Procesamiento de Lenguaje Natural, Revista nº 48 marzo de 2012, pp 123-126 recibido 16-01-12 revisado 06-03-12 aceptado 18-03-12

Metodología para la población automática de ontologías. Aplicación en los dominios de medicina y turismo.

A methodology for populating ontologies. Application in medicine and tourism domains.

Juana María Ruiz-Martínez
Dto. de Ingeniería de la Información y las Comunicaciones.
Facultad de Informática.
Universidad de Murcia
jmruymar@um.es

Resumen: Tesis doctoral en Documentación, con mención de doctorado europeo, realizada por Juana María Ruiz Martínez en la Universidad de Murcia bajo la dirección de los doctores Rafael Valencia García y Rodrigo Martínez Béjar . La defensa tuvo lugar el 7 de Febrero de 2012 ante el tribunal formado por los doctores Piedad Fernández Toledo (Universidad de Murcia), Alessandro Gasparetto (Università di Udine), Jesualdo Tomás Fernández Breis (Universidad de Murcia), Juan Miguel Gómez Berbís (Universidad Carlos III de Madrid) y Ricardo Colomo Palacios (Universidad Carlos III de Madrid). La calificación obtenida fue Sobresaliente *Cum Laude* por unanimidad.
Palabras clave: Población de ontologías, instanciación de ontologías, extracción de información

Abstract: Phd thesis in Information Sciences, with European doctorate mention, written by Juana María Ruiz Martínez at the University of Murcia under the supervision of Dr. Rafael Valencia García and Dr. Rodrigo Martínez Béjar. The author was examined on 7th February by the committee formed by Dr. Piedad Fernández Toledo (University of Murcia), Dr. Alessandro Gasparetto (University of Udine), Dr. Jesualdo Tomás Fernández Breis (University of Murcia), Dr. Juan Miguel Gómez Berbís (University Carlos III de Madrid) and Dr. Ricardo Colomo Palacios (Universidad Carlos III de Madrid). The grade obtained was sobresaliente *Cum Laude*
Keywords: Ontology population, information extraction

1 Introducción

La web semántica y el procesamiento del lenguaje natural (PLN) son áreas de investigación en las que convergen, entre otras ramas del conocimiento, la ingeniería informática, la lingüística y la documentación. Esta tesis, fruto de la colaboración multidisciplinar, afronta desde diversas perspectivas, el problema de la instanciación automática de ontologías a partir de texto en lenguaje natural.

La creación de ontologías y su mantenimiento de la forma más automatizada posible es uno de los principales retos a los que se ha enfrentado desde su nacimiento la web semántica.

La construcción automática de ontologías engloba la creación (semi-) automática de conceptos, relaciones, instancias y atributos ontológicos. Por su parte la población o instanciación (semi-)automática se centra en enriquecer, mediante instancias de las clases y/o de las relaciones, una ontología ya existente. La importancia de la población de ontologías reside fundamentalmente en el hecho de que (1) las ontologías existentes necesitan de actualizaciones periódicas (2) el tejido ontológico es lo suficientemente amplio como para dar cobertura a numerosos dominios especializados y, la instanciación de las ontologías que lo conforman supone un salto cualitativo en tareas como por ejemplo, la recuperación de información. Las metodologías

ISSN 1135-5948

de instanciación de ontologías que encontramos en la literatura combinan técnicas de PLN, tales como extracción y reconocimiento de patrones lingüísticos, POS-tagger y análisis sintáctico (Maynard et al. 2009; Amardeilh et al., 2006; McDowell & Cafarella, 2008; Navigli & Velardi, 2006), junto con otras técnicas de aprendizaje automático (Tanev & Magnini, 2008; Giuliano & Gliozzo, 2008).

2 Contribuciones

Las dos metodologías propuestas en esta tesis abordan la instanciación automática de ontologías desde un punto de vista que combina el análisis lingüístico tradicional y tecnologías para la extracción de conocimiento textual. Así mismo, se recurre a la combinación de recursos lingüísticos y ontológicos ya desarrollados, y ampliamente difundidos entre la comunidad científica, que permiten llevar a cabo el proceso de instanciación. El ensamblaje de estos recursos en las metodologías propuestas, garantiza buenos resultados en tareas previas a las instanciación como por ejemplo, el reconocimiento y clasificación de entidades nombradas.

La lingüística como ciencia del lenguaje se ha desarrollado durante siglos. Ciertamente, los métodos estadísticos y computacionales son necesarios para el desarrollo de herramientas de PLN, pero estos métodos serán más eficaces si consideran las características lingüísticas de los textos objeto del procesamiento, siendo ésta una de las principales contribuciones de la tesis.

Las metodologías propuestas son modulares, es decir, están divididas en fases y en ellas se reutilizan algunos recursos ya desarrollados y disponibles gratuitamente, de amplia difusión, lo que facilita su adaptación a diversos escenarios de aplicación.

2.1 Metodología basada en la distancia cotextual y la ganancia de conocimiento

La primera metodología se basa en la distancia cotextual, como elemento lingüístico, y en la ganancia de conocimiento, como elemento ontológico.

La distancia cotextual se refiere a la distancia física que existe entre dos unidades lingüísticas en el texto. En cuanto a la ganancia de conocimiento es una medida que hace referencia al conocimiento cuantitativo adquirido por el sistema y susceptible de ser incluido en a ontología, es decir, a mayor cantidad de conocimiento adquirido mayor ganancia de conocimiento.

El punto de partida de esta metodología son dos corpora, relativos al dominio del turismo, que se han analizado desde un punto de vista discursivo siguiendo los parámetros propuestos por Bhatia (1993). Los datos obtenidos se han interpretado, de modo que la información que arroja el análisis se ha utilizado para (1) determinar cuáles son las instancias relevantes del dominio (2) determinar cómo dichas instancias aparecen representadas textualmente (3) desarrollar los componentes lingüísticos que permitan la extracción de las mismas y (4) ajustar los parámetros cotextuales de la metodología de instanciación.

Además, se ha desarrollado una ontología de turismo en la que se insertan las instancias.

La metodología de instanciación se divide en cuatro fases secuenciales:

1. Fase de Procesamiento de Lenguaje Natural y de Procesamiento del Corpus.
2. Fase de Reconocimiento e identificación de las Entidades Nombradas.
3. Fase de Población de la Ontología.
4. Verificación de la consistencia de la ontología.

En la primera fase se obtiene la estructura morfosintáctica de cada oración en el corpus con el objetivo de extraer la información lingüística necesaria para las siguientes fases. Para ello, se ha utilizado el framework GATE[1].

Durante la segunda fase se extraen un conjunto de menciones de entidades nombradas. Con este propósito, se han desarrollado listas de gazetteers de carácter general, como por ejemplo localizaciones, código postal o apellidos, así como de carácter específico, como tipos de comida o servicios. Estas listas, combinadas con reglas basadas en expresiones regulares, permiten la anotación de entidades nombradas. Cuanto mayor sea el número de anotaciones en esta fase, mayor será el número potencial de instancias de la ontología, ya que, en la siguiente fase, dichas menciones serán candidatas a instancias o a valores de propiedades de la ontología.

Durante la fase de instanciación de la ontología, el sistema recibe como entrada el conjunto de anotaciones de la fase anterior, convirtiéndose éstas en candidatas a instancias.

[1] http://gate.ac.uk/

Sin embargo, durante el proceso de anotación pueden surgir ambigüedades, como por ejemplo que una anotación o grupo de anotaciones esté relacionado con más de una entidad nombrada o que una entidad esté relacionada con varios recursos de la ontología. En estos casos el sistema procederá a la desambiguación. Para ello, se basa en la distancia que las separa en el texto (distancia cotextual), por un lado, y en la cantidad de conocimiento que aportan a la ontología, por otro.

La ganancia de conocimiento se fundamenta en el hecho de que una instancia no aparece aislada en el discurso, sino que en el texto circundante, es decir, en el cotexto, se pueden localizar otros elementos que aporten información sobre la misma. Esta información es susceptible tanto de ser incluida en la ontología como de ser utilizada para la desambiguación de dicha instancia. La desambiguación se lleva a cabo creando un árbol combinatorio con todas las clases y propiedades que podrían ser instanciadas en función de la instancia ambigua y aquellas más cercanas en el texto con las que se podría relacionar. Finalmente se elige la combinación que aporta mayor conocimiento a la ontología.

Una vez desambiguadas las anotaciones se insertan en la ontología como un individuos de una o más clases. Así mismo, se identifican los valores de sus atributos y relaciones. Si una instancia no había sido insertada previamente, entonces el sistema instancia la ontología con la información extraída, creando así un nuevo individuo. Si por el contrario, una instancia ya existía ésta se enriquece con nuevas relaciones y atributos, en el caso de que hayan sido identificados. Finalmente, en la última fase, se comprueba la consistencia de la ontología mediante un razonador OWL-DL para detectar individuos erróneos y que no cumplan las restricciones de la ontología.

La evaluación se ha llevado a cabo considerando por un lado las entidades anotadas por el sistema y por otro, la cantidad de instancias que se han incluido en la ontología. El número de anotaciones de entidades nombradas correctamente identificado por GATE representa el 99% del total de entidades nombradas anotadas. Esta información se refiere a los grupos que contienen una sola anotación para entidad nombrada dada y, en consecuencia, no presentan ambigüedades, lo que representa la mayoría de los casos. Por el contrario, de entre las anotaciones que presentaban alguna ambigüedad fueron desambiguadas el 66,4%. En cuanto a los resultados de la instanciación, se han utilizado las medidas estándar de exhaustividad, precisión y medida de F, para medir el número de instancias, object properties y datatype properties obtenidas de los corpora. Los resultados globales arrojan una medida de F del 89.51% para uno de los corpus y de 86.85% para el otro.

2.2 Metodología basada en roles semánticos

La integración de diversos recursos ontológicos y lingüísticos es la base de esta metodología en la que se combinan ontologías de alto nivel del dominio biomédico con *frames* semánticos.

Dada la disponibilidad de distintos recursos y herramientas para el PLN cuyo uso está extendido entre la comunidad científica, aunque generalmente no de manera combinada, se ha procedido al diseño de un marco de trabajo en el que se integran dichos recursos.

Estos recursos son por un lado, la ontología de relaciones de OBO (Smith et al., 2005) y la ontología BioTop (Beisswanger et al., 2008) y, por otro, los *frames* de FrameNet (Baker et al., 1998).

OBO y BioTop son dos ontologías de alto nivel o top level ontologies en las que se expresan las relaciones genéricas del domino biomédico. Se han seleccionado algunas de las relaciones de las ontologías y se han definido un conjunto de axiomas. La ontología resultante se ha mapeado con *frames* o marcos semánticos extraídos de FrameNet mediante la asociación de uno o más *frames* a cada relación. Los *frames* son representaciones esquematizadas de situaciones del mundo real en base a las cuales se organiza la información. Las relaciones ontológicas y los *frames* tienen en común que son generalizaciones de situaciones discursivas cuya expresión lingüística consiste fundamentalmente en una forma verbal o su nominalización. En consecuencia, aquellas relaciones ontológicas que tienen asociado un *frame*, contienen tanto expresiones lingüísticas, que representan dichas relaciones a un nivel textual, como unos roles semánticos que describen cada relación. El modelo ontológico resultante es BioOntoVerb OM.

El proceso de instanciación se lleva a cabo, por un lado, mediante la identificación de entidades nombradas con el reconocedor de entidades nombradas de GENIA (Tsuruoka et al., 2005), convirtiéndose las mismas en candidatas a instancias de la ontología, y por otro lado, mediante la identificación de relaciones entre ellas para lo que se utiliza BioOntoVerb OM.

Cuando dos entidades nombradas se relacionan a través de las unidades léxicas incluidas en BioOntoVerb OM, dicha relación se incluye provisionalmente como instancia de una ObjectProperty, y las instancias implicadas en la misma se convierten en candidatas a instancias de la ontología.

Un *frame* conecta las dos entidades más cercanas en el texto siempre y cuando cumplan con ciertos requisitos, como por ejemplo, la distancia que existe entre ellas.

Finalmente un razonador comprueba la consistencia de la ontología, de modo que si la ontología es consistente, las clases y propiedades correspondientes son instanciadas. Además mediante el razonador se pueden obtener nuevas relaciones en base a los axiomas previamente definidos en el modelo ontológico.

La validación del sistema se ha llevado a cabo mapeando el modelo ontológico con una ontología de dominio, de manera que tanto los marcos semánticos como los axiomas definidos en el modelo pasan a formar parte de la ontología de dominio.

De nuevo aquí se evalúan dos aspectos del proceso de instanciación, por un lado las entidades nombradas identificadas y por otro la cantidad de relaciones ontológicas que se han establecido. Las medidas utilizadas son la exhaustividad, precisión y medida de F.

Dado que el prototipo que se ha diseñado utiliza el identificador de entidades de GENIA y la ontología de dominio instanciada está basada también en el corpus GENIA, las medidas de exhaustividad y precisión obtenidas en el proceso de identificación y clasificación son de casi del 97%.

En cuanto al conocimiento instanciado, el prototipo diseñado ha alcanzado una precisión total de 79.59%, una exhaustividad del 69.02% y una medida de F de 73,93%. Estos valores son altos, debido a que el experimento se ha llevado a cabo en un dominio específico, el de la biomedicina, y los roles semánticos seleccionados están estrechamente relacionados con los verbos y otras unidades léxicas que indican las relaciones en dicho dominio.

Bibliografía

Amardeilh, F. 2006. OntoPop or how to annotate documents and populate ontologies from texts. *Proceedings of the Workshop on Mastering the Gap: From Information Extraction to Semantic Representation* (ESWC'06), Budva, Montenegro.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. 1998. The berkeley FrameNet Project. *Proceedings of COLING/ACL-98*, pp. 86-90.

Beisswanger, E., Schulz, S., Stenzhorn, H., & Hahn, U. 2008. BioTop: An upper domain ontology for the life sciences. A description of its current structure, contents and interfaces to OBO ontologies. *Applied Ontology, 3*(4), pp. 205-212.

Bhatia, V. K. 1993. *Analysing genre. language use in professional settings*. London: Longman.

Giuliano, C., & Gliozzo, A. 2008. Instance based lexical entailment for ontology population. En *Proceedings of* EMNLP-CoNLL, pp.265-272.

Maynard, D., Funk, A., & Peters, W. 2009. SPRAT: A tool for automatic semantic pattern-based ontology population. *International Conference for Digital Libraries and the Semantic Web,* Trento, Italy.

McDowell, L. K., & Cafarella, M. 2008. Ontology-driven, unsupervised instance population. *Web Semantics: Science, Services and Agents on the World Wide Web, 6 (3)*, pp. 218-236.

Navigli, R., & Velardi, P. 2006. Enriching a formal ontology with a thesaurus: An application in the cultural heritage domain. *En Proceedings* of *COLING• ACL 2006,* Sydney.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Rosse, C. 2005. Relations in biomedical ontologies. *Genome Biology, 6*(5), R46.

Tanev, H., & Magnini, B. 2006. Weakly supervised approaches for ontology population. *Proceedings of EACL-2006, Trento,* pp. 3-7.

Tsuruoka, Y., Tateishi, Y., Kim, J. D., Ohta, T., McNaught, J., Ananiadou, S., & Tsujii, J. 2005. Developing a robust part-of-speech tagger for biomedical text. *Advances in Informatics,* pp. 382-392.