Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 223-226 recibido 05-04-2013 revisado 16-06-2013 aceptado 21-06-2013

DIANA: Análisis del discurso para la comprensión del conocimiento

DIANA: DIscourse ANAlysis for knowledge understanding

Paolo Rosso
Universitat Politècnica de València
Camino de Vera s/n. 46022
Valencia, España
prosso@dsic.upv.es

M. Antònia Martí, Mariona Taulé
Universitat de Barcelona
Gran Via 585, 08007
Barcelona, España
{amarti, mtaule}@ub.edu

Resumen: DIANA es un proyecto coordinado en el que participan el grupo de *Ingeniería del Lenguaje Natural y Reconocimiento de Formas* (ELiRF) de la Universitat Politècnica de València y el grupo *Centre de Llenguatge i Computació* (CLiC) de la Universitat de Barcelona. Se trata de un proyecto del programa de I+D (TIN2012-38603) financiado por el Ministerio de Economía y Competitividad. Paolo Rosso coordina el proyecto DIANA y lidera el subproyecto DIANA-Applications y M. Antònia Martí lidera el subproyecto DIANA-Constructions.
Palabras clave: Semántica distribucional, construcciones, detección de ironía y de fraude en medios sociales, detección de paráfrasis y plagio

Abstract: DIANA is a coordinated Project involving the research group of *Ingeniería del Lenguaje Natural y Reconocimiento de Formas* (ELiRF) of the Universitat Politècnica de València and the research group of *Centre de Llenguatge i Computació* (CliC) of the Universitat de Barcelona. This is an R&D project (TIN2012-38603) funded by the Spanish Ministry of Economy and Competitiveness. Paolo Rosso coordinates the DIANA project and leads the subproject DIANA-Applications and M. Antònia Marti leads the DIANA-Constructions subproject.
Keywords: Distributional semantics, constructions, irony and deception detection in social media, paraphrasing and plagiarism detection

1 Introducción

La finalidad de este proyecto es avanzar en el área de la Lingüística Computacional (LC) y el Procesamiento del Lenguaje Natural (PLN) con el fin de superar las actuales limitaciones de los sistemas en una doble línea de actuación: (i) Desde la LC proponiendo y evaluando empíricamente nuevos fundamentos teóricos en la concepción de la estructura del lenguaje humano; (ii) desde el PLN desarrollando nuevas técnicas y métodos aprovechando el estado actual de los conocimientos científico-técnicos. Estas nuevas aproximaciones se aplicarán, por un lado, en tareas que impliquen el tratamiento del lenguaje en el marco del discurso (resolución de la correferencia de entidades y eventos, tratamiento de los argumentos implícitos e identificación de paráfrasis en el plagio) y, por otro lado, en el análisis y comprensión de textos subjetivos. Nuestro propósito es dar respuesta a la pregunta que se plantea Wintner (2009) en su artículo 'What Science Underlies Natural Language Engineering?' en el que expone un problema fundamental: 'What branch of science, then, underlies Natural Language Engineering? What is the theoretical infrastructure on which we build our applications? And what kind of mathematics is necessary for reasoning about human languages? (...)'.

Desde la lingüística cognitiva (Croft y Cruse, 2000), la neurociencia (Hawkins, 2004) y la psicología del desarrollo (Tomasello, 2003) se proponen alternativas a lo que constituyen los fundamentos de la estructura lingüística. De esta conjunción de factores emerge un nuevo modelo de lenguaje que toma como base las producciones lingüísticas para inferir la estructura de la lengua. Este nuevo modelo tiene el concepto de construcción como unidad

ISSN 1135-5948

básica (Goldberg, 1995) y el discurso como marco del análisis. Nuestro objetivo es complementar el conocimiento lingüístico tradicional -análisis morfológico (PoS), sintáctico y semántico- con la identificación de construcciones, unidades complejas de tipo sintagmático que corresponden de manera aproximada a las unidades que en PLN se tratan bajo el nombre de expresiones multipalabra (EM). Se entiende por construcción, la conjunción de una forma y un significado que expresan una determinada función comunicativa. Sag et al. (2001) incluyen dentro de las construcciones o EM tanto expresiones fijas, semifijas (locuciones, compuestos nominales, nombres propios), como expresiones sintácticamente flexibles (construcciones verbo-partícula, locuciones descomponibles, construcciones con verbos 'light', expresiones estereotipadas). Se parte de la hipótesis de que las construcciones son unidades lingüísticas que almacenamos en memoria a las que se accede tanto en la producción como en la comprensión de lenguaje. Jackendoff (1999) estima que el 50% del léxico de los hablantes está constituido por este tipo de construcciones.

Esta concepción del lenguaje conlleva la necesidad de explorar nuevos caminos desde la LC (modelos formales computacionalmente tratables) y desde el PLN (desarrollo de técnicas y métodos) que superen el umbral de calidad al que han llegado las aproximaciones que se han desarrollado hasta este momento.

2 Objetivos

El objetivo general del proyecto DIANA es desarrollar nuevas herramientas de análisis lingüístico coherentes con las nuevas formulaciones teóricas y las necesidades aplicadas desde el PLN y ensayar nuevas aproximaciones en la hibridación de estas herramientas y nuevos recursos para conseguir unos mejores resultados en el análisis del lenguaje. La consecución de este objetivo general se concreta en los siguientes objetivos más específicos:

1. Desarrollo de sistemas híbridos de PLN que combinen conocimiento lingüístico (corpus anotados a diferentes niveles) con técnicas de aprendizaje automático incorporando medidas de similitud semántica que extiendan el conocimiento disponible a casos no tratados.
2. Aplicación de estas tecnologías para superar las actuales limitaciones de los sistemas de resolución de la correferencia, la identificación de paráfrasis y estructura argumental. En concreto, en resolución de la correferencia nos centraremos en el tratamiento de los sintagmas nominales no coincidentes, la identificación de eventos correferentes y argumentos implícitos (Peris, Taulé y Rodríguez, 2013). En cuanto a la paráfrasis y el plagio, nuestro objetivo es ampliar el campo de tratamiento abordando el problema de su detección a partir de estrategias de análisis del discurso (Barrón-Cedeño et al., 2013).
3. Aplicación de estas tecnologías para la identificación de construcciones lingüísticas con el objetivo de mejorar el análisis y comprensión de textos subjetivos, para la identificación de estados de ánimo (estrés, frustración, depresión, neurosis y agresividad), para la detección de pedofilia y acoso en los medios sociales de comunicación (Bogdanova et al., 2013), así como de engaño en los mismos (detección de opiniones fraudulentas artificialmente construidas (Hernández et al., 2013)). Se desarrollarán aplicaciones basadas en este tipo de conocimiento.
4. Inferencia de construcciones lingüísticas representativas en textos figurados de los medios sociales de comunicación para la interpretación de su verdadero sentido. En especial, el reconocimiento de humor y la identificación de ironía en opiniones (Reyes y Rosso, 2012).

3 Metodología

El proyecto dará como resultado un entorno de PLN para el tratamiento de textos a nivel discursivo que detectará construcciones de carácter general y específicas de determinados dominios semánticos, con el objetivo de avanzar en la resolución de la correferencia, detección de paráfrasis y plagio, y en la identificación de la actitud subjetiva en los textos. Las herramientas del entorno serán independientes de la lengua y dispondrán de una interfaz para su extensión a diferentes formatos. Este entorno de PLN tendrá asociadas diferentes estructuras de datos lingüísticos que se obtendrán durante el desarrollo del proyecto: léxicos de construcciones y grafos de palabras semánticamente relacionadas tanto de carácter

general como específicos de los dominios tratados (Franco-Salvador et al., 2013).

Se aplicarán diferentes técnicas de aprendizaje automático para el desarrollo de un *parser* semántico. Se aplicarán modelos geométricos (*Vector Space Models*) para la representación semántica a partir del contexto basándonos en las propuestas de Turney y Pantel (2010), Padó y Lapata (2007) y Baroni y Lenci (2010). Para la inferencia y generalización de construcciones se utilizarán técnicas de inducción, generalización y jerarquización de patrones (Zuidema, 2007; Gries 2003). Esta metodología se refleja analíticamente en los puntos que detallamos a continuación:

a) Recopilación de los corpus disponibles y de corpus de nueva creación.
b) Estandarización de los corpus: formato común en XML con etiquetas identificadoras.
c) Procesamiento de los corpus a nivel básico con las herramientas disponibles.
d) Aplicación de medidas de semántica distribucional para la extracción de relaciones de similitud entre palabras basados en diferentes tipos de modelización del contexto. Nos centraremos en los predicados nominales y verbales.
e) Extensión de la información contenida en los léxicos AnCora-Nom y AnCora-Verb (léxicos semilla) a aquellas palabras del castellano semánticamente relacionadas, dando lugar a los léxicos DIANA-Nom y DIANA-Verb. Los corpus de cada tarea tendrán su léxico específico.
f) A partir del recurso AnCora-Net, que vincula los léxicos nominal y verbal del castellano a los correspondientes del inglés y catalán, se extenderá al conocimiento de DIANA-Nom y DIANA-Verb al catalán y al inglés.
g) Se desarrollará un parser semántico, DIANA-Parser, que tomará como datos de entrada los léxicos DIANA y el corpus analizado automáticamente con dependencias. El resultado será el análisis parcial de corpus con argumentos y papeles temáticos. El parser se desarrollará para las tres lenguas implicadas en el proyecto.
h) Análisis de los corpus con DIANA-Parser.
i) Obtención de léxicos de contextos sintáctico-semánticos asociados a los ítems léxicos nominales y verbales a partir de los árboles de dependencias semánticas del corpus. Estos léxicos de contextos constituirán la base para la obtención de construcciones.
j) Obtención de construcciones de base léxica y de base no léxica a partir de los léxicos de contextos sintáctico-semánticos. Se aplicarán técnicas de generalización y jerarquización de árboles. Se elaborará una jerarquía de construcciones de manera que se puedan establecer equivalencias entre las mismas de cara a su explotación en las diferentes aplicaciones.
k) Aplicación de la tecnología DIANA al desarrollo de aplicaciones que implican la comprensión de textos subjetivos.

4 Resultados esperable del proyecto

Los resultados del proyecto DIANA podrán incidir favorablemente en el desarrollo de aplicaciones en el marco web. Los avances en la tecnología web han puesto a nuestro alcance contenidos generados por los propios usuarios en forma de blogs, opiniones y todo tipo de interacciones en los medios de comunicación social, que se encuentran en forma no estructurada y expresados en fragmentos de texto donde se combina la simple narración de hechos con la toma de decisiones, experiencias aleccionadoras y opiniones sobre todo tipo de eventos. Se trata de una fuente de información valiosa que, si se dispone de los medios necesarios, se puede utilizar para la resolución de problemas sobre la base de conocimiento compartido y reutilizado. Las tecnologías que se desarrollarán permitirán avanzar en la captación de contenidos basados en la experiencia de los propios usuarios y en su aprovechamiento colectivo. Destacamos el desarrollo de diferentes técnicas y métodos susceptibles de ser incorporados en aplicaciones para:

- La identificación de estados de ánimo en textos subjetivos (entusiasmo, fatiga, relajación, estrés, frustración, depresión, agresividad), que permitirán captar el grado de satisfacción de los usuarios respecto de los servicios y aplicaciones que se ofrecen on-line. En los sistemas de tutorización un aspecto clave es conocer los diferentes estados de ánimo del alumno (grado de satisfacción, desorientación, motivación, comprensión de la materia, o ciertos síntomas relevantes en el proceso de aprendizaje).

- La detección en los medios sociales de comunicación de usuarios con potenciales trastornos de la personalidad, a nivel de agresividad o de neurosis, que podrían alertar sobre posibles casos de acoso y pedofilia. En lugar de rastrear manualmente la red para encontrar potenciales acosadores, un sistema automático de alerta ayudaría a los expertos en la identificación de los acosadores potenciales.
- La detección de opiniones fraudulentas creadas por parte de personas específicamente contratadas para este fin. Nuestra contribución consistirá en el desarrollo de técnicas para la extracción de las construcciones más recurrentes en la expresión de opiniones fraudulentas sobre personas, organizaciones, productos y servicios.
- La incorporación de un detector de ironía que mejorara las prestaciones de los sistemas de análisis de opinión. En el lenguaje figurado, el sentido literal del texto no coincide con el sentido que se quiere comunicar, de ahí la importancia de poder identificar las construcciones prototípicas de la expresión de la ironía en los medios sociales de comunicación.

Los nuevos recursos y herramientas que se prevé desarrollar serán utilizables en diferentes entornos de PLN y, muy especialmente, en aplicaciones como las que acabamos de describir. En concreto, los grafos de similitud léxica basados en el contexto y el analizador semántico serán de utilidad para superar las actuales limitaciones en temas como la resolución de la correferencia, la detección de paráfrasis, la obtención de la estructura argumental y la caracterización de usos estereotipados del lenguaje en casos como los que acabamos de apuntar.

Bibliografía

Baroni M. y A. Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Barrón-Cedeño, A., M. Vila, M. A. Martí y P. Rosso. 2013. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics.* [doi:10.1162/COLI_a_00153].

Bogdanova D., Rosso P., Solorio T. Exploring High-Level Features for Detecting Cyberpedophilia. *Computer Speech and Language* (aceptado).

Croft, W. y D. A. Cruse. 2004. *Cognitive linguistics*, Cambridge Textbooks in Linguistics, Cambridge University Press.

Franco-Salvador M., Gupta P., Rosso P. 2013. Cross-Language Plagiarism Detection Using Multilingual Semantic Network. *Proc. 35th, ECIR-2013*, Springer-Verlag, LNCS(7814).

Goldberg, A. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Gries, S. Th. 2003. Collostructions: Investigating the interaction of words and constructions, *International Journal of Corpus Linguistics*, 8:2.

Hawkins, J. 2004. *On Intelligence*, Henry Holt and Company, LLC, New York, USA.

Hernández D., Guzmán R., Montes-y-Gómez M., Rosso P. 2013. Using PU-Learning to Detect Deceptive Opinion Spam. *WASSA*-2013.

Jackendof, R. 1999. Possible stages in the evolution of the language capacity. *Trends in Cognitive Sciences*, 3(7): 272-279.

Padó, S. y M. Lapata. 2007. Dependency-Based Construction of Semantic Space Models, *Computational Linguistics*, 33(2).

Peris, A., M. Taulé, H. Rodríguez y M. Bertran. 2013. LIARc: Labeling Implicit ARguments in Spanish deverbal nominalizations, *CICLING-2013.* Springer.

Reyes A. y P. Rosso 2012. Making Objective Decisions from Subjective Data: Detecting Irony in Customers Reviews. *Journal on Decision Support Systems*, 53(4):754–760.

Sag I., T. Baldwin, F. Bond, A. Copestake y D. Flickinger. 2001. Multiword Expressions: A Pain in the Neck for NLP. *CICLING-2002*. Springer, LNCS (2276): 1-15.

Tomasello, M. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition.* Harvard University Press.

Turney, P. y P. Pantel. 2010. From Frequency to meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141-188.

Wintner, S. 2009. What science underlies natural language engineering? *Computational Linguistics*, 35 (4): 641-644.

Zuidema, W. 2007. Parsimonious Data-Oriented Parsing, Proc. *EMNLP-CoNLL.*