporte. Un análisis de los modelos existentes muestra que ninguno reúne conjuntamente requisitos de potencia, flexibilidad y legibilidad. Se diseña entonces un modelo de representación basado en Gramáticas Independientes del Contexto (GIC), con su misma potencia y legibilidad pero que además permite el uso de diferentes atributos a la vez. Para ello este nuevo modelo introduce funciones asociadas a los símbolos no terminales. Estas funciones permiten representar condiciones (pares atributo-valor) sobre las cadenas de entrada de la gramática (Figura 1). A este nuevo modelo se le da el nombre de Information Extraction Grammar (IEG).

El modelo IEG permite representar lenguajes regulares e independientes del contex-

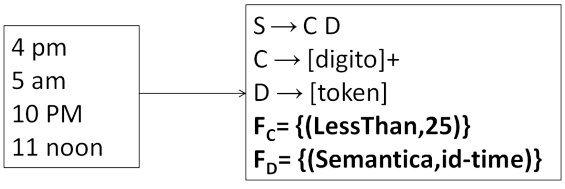

Figura 1: Las entidades a la izquierda pueden representarse con el patrón IEG de la derecha. El no terminal C tiene asociada una función que indica que su valor ha de ser menor que 25 y el no terminal D tiene asociada una función que indica que el atributo *Semántica* de las cadenas reconocidas ha de ser *id-time*.

to. Se comprueba que en ambos casos la complejidad en tiempo de reconocimiento no aumenta respecto a las gramáticas tradicionales (i.e. $O(n)$ si puede transformarse en una expresión regular, $O(n^3)$ en caso contrario, siendo n el tamaño de la cadena de entrada). La condición para ello es que el tiempo de comprobación de las funciones sea constante. Esto supone indizar previamente el valor de los atributos sobre el texto de aplicación, labor que es habitual para optimizar el tiempo de reconocimiento al utilizar cualquier tipo de patrón.

En cuanto a la sintaxis, se busca una notación sobre la que la introducción de las nuevas funciones tenga el menor impacto posible. Para ello se adapta el Speech Recognition Grammar Specification (SRGS) (Hunt y McGlashan, 2004), estándar del W3C que mapea la notación Augmented Backus-Naur a XML. Las modificaciones sobre la DTD son mínimas (se introduce un único elemento), con lo que se mantiene un alto grado de estandarización (Marrero et al., 2010b).

Como resultado, el modelo diseñado resulta en conjunto más potente, flexible y fácilmente editable que los existentes, y la orientación Web del estándar SRGS supone una ventaja adicional. La sintaxis XML facilita la transferencia y almacenamiento de patrones en este medio, mientras que el uso de identificadores uniformes de recursos permite la creación de repositorios distribuidos de funciones y patrones completos o parciales.

2.2 Método de Generación

Se diseña un método ad-hoc para la generación semi-automática de patrones IEG-SRGS sin usar corpus anotados previos. A diferencia de otros métodos en el estado del arte, es lo suficientemente flexible para incorporar cualquier atributo, con el límite de que han de ser atributos de granularidad token (e.g. lema de un token, longitud en caracteres, etc.). Adicionalmente es capaz de combinarlos con atributos predefinidos de granularidad carácter, que son introducidos de forma automática.

La combinación de ambos tipos de atributo genera patrones de tipo IEG, aunque el método no es exclusivo de éstos. Por ejemplo, los atributos de granularidad carácter y token pueden representarse en dos autómatas distintos mediante autómatas finitos en cascada, mientras que el uso tan sólo de atributos de granularidad carácter lleva a la generación de expresiones regulares estándar. En todos los casos la potencia expresiva de los patrones generados es la de las expresiones regulares. independientes del contexto.

El método utiliza técnicas de aprendizaje activo para determinar si los patrones generados son válidos. La unidad de anotación utilizada es la propia entidad, es decir, el sistema selecciona potenciales entidades de un corpus no anotado que el usuario deberá validar. Esto supone una ventaja adicional frente a la aplicación de estas técnicas en otros trabajos de NER, donde lo habitual es pedirle al usuario la anotación de documentos completos. De este modo se independiza el método de generación de las características documentales del corpus.

3 Evaluación y Resultados

Para comprobar la relación efectividad/coste del método diseñado se llevan a cabo diversos experimentos sobre el Seminar Announcement Corpus (SAC) y el Software Jobs Corpus[1] (SJC), que tienen 4 y 17 tipos diferentes de entidades anotadas respectivamente (e.g. ponentes de las conferencias, compañía ofertante de empleo, años de experiencia exigidos, etc.). Los resultados son comparados con los del trabajo de Li, Bontcheva, y Cunningham (2008). En él se muestran los mejores resultados conocidos sin utilizar corpus anotados de partida, utilizando además diversas unidades de anotación: entidades, fragmentos predefinidos de 11 tokens y documentos completos.

En la generación de patrones para el reconocimiento de entidades se obtienen resultados similares sobre SAC, sin embargo sobre SJC, con mayor variedad y cantidad de entidades anotadas, ante un mismo coste de anotación se mejora la efectividad entre un 36 % y un 62 %, mayor cuanto mayor es el número de entidades aportadas de partida. Se observan mejoras incluso al comparar estos resultados con los obtenidos por el trabajo de referencia al utilizar unidades de anotación mayores. Concretamente el método propuesto en esta tesis supone un ahorro en costes de anotación del 54 % al utilizar 10 entidades de partida si lo comparamos con el trabajo de referencia al utilizar el documento como unidad de anotación y dos documentos anotados de partida.

[1] http://www.isi.edu/info-agents/RISE/

Experimentos similares prueban también la utilidad del método en la generación de expresiones regulares sobre entidades que responden a patrones de granularidad carácter, lográndose tasas superiores a 0.8 de *F-measure* con 3 entidades de partida y un coste de anotación de entre 20 y 50 entidades.

Por último, se prueba la utilidad del método en la anotación de recursos. Para ello se suma al coste de anotación durante el proceso de generación de patrones el coste de revisar las entidades capturadas por esos patrones, de manera que el corpus quede perfectamente anotado. De este modo se consigue anotar el 70 % de media de cualquier tipo de entidad en el SAC con 50 entidades de partida y un coste equivalente al 12 % del corpus. Para el SJC estas tasas aumentan al 80 % con un coste de tan sólo el 6 % del corpus.

Adicionalmente se observa que los resultados obtenidos en las diferentes ejecuciones de los experimentos son estables, las tasas de *precision* y *recall* son equilibradas, y que el método es capaz de obtener resultados satisfactorios comenzando el proceso de aprendizaje con un pequeño conjunto de entidades de partida. Una ventaja adicional es que estas entidades pueden no encontrarse en el corpus no anotado, y los experimentos realizados muestran que el método resulta ser robusto frente a la posible introducción de entidades con patrones poco frecuentes respecto al conjunto (i.e. *datos atípicos*).

4 Conclusiones y Trabajos Futuros

La tesis profundiza en el concepto de Entidad Nombrada y muestra que NER no sólo no puede considerarse un área resuelta, sino que necesita técnicas más dinámicas ante la variedad de tipos de entidad y dominios que se exigen actualmente. Se presenta una alternativa que facilita la adaptación de los sistemas NER a nuevas semánticas, donde además es habitual no contar con corpus anotados. El método logra mejorar la relación coste/efectividad frente a métodos similares del estado del arte.

Como trabajos futuros se plantean, entre otros, su aplicación a diversos idiomas y su adaptación para la generación de patrones basados en autómatas finitos en cascada de uso extendido en Extracción de Información, como es el caso de JAPE, incluido en el framework GATE (Cunningham et al., 2012).

Bibliografía

Cunningham, H., 2005. *Information extraction, automatic*, páginas 665–677. Elsevier, 2 edición.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M.A. Greenwood, H. Saggion, J. Petrak, Y. Li, y W. Peters. 2012. *Text Processing with GATE (Version 7.1)*. The University of Sheffield.

Grishman, R. y B. Sundheim. 1996. Message Understanding Conference-6: a brief history. En *Conference on Computational Linguistics*, páginas 466–471.

Hunt, A. y S. McGlashan. 2004. Speech Recognition Grammar Specification (v. 1.0). Informe técnico, World Wide Web Consortium (W3C).

Li, Y., K. Bontcheva, y H. Cunningham. 2008. Adapting SVM for data sparseness and imbalance: a case study in information extraction. *Natural Language Engineering*, 15(2):241–271.

Marrero, M., S. Sánchez-Cuadrado, J. Morato, y Y. Andreadakis. 2009. Evaluation of named entity extraction systems. *Research in Computing Science*, 41:47–58.

Marrero, M., S. Sánchez-Cuadrado, J. Urbano, J. Morato, y J.A. Moreiro. 2010a. Tratamiento léxico en los sistemas de recuperación y representación de información en biomedicina. *El Profesional de la Información*, 19(3):246–254.

Marrero, M., J. Urbano, J. Morato, y S. Sánchez-Cuadrado. 2010b. On the definition of patterns for semantic annotation. En *CIKM Workshop on Exploiting Semantic Annotations in Information Retrieval*, páginas 15–16. ACM.

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

Sang, E.F.T.K. y F. Meulder. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. En *Conference on Natural Language Learning*, páginas 142–147.