I *Workshop* en Procesamiento Automatizado de Textos y Corpora, Viña del Mar – Chile, 19 y 20 de julio de 2012

Por: NELLY ESPERANZA MORENO
Universidad Nacional de Colombia

EN EL CAMPUS Sausalito, ubicado en Viña del Mar, Chile, de la Pontificia Universidad Católica de Valparaíso, se llevó a cabo el *I Workshop en Procesamiento Automatizado de Textos y Corpora*, un espacio destinado a la socialización y discusión de trabajos e investigaciones desarrollados en ámbitos como la lingüística computacional, la lingüística de corpus y el procesamiento del lenguaje natural.

Con la participación de investigadores de Chile, Argentina, México, Colombia y España, se realizaron diversas conferencias, entre el 19 y 20 de julio, que abordaron tópicos como la clasificación automática de textos, la detección automática del plagio, el análisis automático de textos y la modelización computacional del lenguaje. El *Workshop* se dividió en tres grandes momentos: la conferencia inaugural, el panel de discusión entre lingüistas e ingenieros computacionales y la presentación de las conferencias por parte de los ponentes.

La conferencia inaugural estuvo a cargo del Ph.D Manuel Montes y Gómez, quien presentó el panorama general de la clasificación automática de textos, una tarea de minería de datos que consiste en organizar información —documentos—, de acuerdo con ciertas categorías predefinidas, teniendo en cuenta su contenido. Entre las aproximaciones más comunes de esta tarea, se encuentra, de un lado, el enfoque supervisado que se refiere a la representación de textos y documentos, mediante árboles, reglas o probabilidades, con base en las palabras empleadas en dichos textos. Este acercamiento emplea el modelo de representación denominado *bolsas de palabras —bags of words—*, con el cual se evalúa qué tan discriminativo es un término para la tarea de clasificación, es decir, si una palabra puede o no constituirse como un rasgo que determine la pertenencia o no de un documento a una determinada clase. Como consecuencia de esta representación, según Montes y Gómez, se acarrea la pérdida de información contextual de las palabras, debido a que cada una se convierte en un vector, razón por la cual se utilizan técnicas como los n-gramas y la identificación de frases, que sí incluyen los términos que coocurren con una determinada palabra.

Montes y Gómez expuso otra aproximación muy reciente a esta tarea. denominada *clasificación consensuada*, en la cual la información se categoriza teniendo en cuenta no solo el contenido del documento a clasificar, sino también los documentos vecinos que lo circundan y que, como supuesto fundamental, podrían pertenecer a la misma categoría. Con este método, se establece un documento prototipo, es decir, un ejemplo virtual que reúne en promedio las características de un grupo de documentos. Este prototipo se compara con los documentos por clasificar y, como resultado de la comparación, es posible evaluar la similitud promedio de cada uno de los documentos, de tal manera que los documentos con las mayores similitudes entran a pertenecer a la misma categoría. Este novedoso método tiene como ventaja el hecho de que clasifica un documento con base en sí mismo y también comparándolo con otros que son similares —en términos de tipología o géneros textuales, autor, temática, etc.—.

Finalmente, Montes y Gómez evaluó los resultados de este enfoque en la clasificación de documentos cortos, constituidos por títulos de noticias, y de documentos en diferentes lenguas. Su conferencia concluyó con la formulación de diversos retos de la clasificación automática de textos, entre los cuales se encuentra el desarrollo de técnicas que permitan realizar, de manera más precisa, esta tarea, tanto en documentos largos como en documentos cortos; que se adapten a la identificación de dominios —géneros textuales—, y que posibiliten, por ejemplo, organizar la información con base en palabras locales —conjuntos de palabras que caracterizan secciones de un texto—. La principal invitación de este investigador fue motivar a la comunidad académica y científica a tomar el riesgo de emplear la gran cantidad de información disponible en Internet —artículos, *tweets*, correos electrónicos, blogs, entre otros—.

En segundo lugar, otro de los momentos más representativos del *Workshop* fue el *Panel de convergencia interdisciplinar* en el cual participaron lingüistas e ingenieros informáticos. El debate se dio en torno a algunas preguntas que pretendían examinar, por un lado, el aporte de las técnicas y métodos de las ciencias de la computación a los estudios lingüísticos y, por otro, la contribución de la lingüística al desarrollo de sistemas y algoritmos más eficientes para el procesamiento del lenguaje natural.

En general, los panelistas indicaron que, a propósito de la primera cuestión, las ciencias computacionales han proporcionado diversos recursos para el análisis automático de textos. En el campo de la terminología, por ejemplo, la tecnología se ha constituido en un asistente del traductor que permite contar con diversos corpus —comparables

o paralelos— y detectar terminología —conjunto de palabras que representan un cuerpo de conocimientos especializados— mediante la asociación estadística de términos y la formulación automática de patrones sintácticos. Sin embargo, a pesar de que las ciencias computacionales facilitan la labor del traductor y del lingüista, se presentan algunos inconvenientes que requieren más finura de estos métodos a nivel lingüístico, por lo cual los panelistas plantearon la necesidad de implementar miradas interdisciplinares que faciliten la vinculación entre la descripción y el análisis de un lingüista, y las herramientas computacionales diseñadas por los ingenieros.

Sobre la segunda cuestión, los panelistas plantearon que si bien algunos modelos computacionales formulados por los ingenieros son válidos para dar solución a problemas de análisis del lenguaje, muchos de estos no son interpretables. En consecuencia, el principal aporte de la lingüística a la computación consiste en la formulación de mejores modelos, con acercamientos más adecuados al lenguaje, los cuales puedan facilitar una interpretación significativa de los resultados obtenidos. Los panelistas indicaron que campos como el procesamiento automático de textos o la minería de textos no implican una separación entre la lingüística y la computación. El principal objetivo de estas áreas es dejar a las máquinas las tareas más complejas para que el lingüista se dedique a la interpretación.

De otro lado, las conferencias presentadas por los ponentes abordaron interesantes temáticas del procesamiento automático del lenguaje y la minería de textos. A propósito de la detección automática de plagio, Parth Gupta, de la Universidad Politécnica de Valencia, presentó un novedoso método para identificar casos de plagio por parafraseo entre dos lenguas —inglés y alemán—, mediante una técnica que realiza un mapeo semántico-sintáctico entre el texto original y la paráfrasis, y que permite, de manera preliminar, identificar aquellos segmentos en los cuales el autor ha cometido plagio.

Sobre la clasificación automática, Walter Koza, de la Universidad Nacional de Rosario, propuso un método para clasificar automáticamente los usos gramaticales y sintácticos de la coma —elipsis, numeración, delimitación, desambiguación, entre otros—. La extracción de los rasgos de análisis se realizó mediante las herramientas *Smorph* y *Post-Smorph*, con las cuales se segmentaron y analizaron morfológicamente las cadenas de texto y se etiquetaron las funciones gramaticales de la coma. Con base en un corpus de 10 000 palabras, y teniendo como base las medidas de evaluación, la precisión y cobertura, el ponente indicó que, en términos generales, el modelo de clasificación obtenido arroja buenos resultados en la identificación automática de las funciones de la coma, especialmente, cuando esta se emplea para señalar marcadores discursivos.

Otra investigación sobre clasificación automática fue presentada por Mauricio Martis, de la Universidad Católica de Valparaíso, quien en su trabajo intentó clasificar la intención comunicativa —chat, opinión, localización espacial, mención, noticias, entre otros— de un corpus constituido por *tweets*. Empleando los algoritmos *Naive Bayes* y *SVM*, en un modelo de representación vectorial de los documentos —bolsas de palabras—, se determinó que los *tweets* mejor clasificados fueron aquellos que dieron información sobre la localización espacial del autor.

En la misma línea de clasificación automática de *tweets*, otra investigación, presentada por Juan Pablo Cárdenas, de la Universidad Católica de Valparaíso, se ocupó de la clasificación automática de este tipo de textos mediante el uso de redes de palabras y sus coocurrencias. Lo novedoso de esta técnica consiste en que la relación entre palabras se emplea como un rasgo lingüístico que permite clasificar diferentes textos de acuerdo con su dominio. Esta relación de dependencia se puede visualizar como un enlace en una red que, a su vez, indica qué transiciones entre palabras son más probables y caracterizan a un determinado tipo de textos. Los resultados de esta investigación, obtenidos para un corpus de *tweets*, superan los logrados con métodos de clasificación automática que usan, por ejemplo, el modelo de representación de bolsa de palabras.

De otro lado, en una investigación realizada sobre artículos científicos del campo de la medicina, Concepción Pérez de Celis, de la Universidad Autónoma de Puebla, empleó, como método de clasificación, una *ontología* definida, desde la gestión del conocimiento y las bases de datos, como un modelo de representación del conocimiento que organiza de manera jerárquica un conjunto de conceptos. Utilizando una ontología de síndromes, enfermedades y genes, la investigadora presentó buenos resultados en la clasificación automática de un corpus de artículos biomédicos relacionados con la sordera o hipoacusia.

Sobre el análisis automático de la cohesión, René Venegas, de la Universidad Católica de Valparaíso, expuso los resultados de la investigación que realizó en torno a los mecanismos de cohesión presentes en los discursos presidenciales de diferentes mandatarios chilenos del periodo dictatorial —1974-1989— y del periodo democrático —1990-2012—. La extracción de los rasgos lingüísticos y su análisis se llevó a cabo con la herramienta *Spanishtool*. Entre los resultados más importantes, los cuales se obtuvieron con el método de *Análisis Semántico Latente*, se encuentra una diferenciación radical entre el discurso dictatorial y el discurso democrático, ya que en el primero se privilegia el uso de oraciones mucho más extensas, en las cuales son comunes la adjetivización, como recurso de expresión de apreciaciones

y afectos, y la reiteración léxica. Asimismo, otro aspecto que diferencia a este tipo de discurso es la similitud semántica entre las oraciones que lo componen. René Venegas indica que estos resultados pueden deberse al hecho de que el discurso de la dictadura se preocupa más por la elaboración retórica y la recurrencia de temas.

Ahora, a propósito de la evolución del lenguaje, la investigación presentada por Nelly Moreno, de la Universidad Nacional de Colombia, se ocupó del análisis automático de la evolución de la competencia para escribir textos narrativos en español. En este trabajo, la extracción de los rasgos lingüísticos —palabras, etiquetas POS y n-gramas— se llevó a cabo con la ayuda de herramientas computacionales como *TreeTagger*, *FreeLing* y *WEKA*. En general, se realizó un análisis de correlación entre la edad y el lenguaje empleado en las narraciones, con base en el cual se determinó que algunos rasgos lingüísticos que dan cuenta de la evolución de la habilidad para narrar son los pronombres —estrategias anafóricas—, los signos de puntuación y una morfología verbal mucho más enriquecida. De otro lado, entre los rasgos con correlación negativa, se identificaron frases nominales —estrategias nominales—, la conjunción coordinante "y" y procesos de derivación diminutiva.

Finalmente, otras conferencias abordaron la automatización de técnicas como el *Background Reading* para la representación de textos técnicos —Bell Manrique, de la Universidad Nacional de Colombia— y la modelización computacional del lenguaje mediante el modelo de *Sintaxis Dinámica* —Nicolás Saavedra, de la Pontificia Universidad Católica de Chile—.

El *Workshop* finalizó con una discusión sobre las perspectivas de trabajo y limitaciones del desarrollo en áreas como la lingüística de corpus y el procesamiento automático de textos en Hispanoamérica. En general, conferencistas y expertos indicaron que es necesario fortalecer las redes de trabajo entre investigadores de diferentes latitudes para visibilizar los resultados que, en estas áreas, se han logrado, y para socializar los problemas o soluciones encontrados en los procesos de investigación en torno al procesamiento automático del lenguaje natural.

La invitación más importante de este *Workshop* fue a desarrollar trabajos interdisciplinares que involucren una intervención de estudiosos del lenguaje y de la computación. Esto garantizaría un adecuado acercamiento a los problemas de investigación planteados, que permita no solo un análisis lingüístico apropiado, sino también la implementación de mejores técnicas y herramientas computacionales. Asimismo, el *Workshop* mostró el amplio abanico de posibilidades de investigación en lingüística de corpus, el cual va desde aspectos microestructurales, como la cohesión, hasta discursivos y pragmáticos, como la caracterización de dominios y la identificación de la intención.