ARTÍCULO ACEPTADO

Semantic Analyzer for Spanish, Using Ontologies

por Alma Delia Cuevas Rasgado y Yedid Erandini Niño Membrillo

Resumen

Este artículo presenta el diseño e implementación de un analizador semántico que identifica la cercanía semántica entre las palabras de una oración usando una ontología. Una definición común de ontología es una especificación explícita y formal de una conceptualización compartida. Esta ontología puede ser una herramienta importante al representar un texto descriptivo y analizar textos en español.

El analizador semántico extrae datos de una ontología de conocimiento común que consta de relaciones, conceptos y valores almacenados con base en su significado. La ontología básica comienza con datos cualitativos alimentados manualmente por un usuario. El analizador responde "verdadero" si una oración está semánticamente relacionada o "falso" si no lo es. Su utilidad radica en hacer preguntas en lenguaje natural a una biblioteca digital y para que los robots de servicios entiendan el significado de la acción a realizar. Se ha probado su eficiencia usando el corpus CONLL v.2009 obteniéndose una buena precisión en sus resultados.

Introducción

Analizar un texto de algún tema particular en lenguaje natural y convertirlo a una estructura entendible por las computadoras, no es tarea trivial. En parte porque en el proceso de conversión se debe comprobar la gramática, semántica, y la pragmática, al menos. Éste es uno de los retos de este proyecto.

Con el abaratamiento de las computadoras y el acceso a una gran cantidad de documentos (por ejemplo, en Internet), resulta posible (y deseable) utilizar una computadora para analizar información relevante en documentos que están escritos en lenguaje natural, específicamente en Español. Los analizadores actuales utilizan principalmente la sintaxis, por ejemplo., atienden a las reglas gramaticales del lenguaje en español, pero no toman en cuenta su semántica.

Por esta razón, nos hemos dado a la tarea de construir analizadores que usen tanto la información gramatical como la información semántica. Otro reto es la extracción automática de conocimiento en textos. Las herramientas necesarias para esta conversión son etiquetadores, analizadores sintácticos, morfológicos y semánticos.

El lenguaje natural tiene como característica inherente la ambigüedad, que en los humanos se resuelve usando el sentido común; sin embargo, estás no son parte de las computadoras. Por ejemplo, “la puerta es amarilla”, en este caso “puerta” se puede referir a la puerta de un carro o a la puerta de una casa, pero ¿cómo puede saber esto la computadora? En este artículo se muestra una herramienta de análisis semántico que permite a la computadora resolver estos problemas.

En este documento describimos el analizador semántico cuya operación inicia analizando tres entradas: Relación, Concepto y Valor. Por ejemplo en la frase: “puerta con material de plástico” existen los siguientes elementos que nos interesan: Concepto (puerta), Relación (material) y Valor (plástico), como se observa en la frase se han ignorado las preposiciones, también otros vocablos como artículos, adjetivos, adverbios, dando preferencia a los sustantivos, verbos y nombres propios.

Un analizador sintáctico y morfológico como Freeling [1] hace un reconocimiento limitado al número de sus reglas gramaticales y por tanto, no utiliza el conocimiento semántico.

El idioma español también sufre de ambigüedad (palabras que morfológicamente son iguales, pero semánticamente diferentes), por ejemplo "Toma el gato que está dentro del carro … " en español “gato” es un animal o “gato” es una palanca mecánica para cambiar un neumático, ambos conceptos pueden estar relacionados con auto, otro ejemplo que tiene ambigüedad en español y en inglés: “Juan es un alto oficial” puede significar que Juan es alto, o que él es un oficial. Otros trabajos [2], [3] resuelven esto por proximidad semántica.

En este artículo se usan marcos semánticos en la solución [4], que representan mejor el significado de un concepto. Un marco semántico, es un escenario, una descripción de un objeto, proceso, acción o situación. Por ejemplo, el automóvil es un marco que describe el escenario en el cual ocurren los eventos (viaje, conducción, descanso, carretera, casetas de cobro, autopista, calle, etc.) y sus partes (asiento, pistón, gato neumático, palanca mecánica, motor, motores de combustión interna, etc.). La ontología se construye con marcos, para mantener las relaciones semánticas entre dos conceptos. Algunos proyectos como (FrameNet, [5]) los utilizan.

La ontología consiste en relaciones o enlaces (que representan acciones, típicamente verbos y propiedades) entre nodos (que representan objetos, típicamente sustantivos). Ambos nodos y enlaces representan

conceptos. Usamos la notación OM [6] para representar la ontología, ver Tabla 1, porque sus relaciones también pueden ser nodos. OM es una notación de manera similar a XML, vea la Figura 1. Para obtener una definición más formal de la ontología, consultar [6].

La creación de marcos semánticos pretende guiar al analizador para que provea la semántica de las palabras y otros elementos como: sinonimia, hiperonimia, meronimia, herencia de propiedades y la glosa o descripción del concepto.

La figura 1 define el concepto: “acción”. Cada concepto es un nodo en la ontología, un concepto es conocido como una clase o conjunto de cosas del mismo tipo. Un concepto puede tener descendientes o subconjuntos (conceptos que dependen de éste). Por ejemplo, el subconjunto "acción" contiene al subconjunto "tiene". Quiere decir que “tener” es una “acción”. Luego el subconjunto "acción" está contenido en “cosa”. En la ontología cosa se puede dividir en cosas concretas y cosas abstractas, por ejemplo, las acciones, emociones, procesos mentales, software son cosas abstractas, mientras que los objetos palpables, hardware, etc. son cosas concretas. Las relaciones son las propiedades del conjunto en la cual se encuentran, por ejemplo, “carro” tiene como propiedad: “color”.

En la notación OM para representar ontologías no solo representa una jerarquía de conceptos sino incluye sus relaciones, por ejemplo, “carro” se relaciona con “turista” y “conductor”. Esta relación semántica no es mostrada en WordNet. El modulo semántico reconoce relaciones semánticas y taxonómicas.

En la notación OM, los sinónimos aparecen dentro de la etiqueta <language> que define el idioma en que están definidos los conceptos, luego dentro de <word>. Por lo tanto, “*tiene*” y “*cuenta con*” denotan el mismo concepto (son sinónimos), véase la figura 1.

```
<Language>Spanish<word>Cosa, algo</word> </Language>
<concept>acción
<Language>Spanish<word>acción</word></Language>
        <subset> Cosa</subset>
        <concept>tiene
            <Language>Spanish<word>tiene, cuenta con
            <subset> acción</subset>
        </concept>
</concept>
```

Figura 1. Fragmento de ontología en notación OM

El analizador semántico tiene una manera heurística de hallar la cercanía semántica de los conceptos usando marcos semánticos.

Un objetivo importante de nuestra semántica es combinarla con Freeling (u otro analizador) para obtener un análisis más preciso de las frases. La organización de este documento es el siguiente: Primero, presentamos el marco teórico. Después, describimos el algoritmo del analizador. Posteriormente se muestran los resultados. Se finaliza con las conclusiones y las referencias.

Marco teórico

La característica central de un sistema es que puede ser considerado inteligente desde una perspectiva humana. Para Turing, es la habilidad de comunicarse con las personas en su propio lenguaje [7].Un aspecto clave a considerar es que la mayoría del conocimiento humano está escrito en lenguaje natural, pero hacer que todo este conocimiento sea introducido en una computadora y sea consultado por el humano de forma natural es necesario construir programas que utilicen las mismas habilidades y los principios humanos, a través del uso de herramientas de inteligencia artificial [8].

Procesamiento del lenguaje natural

Para entender el significado de un texto se usa el procesamiento del lenguaje natural. En este contexto se presentan al menos tres problemas: 1) ambigüedad en sustantivos, verbos y preposiciones, 2) resolución de anáforas pronominales y 3) relaciones gramaticales. En este artículo se trata el punto 3.

Representación del conocimiento

La representación del conocimiento es un área de la Inteligencia artificial cuyo principal objetivo es facilitar la búsqueda del conocimiento y realizar deducciones. La idea es que la computadora reciba frases y sentencias a través de un texto y pueda ser “entendida” por la computadora. Es decir, encontrando sus relaciones morfológicas, sintácticas y semánticas tanto como las personas lo hacen. [9]. Por lo tanto, la representación conceptual del conocimiento es relevante porque permite hacer deducciones desde un concepto. Hay varias técnicas para representar el conocimiento tales como marcos semánticos, reglas, redes semánticas y ontologías.

Trabajos relacionados

El análisis semántico está enfocado en la solución de varios problemas del procesamiento del lenguaje natural. Estas tareas incluyen desambiguaciones del sentido de la palabra (WSD), desambiguador de preposiciones, resolución de anáforas y extracción de relaciones entre palabras [10]. De esta forma, la construcción de un analizador semántico completo es muy complicado, ya que debe resolver problemas léxicos y reglas sintácticas, que requieren conocimiento del contexto.

Para el idioma inglés, hay muchos más trabajos, que para el español; en Panchenko [11] se presenta un método para extraer relaciones semánticas entre palabras usando el algoritmo K vecino más cercano (KNN) y medidas de similitud semántica, sobre resúmenes de Wikipedia. El procedimiento recibe un conjunto de términos como: {caimán, animal, edificio, casa, teléfono} y genera salidas que muestren las posibles

relaciones, que existen, en ese caso: {{caimán, animal}, {edificio, casa}}.

Melcuk en [12] describe un procedimiento automático para identificar patrones léxicos que representen relaciones semánticas entre conceptos en una enciclopedia en línea. Algunos conceptos de las entradas existentes son extraídos de Wikipedia y buscados en WordNet para determinar su significado. Posteriormente, se buscan otras palabras dentro de estas entradas, para relacionarlas al concepto y de esta manera, crear un contexto más amplio. Se crea un patrón de contexto. Una vez creado, busca estos patrones en otras entradas de Wikipedia, intentando generalizar. Finalmente, se verifica si los patrones generalizados contienen relaciones que inicialmente no existían en WordNet y, de ser así, se agregan.

Agrawal y Kakde en [13] describen un método de análisis semántico de consultas en lenguaje natural utilizando una ontología del dominio de trenes. La estructura de la ontología contiene conceptos que describen el conjunto de propiedades asociadas con un concepto, el cual al estar enlazado con otros conceptos, define relaciones entre ellos. La ontología tiene tres conceptos principales: objetos (por ejemplo: tren); mental o abstracto (concesión) y social (departamento). También tiene eventos que son acciones o situaciones que ocurren en un periodo de tiempo y propiedades usadas para definir conceptos de la ontología. Se usa el método de tripleta de palabras para obtener la semántica utilizada en la ontología, al igual que como proponemos en este artículo.

En Thakor y Sasi [14] se describe un enfoque para construir un modelo de ontología que identifique problemas asociados con la insatisfacción de los clientes en divisiones o áreas del servicio postal. Los sustantivos son considerados como objetos y los verbos como propiedades de objetos. Las clases, objetos y las propiedades de los objetos se usan para construir un modelo de ontología. Los resultados del análisis podrían ser usados por la compañía para tomar medidas correctivas para resolver problemas. No obstante, el modelo necesita ser refinado, así como los objetos, las clases y las propiedades de los objetos, mismos que requieren irse actualizando con nueva información.

Hay pocos estudios que consideran la semántica como un recurso importante para estructurar la información. La mayoría de ellos trabajan solo con sintaxis.

Metodología

En este apartado se presenta un método del análisis semántico dado en forma de una tripleta: R(C,V) donde: R: relación, C: Concepto, V: valor, que es enviada a la ontología para buscar su significado. Este significado puede ocurrir de en diferentes formas:

1. *Puede buscar el valor de la relación de un concepto*,

Por ejemplo "Color(Carro de Juan, nulo)" donde la relación "Color" es asociado con nulo. Nulo significa que hay un espacio en la tripleta, que no tiene datos y esos datos se van a obtener con la búsqueda del analizador.

Al ser hallado el concepto Carro de Juan en la ontología nulo puede casar con el valor "amarillo" (amarillo en este caso es el color del Carro de Juan).

2. *Puede buscar el concepto a partir de sus propiedades*

La tripleta: "cuenta con(Carro, puerta)", en esta tripleta no existe un nulo, quiere decir que el analizador no devolverá un dato sino verificará si hay relación semántica entre los elementos de la tripleta.

Después de buscar el concepto "Carro", el analizador verifica sus propiedades y obtiene que la relación "cuenta con" no se halla en la ontología pero existe una relación "tiene", ahora verifica si estos vocablos ("tiene" y "cuenta con") tienen alguna correspondencia. Al comprobar la relación "tiene" se identifica que su sinónimo es "cuenta con" por lo tanto casa con la relación "cuenta con" de la tripleta, lo mismo pasa con "puerta" y "portezuela", al final devuelve verdadero, véase la figura 2.

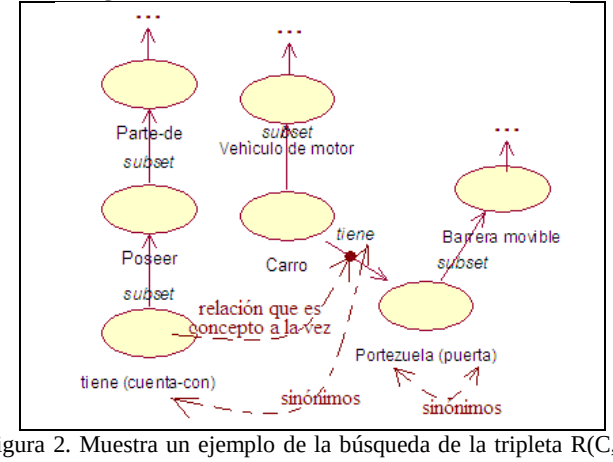

Figura 2. Muestra un ejemplo de la búsqueda de la tripleta R(C,V) donde R=cuenta con, C=Carro, V=puerta. En la ontología no aparecen "cuenta con" y "puerta" sino "tiene" y "portezuela". En analizador identifica los sinónimos y reconoce la relación semántica entre estos.

3. *Puede buscar el concepto a través de la herencia de propiedades*

Por ejemplo, la tripleta a buscar es: tiene (carro, rueda), el analizador no halla la tripleta indicada pero busca entre los antecesores de "carro" y encuentra que: tiene (transporte, rueda) es decir, la propiedad no está asociada con el concepto directamente sino con el concepto antecesor o hiperónimo. Si el analizador sigue la siguiente propiedad: si "carro es subconjunto de "transporte" y "transporte" tiene "ruedas" entonces "carro" tiene "ruedas", nos referimos a la propiedad transitiva.

4. *Puede buscar la glosa del concepto*

Por ejemplo: nulo (carro, nulo). En la tripleta solo se tiene el concepto "carro", no se define la relación ni el valor a buscar, el analizador no tiene forma de buscar entre las relaciones, tampoco entre los valores de sus relaciones por lo tanto devuelve al usuario la glosa del concepto "carro".

Nótese que en todos los casos, se pueden obtener las semánticas de lo que se busca, siempre y cuando los conceptos se encuentren en la ontología del conocimiento común y que además estén definidos sus marcos.

La figura 3 muestra el funcionamiento general del analizador semántico.

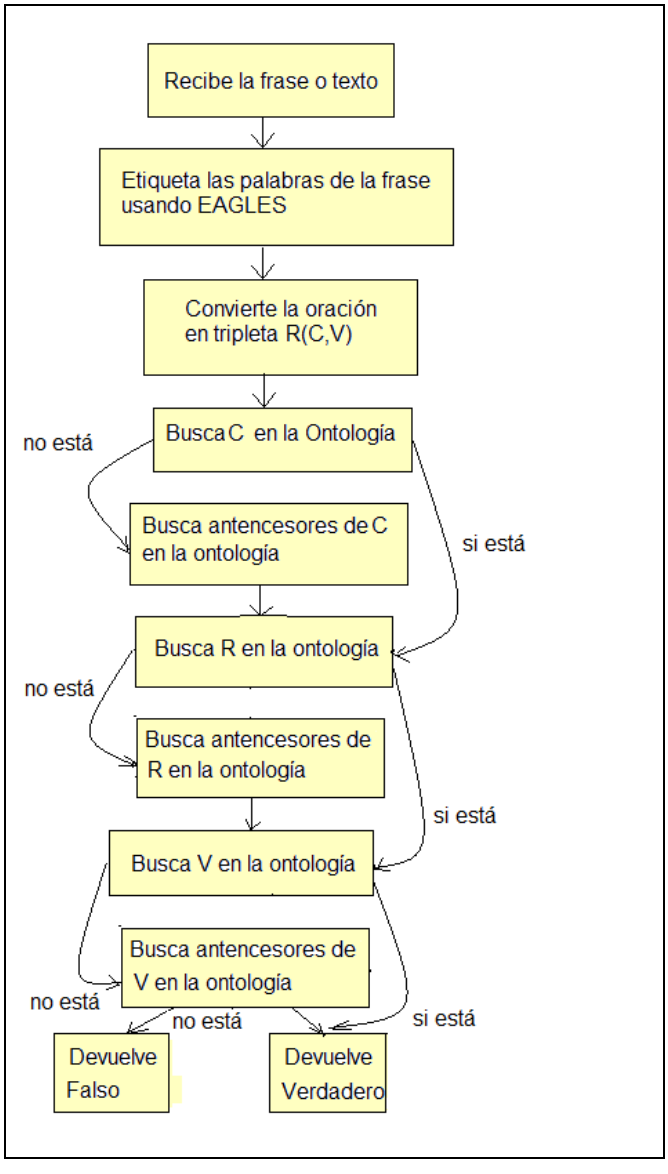

Figura 3. Diagrama general del analizador semántico, en la cual la tripleta está identificada como R(C,V), donde: R es la relación o enlace, C es el concepto y V es el valor de la relación, por ejemplo: cuenta con(Carro, Llantas), R=cuenta con, C=Carro, V=Llantas

Pruebas y Resultados

Para confirmar el funcionamiento del algoritmo, se realizaron un conjunto de pruebas tomando como datos de entrada R (C, V). Se usó el corpus CONLL (Conference on Computational Natural Learning) este es el resultado de un esfuerzo entre grupos de trabajo, durante 5 años. Estos grupos tienen el propósito de impulsar aplicaciones de procesamiento de lenguaje natural. El corpus versión 2009 está enfocado exclusivamente al análisis sintáctico sin embargo es complicado el análisis semántico, esta es la razón por la cual ha sido elegido para estas pruebas.

El corpus completo se puede consultar en https://www.dropbox.com/s/03p2uvutrcmgemt/CORPUS%20CONLL.pdf?dl=0.

En el corpus hay palabras como Tribunal_Supremo_de_Justicia que se reconocen como nombres propios, es decir tienen un solo significado, el etiquetado no se realiza por cada palabra sino por toda la cadena de palabras (por eso se incluye el guion bajo), al final se etiqueta como un nombre propio y se toma como sustantivo. Para evaluar la precisión del analizador semántico con el corpus, comparamos los resultados con respecto al método manual, comparamos si las tripletas halladas por el analizador resultaban ser las mismas que el método manual y que las consultas arrojaran los mismos resultados.

La cantidad de conceptos y valores de la ontología fueron: 367, la cantidad de relaciones o verbos: 249, la cantidad total de nodos en la ontología: 689 (se suman los conceptos, relaciones y conceptos que son antecedentes de los conceptos buscados en la tripleta).

Manualmente se hallaron 77 tripletas completas en el total de 689 nodos de la ontología. La precisión total del analizador si hubiese hallado las 77 fuera de 100%, pero 4 tripletas fueron incompletas, por lo que 73 tripletas halladas resulta un total de: 94% de precisión. Las 4 tripletas que se no hallaron fue porque el concepto estaba demasiado lejos (semánticamente) de la relación y el valor, de todas maneras 94% es una precisión aceptable.

El desarrollo del analizador semántico permite la realización de varias actividades tales como: la definición de una ontología (conjunto de elementos interrelacionados, como una red de neuronas en los humanos), usando marcos semánticos (definición de escenarios, contextos o dominios) y obteniendo el significado (semánticas).

Porqué decimos que el analizador es semántico? Porque los nodos en la ontología están conectados de acuerdo a su significado usando marcos semánticos y el analizador realiza la búsqueda de acuerdo a este significado.

Este resultado de analizador semántico va dirigido a sistemas que puedan incluir un módulo para responder

preguntas complejas en lenguaje natural o la interpretación de las acciones de un robot de servicios.

Observaciones Finales

Se ha diseñado un analizador semántico que identifica la cercanía semántica entre los elementos de una oración. Este no es para resolver sustantivos ambiguos, tampoco identifica anáforas, sin embargo se recomienda que sea usado después de estas herramientas.

REFERENCIAS

1 Padró, L. and Stanilovsky, E. (2012) Freeling 3.0: Toward wider multilinguality". In Proceedings of the Eight International Conference on Language Resources and Evaluation (Istambul, Turkey May 23-25, 2012). LREC'12, 2473-2479.

2 McInnes, T. and Pedersen, T. (2013) Evaluating Measures of Semantic Similarity and Relatedness to Disambiguate Terms in Biomedical Text. Journal of Biomedical Informatics. Elsiever Science. 6 (San Diego, USA, December), 1116-1124. DOI = http://dx.doi.org/10.1016/j.jbi.2013.08.008

3 Colorado, F. (2008) Mapeando palabras a conceptos. M. Sc. Thesis. In Spanish. CIC-IPN.

4 Minsky, M. (1974) A Framework for Representing Knowledge. Memo. 306MIT. AI Laboratory.

5 Collin, B. (2008) FrameNet, Present and Future. In Proceedings of the First International Conference on Global Interoperability for Language Resources. (Hong Kong: City University), 12-17.

6 Cuevas, A., and Guzmán, A. (2010) Knowledge accumulation through automatic merging of ontologies. Expert Systems with Applications, Elsevier Editorial. 37, 3. System, USA. ISSN: 0957-41741991-2005.

7 Aliseda, A. (2013) ¿Inteligencia Mecánica? La pregunta de Alan Turing. Ciencia. Revista Mexicana de la Academia de Ciencias, 64, 4. ISSN 1405-6550. In Spanish. 10-17.

8 Pino, R., Gómez, A., and de Abajo, N. (2001) Introducción a la Inteligencia Artificial: Sistemas expertos, redes neuronales artificiales y computación evolutiva. Universidad de Oviedo. ISBN 84-8317-249-6. 106.

9 Davis, R., Shrobe, R., and Szolovits, P. (1993) What is a knowledge Representation? Artificial Intelligence Magazine. 14, 1. 17-33.

10 López, Y. (2012) System for extracting and representing knowledge from descriptive texts. M. Sc. Thesis. In Spanish. CIC-IPN

11 Panchenko, A. (2013) Similarity measures for semantic relation extraction. Doctoral Thesis. Université Catholique de Louvain and Bauman Moscow State Technical University.

12 Melcuk, I. (1996) Lexical functions: A tool for the description of lexical relations in a lexicon. In Lexical functions in lexicography and natural language processing. John Benjamin, (Amsterdam, Philadelphia, 1996). 37–102.

13 Agrawal, A., and Kakde, O. (2013) Semantic Analysis of Natural Language Queries Using Domain Ontology for Information Access from Database. International Journal of Intelligent systems and Aplications. (December, 2013), 81-90. .DOI= DOI: 10.5815/ijisa.2013.12.07

14 Thankora, P. and Sasi, S. (2015). Ontology-based Sentiment Analysy Process for Social Media Content Pratik. Computer Science., 199-207. INNS Conference on Big Data, ELSEVIER.

SOBRE LOS AUTORES

Alma Delia Cuevas Rasgado Obtuvo su Maestría y Doctorado en Ciencias de la Computación del CIC-IPN (Centro de Investigación en Computación) en México en 2006. Sus líneas de investigación son ingeniería de software, calidad de software e inteligencia artificial específicamente la representación del conocimiento y fusión de ontologías. Actualmente es profesora investigadora en la Universidad Autónoma del Estado de México. Imparte cursos en las áreas de: base de datos, sistemas de información, programación y tecnologías para la web e inteligencia artificial. Es miembro de la Sociedad Mexicana de Ciencias de la Computación. Miembro de la Académica Mexicana de Computación.

Yedid Erandini Niño Membrillo recibió el grado de Maestra en Ciencias en el Colegio de Postgraduados. Trabaja en la Universidad Autónoma del Estado de México UAEM como profesora. Sus intereses actuales de investigación son aplicaciones de inteligencia artificial e ingeniería de software. Es miembro de la Sociedad Mexicana de Ciencias de la Computación.