novática
CEPIS UPGRADE Revista de la Asociación de Técnicos de Informática

Nº 204, marzo-abril 2010, año XXXVI

sumario

Monografía del próximo número:
"Ciencia y tecnología de los servicios informáticos"

Novática, revista fundada en 1975 y decana de la prensa informática española, es el órgano oficial de expresión y formación continua de **ATI** (Asociación de Técnicos de Informática), organización que edita también la revista **REICIS** (Revista Española de Innovación, Calidad e Ingeniería del Software). ***Novática*** edita asimismo **UP**GRADE, revista digital de **CEPIS** (*Council of European Professional Informatics Societies*), en lengua inglesa, y es miembro fundador de **UP**ENET (**UP**GRADE *European* ***NET****work*).

<http://www.ati.es/novatica/>
<http://www.ati.es/reicis/>
<http://www.upgrade-cepis.org/>

ATI es miembro fundador de **CEPIS** (*Council of European Professional Informatics Societies*) y es representante de España en **IFIP** (*International Federation for Information Processing*); tiene un acuerdo de colaboración con **ACM** (*Association for Computing Machinery*), así como acuerdos de vinculación o colaboración con **AdaSpain**, **AI2, ASTIC, RITSI** e **Hispalinux**, junto a la que participa en **ProInnova**.

Consejo Editorial
Joan Batlle Montserrat, Rafael Fernández Calvo, Luis Fernández Sanz, Javier López Muñoz, Alberto Llobet Batllori, Gabriel Martí Fuentes, Josep Molas i Bertran, José Onofre Montesa Andrés, Olga Pallás Codina, Fernando Piera Gómez (Presidente del Consejo), Ramon Puigjaner Trepat, Miquel Sarries Griñó, Adolfo Vázquez Rodríguez, Asunción Yturbe Herranz

Coordinación Editorial
Llorenç Pagés Casas <pages@ati.es>
Composición y autoedición
Jorge Llácer Gil de Ramales
Traducciones
Grupo de Lengua e Informática de ATI <http://www.ati.es/gt/lengua-informatica/>
Administración
Tomás Brunete, María José Fernández, Enric Camarero, Felicidad López

Secciones Técnicas - Coordinadores
Acceso y recuperación de la información
José María Gómez Hidalgo (Optenet), <jmgomezh@yahoo.es>
Manuel J. Maña López (Universidad de Huelva), <manuel.mana@diesia.uhu.es>
Administración Pública electrónica
Francisco López Crespo (MAE), <flc@ati.es>
Arquitecturas
Enrique F. Torres Moreno (Universidad de Zaragoza), <enrique.torres@unizar.es>
Jordi Tubella Morgadas (DAC-UPC), <jordit@ac.upc.es>
Auditoría SITIC
Marina Touriño Troitiño, <marinatourino@marinatourino.com>
Manuel Palao García-Suelto (ASIA), <manuel@palao.com>
Derecho y tecnología
Isabel Hernando Collazos (Fac. Derecho de Donostia, UPV), <isabel.hernando@ehu.es>
Elena Davara Fernández de Marcos (Davara & Davara), <edavara@davara.com>
Enseñanza Universitaria de la Informática
Cristóbal Pareja Flores (DSIP-UCM), <cpareja@sip.ucm.es>
J. Ángel Velázquez Iturbide (DLSI I, URJC), angel.velazquez@urjc.es>
Entorno digital personal
Andrés Marín López (Univ. Carlos III), <amarin@it.uc3m.es>
Diego Gachet Páez (Universidad Europea de Madrid), <gachet@uem.es>
Estándares Web
Encarna Quesada Ruiz (Pez de Babel) <equesada@pezdebabel.com>
José Carlos del Arco Prieto (TCP Sistemas e Ingeniería) <jcarco@gmail.com>
Gestión del Conocimiento
Joan Baiget Solé (Cap Gemini Ernst & Young), <joan.baiget@ati.es>
Informática y Filosofía
José Angel Olivas Varela (Escuela Superior de Informática, UCLM) <joseangel.olivas@uclm.es>
Karim Gherab Martín (Harvard University) <kgherab@gmail.com>
Informática Gráfica
Miguel Chover Sellés (Universitat Jaume I de Castellón), <chover@lsi.uji.es>
Roberto Vivó Hernando (Eurographics, sección española), <rvivo@dsic.upv.es>
Ingeniería del Software
Javier Dolado Cosín (DLSI-UPV), <dolado@si.ehu.es>
Luis Fernández Sanz (Universidad de Alcalá), <luis.fernandezs@uah.es>
Inteligencia Artificial
Vicente Botti Navarro, Vicente Julián Inglada (DSIC-UPV)
<{vbotti, vinglada}@dsic.upv.es>
Interacción Persona-Computador
Pedro M. Latorre Andrés (Universidad de Zaragoza, AIPO) <platorre@unizar.es>
Francisco L. Gutierrez Vela (Universidad de Granada, AIPO) <fgutierr@ugr.es>
Lengua e Informática
M. del Carmen Ugarte García (IBM), <cugarte@ati.es>
Lenguajes Informáticos
Oscar Belmonte Fernández (Univ. Jaime I de Castellón), <belfern@lsi.uji.es>
Inmaculada Coma Tatay (Univ. de Valencia), <Inmaculada.Coma@uv.es>
Lingüística computacional
Xavier Gómez Guinovart (Univ. de Vigo), <xgg@uvigo.es>
Manuel Palomar (Univ. de Alicante), <mpalomar@dlsi.ua.es>
Mundo estudiantil y jóvenes profesionales
Federico G. Mon Trotti (RITSI) <gnu.fede@gmail.com>
Mikel Salazar Peña (Area de Jovenes Profesionales, Junta de ATI Madrid), <mikeltxo_uni@yahoo.es>
Profesión Informática
Rafael Fernández Calvo (ATI), <rfcalvo@ati.es>
Miquel Sárries Griñó (Ayto. de Barcelona), <msarries@ati.es>
Redes y servicios telemáticos
José Luis Marzo Lázaro (Univ. de Girona), <joseluis.marzo@udg.es>
Juan Carlos López López (UCLM), <juancarlos@uclm.es>
Seguridad
Javier Areitio Bertolín (Univ. de Deusto), <jareitio@eside.deusto.es>
Javier López Muñoz (ETSI Informática-UMA), <jlm@lcc.uma.es>
Sistemas de Tiempo Real
Alejandro Alonso Muñoz, Juan Antonio de la Puente Alfaro (DIT-UPM),
<{aalonso,jpuente}@dit.upm.es>
Software Libre
Jesús M. González Barahona (GSYC-URJC), <jgb@gsyc.es>
Israel Herráiz Tabernero (UAX), <isra@herraiz.org>
Tecnología de Objetos
Jesus García Molina (DIS-UM), <jmolina@um.es>
Gustavo Rossi (LIFIA-UNLP, Argentina), <gustavo@sol.info.unlp.edu.ar>
Tecnologías para la Educación
Juan Manuel Dodero Beardo (UC3M), <dodero@inf.uc3m.es>
César Pablo Córcoles Briongo (UOC), <ccorcoles@uoc.edu>.
Tecnologías y Empresa
Didac López Viñas (Universitat de Girona), <didac.lopez@ati.es>
Francisco Javier Cantais Sánchez (Indra Sistemas), <fjcantais@gmail.com>
Tendencias tecnológicas
Alonso Alvarez García (TID), <aad@tid.es>
Gabriel Martí Fuentes (Interbits), <gabi@atinet.es>
TIC y Turismo
Andrés Aguayo Maldonado, Antonio Guevara Plaza (Univ. de Málaga)
<{aguayo, guevara}@lcc.uma.es>

Las opiniones expresadas por los autores son responsabilidad exclusiva de losmismos. ***Novática*** permite la reproducción, sin ánimo de lucro, de todos los artículos, a menos que lo impida la modalidad de © o *copyright* elegida por el autor, debiéndose en todo caso citar su procedencia y enviar a ***Novática*** un ejemplar de la publicación.

Coordinación Editorial, Redacción Central y Redacción ATI Madrid
Padilla 66, 3º, dcha., 28006 Madrid
Tlfn. 914029391; fax 913093685 <novatica@ati.es>
Composición, Edición y Redacción ATI Valencia
Av. del Reino de Valencia 23, 46005 Valencia
Tlfn./fax 963330392 <secreval@ati.es>
Administración y Redacción ATI Cataluña
Via Laietana 46, ppal. 1ª, 08003 Barcelona
Tlfn.934125235; fax 934127713 <secregen@ati.es>
Redacción ATI Aragón
Lagasca 9, 3-B, 50006 Zaragoza.
Tlfn./fax 976235181 <secreara@ati.es>
Redacción ATI Andalucía <secreand@ati.es>
Redacción ATI Galicia <secregal@ati.es>
Suscripción y Ventas <http://www.ati.es/novatica/interes.html>, ATI Cataluña, ATI Madrid
Publicidad
Padilla 66, 3º, dcha., 28006 Madrid
Tlnf.914029391; fax.913093685 <novatica@ati.es>
Imprenta: Derra S.A., Juan de Austria 66, 08005 Barcelona.
Depósito legal: B 15.154-1975 -- ISSN: 0211-2124; CODEN NOVAEC
Portada: ADN enamorado - Concha Arias Pérez / © ATI
Diseño: Fernando Agresta / © ATI 2003

Tendencias en Procesamiento del Lenguaje Natural y Minería de Textos

Javier Pueyo[1], José Antonio Quiles Follana[2]
[1]GSyC/LibreSoft (URJC), CSIC; [2]Telefónica I+D

<javier.pueyo@gmail.com>, <quiles@tid.es>

Resumen: *los procesos de comunicación o de información, las opiniones, e incluso los sentimientos son compartidos, almacenados y codificados por los seres humanos y por sus instituciones en lenguaje natural (en oposición a los lenguajes artificiales, estructurados o de programación utilizados por las computadoras). Los lingüistas han destacado desde hace siglos la complejidad del análisis y la decodificación del lenguaje humano: baja precisión, altamente contextual y ambiguo. El uso generalizado de las computadoras y las redes mundiales de comunicación han hecho que la mayoría de nuestra comunicación en lenguaje natural (correo electrónico, mensajería instantánea, informes, documentación, e incluso las ideas, las aficiones o las historias personales) se codifiquen y se almacenen en formato digital, y que se compartan a través de los sistemas informáticos. Las técnicas de Procesamiento del Lenguaje Natural (PLN), desarrolladas en el campo de la lingüística computacional, pueden ciertamente sacar partido de este hecho, y ya están siendo ampliamente utilizadas en áreas tales como la minería de textos, de la recuperación de la información, clustering documental, análisis de opiniones o gestión del conocimiento. En este artículo ofrecemos una panorámica de los recursos de conocimiento externo que empiezan ahora a ser explotados para mejorar y enriquecer los mecanismos de PLN. También exploramos los usos futuros de dichos procedimientos, mediante la combinación del conocimiento interno extraído de los documentos con la información externa que tendremos disponible, a través de bancos de datos especializados y estructurados o de diccionarios semánticos o conceptuales.*

Palabras clave: *agrupamiento automático, análisis de opiniones y sentimientos, aprendizaje automático, clasificación automática, Gestión del Conocimiento, Lingüística Computacional, Minería de Texto, Procesamiento de Lenguaje Natural.*

Autores

Javier Pueyo es investigador en el Instituto de Lenguas y Culturas (CSIC, España). Es licenciado en Filología Hispánica por la Universidad de Deusto, España (1990), *Master of Arts* en Lingüística Hispánica por la *University of Southern California* (EEUU) habiendo completado su Doctorado en Filología Española en 1996 (UD). Las áreas de conocimiento en las que trabaja incluyen Lingüística, Lengua y Literatura Sefardíes, y desarrollo y explotación de corpus lingüísticos y técnicas de PLN para el análisis diacrónico de la lengua. Se ha incorporado recientemente al grupo de investigación GSyC/LibreSoft (Universidad Rey Juan Carlos, España) en donde desarrolla y aplica técnicas de PLN al análisis de proyectos de Software Libre y Cultura Libre.

José Antonio Quiles Follana es Ingeniero de Telecomunicación por la Universidad Politécnica de Madrid (1990). Trabajó en el Instituto Nacional de Industria (INI, España) en proyectos de guerra electrónica para el Ministerio de Defensa. Posteriormente, en el Grupo de Bioingeniería y Telemedicina de la Universidad Politécnica de Madrid, dirigiendo y desarrollando proyectos de investigación relacionados con las imágenes médicas y la telemedicina. Actualmente es especialista tecnológico en Telefónica I+D (España), donde ha trabajado en áreas como planificación de redes móviles, sistemas de información geográfica, gestión y optimización de fuerzas de trabajo, gestión del conocimiento, detección del fraude en sistemas prepago de telefonía móvil, buscadores, evaluación de recursos humanos, infraestructura de servicios web y arquitecturas orientadas a servicios (SOA), web semántica, clasificación automática de texto libre y buscadores conceptuales.

1. Introducción

Los lenguajes naturales, en oposición a los lenguajes artificiales, proporcionan a los humanos y a sus instituciones (agencias legales, organizaciones financieras y políticas, negocios, universidades, hospitales, administraciones públicas, industrias o comunidades de ciudadanos) un sistema muy sofisticado para compartir información, hechos, opiniones, pensamientos, juicios, creencias e incluso sentimientos.

Durante miles de años, los seres humanos han utilizado las gramáticas y palabras de sus lenguajes para codificar los mensajes. Estos lenguajes son muy efectivos para la comunicación humana, pero los lenguajes naturales son léxicamente [1] y estructuralmente [2] ambiguos, contextuales tanto verbal como socialmente [3], e incluso en algunos casos metafóricos [4] por naturaleza.

Así, aunque para muchas lenguas existen descripciones comprensivas de su fonología, morfología, sintaxis, semántica o vocabularios, todavía estamos lejos de desarrollar sistemas de inteligencia artificial (IA) que las comprendan automáticamente y que además puedan producir un lenguaje de forma similar a como lo hacemos nosotros [5].

El campo del Procesamiento del Lenguaje Natural (PLN) ofrece diversas técnicas que han demostrado ser muy útiles para analizar automáticamente, extraer y proporcionar conocimiento de las fuentes de información intrínsecamente no estructuradas codificadas en el lenguaje humano. Durante los últimos 20 años, las técnicas de PLN han evolucionado desde métodos con reglas simbólicas (programados en Lisp y Prolog) basados fundamentalmente en lógica e introspección lingüística, es decir, nuestro propio conocimiento lingüístico interno de la lingüística, hasta métodos con procesamiento intensivo de datos, estadísticos y probabilísticos, orientados al procesamiento del lenguaje (y también más independientes de los lenguajes de programación), basados en la producción y almacenamiento masivo que actualmente disponemos.

2. Corpus lingüísticos

La aparición de la disciplina de corpus lingüísticos a finales de los años sesenta [6] y la disponibilidad de bases de documentos anotados y categorizados lingüísticamente (desde un millón de palabras en el Brown Corpus [7], hasta los 100 millones de palabras en el BNC [8], ANC [9], CREA para español [10], e incluso parseados como el Penn Treebank [11], o clasificados por categorías [12]) permitieron a los investigadores cambiar el enfoque desde una visión más de teoría lingüística a empezar a trabajar con información lingüística masiva almacenada en computadores, tanto para entrenamiento y pruebas como en la implantación de sus modelos.

El proceso de creación de un corpus de documentos anotado es básicamente manual, una tarea costosa pero necesaria para desarrollar herramientas de anotación automática. La existencia de estas bases documentales recogidas y anotadas permitió la creación de herramientas actuales para crear, anotar y analizar nuevas colecciones de documentos en muchos lenguajes.

Estos corpus de referencia para una lengua

completa junto con las herramientas desarrolladas para este tipo de datos son realmente demasiado generales para poder aplicarse con efectividad en algunas áreas muy específicas. Sin embargo, son la base que permite abarcar corpus altamente especializados (y probablemente mucho más pequeños) para su uso en las áreas en las que operan las instituciones y organizaciones: biomédicas [13], legales [14], etc.

3. Métodos y software para PLN

Los procesos básicos de PLN aplicados a texto libre para conseguir una estructura inicial de éste son: *tokenización*, segmentación en sentencias, etiquetado gramatical (PoS, *part of speech*), eliminación de palabras sin aporte semántico (*stop words*), extracción de raíces (*stemming*), lematización, detección de subsentencias (frases nominales o frases verbales).

Existen otras técnicas más avanzadas tales como análisis sintáctico, resolución de anáforas, reconocimiento de entidades (NER) y otros tipos de anotaciones semánticas, que son aplicables a la gran parte de la información lingüística contenida en el texto libre con el objetivo de extraer significado. Existen recursos como gramáticas basadas en reglas [15] [16] [17], sistemas de aprendizaje automático utilizando diversos modelos (árboles de decisión, modelos de Markov, redes de Bayes, *Naïve Bayes* y más recientemente *Support Vector Machines*), que se están aplicando por ejemplo para etiquetado PoS. También existen otros recursos léxicos muy valiosos, tales como WordNet [18] o EuroWordNet [19] por ejemplo, que se están utilizando para mejorar los pasos de desambiguación que son necesarios en las distintas etapas del PLN.

La utilización cada vez mayor de modelos estadísticos y probabilísticos requiere el cálculo detallado de frecuencias y el uso de medidas de relevancia: ganancia de información (GI), frecuencias de términos-inversa de frecuencia en documentos (TF-IDF), información mutua (IM), distribución Chi cuadrado, etc.

Otra información lingüística como N-gramas (palabras y secuencias de palabras), lemas, sentencias, entidades, PoS, métricas (longitudes de palabras y sentencias), relaciones entre términos (colocados y coligados), son las características que alimentan los algoritmos de aprendizaje automático para construir tokenizadores eficientes, etiquetadores PoS, analizadores sintácticos o clasificadores [20] [21] [22].

Todas estas técnicas se han desarrollado y mejorado en muchos lenguajes y entornos. Y dependiendo de la naturaleza de los datos que se están analizando, los resultados se acercan al 100% de precisión en los procesos de PLN más comunes.

Aunque existe software comercial ya disponible que realiza muchas de las tareas necesarias para un tratamiento de PLN para texto libre, y PLN es una de las áreas más activas en los laboratorios de investigación de la industria del software [23] [24] [25] [26], sin embargo, una de las tendencias más prometedoras en el campo de implementaciones de PLN, como en muchas otras áreas de la computación, viene de las comunidades de software libre y abierto. Incluso aunque los programas y librerías libres podrían parecer incompletos y en algunos casos quedan por detrás del software comercial, sin embargo, la flexibilidad para integrar herramientas de PLN, desplegar y distribuir libremente, así como la disponibilidad del código fuente, permitirá que muchos campos de investigación en comunidades fuera del PLN puedan beneficiarse de la mayoría de las herramientas libres e integrarlas en muchos de sus productos, soluciones e investigación. Ejemplos de comunidades y herramientas de PLN libres: FreeLing, una suite completa consistente en programas y librerías para análisis del lenguaje [27], Weka, una colección de algoritmos de aprendizaje automático [28], GATE, para anotación y otras tareas de procesamiento del lenguaje [29], NLTK, una colección de módulos Python y datos lingüísticos para PLN [30]. Estos son sólo algunos ejemplos representativos de implementaciones FLOSS.

4. Áreas de utilización de PLN

Como se mencionó antes, el lenguaje lo es todo, y cada aspecto de la interacción humana está rodeado de expresiones verbales no estructuradas, de forma que no hay límites en las áreas donde potencialmente se podrían utilizar las técnicas de PLN para procesar información automáticamente. Cualquier aplicación que de alguna manera trate con texto puede beneficiarse de ellas. Algunas de las áreas y aplicaciones en las cuales el uso de PLN está muy establecido o está apareciendo ahora se mencionan a continuación [31].

La corrección automática de documentos es un área que está empezando a emerger en industrias tales como editoriales y departamentos legales de las empresas. Los correctores manuales existentes actualmente dejarían paso a correctores automáticos sin ninguna intervención de los usuarios. Para poder alcanzar este objetivo es necesario que el sistema comprenda la semántica y los conceptos de los que trata el texto para poder tomar decisiones de corrección. Además de la corrección ortográfica a la que estamos acostumbrados es posible la separación automática de palabras en líneas (*hyphenation*), chequeo gramatical y chequeo de estilo.

Existen otras tareas típicas en el procesamiento de colecciones de documentos tales como la clasificación automática. Aquí se dispone de un conjunto de categorías o tópicos predefinidos, y se trata de asignar a cada uno de los documentos, una (o más) categorías basándose únicamente en su contenido.

Otra aplicación típica cuando se procesan grandes cantidades de documentos textuales es el *clustering*. En este caso, el sistema no dispone de un conjunto de categorías definidas a priori, sino que intenta descubrir estas categorías buscando documentos similares y agrupándolos en grupos de documentos que tienen características comunes (siempre basándose en el contenido textual). La extracción de resúmenes es otra tarea de gran utilidad sobre todo en la sociedad actual donde tenemos acceso a inmensas cantidades de textos y no disponemos del tiempo necesario para procesarlos. Los resúmenes automáticos podrían permitir la selección o descarte casi inmediato de grandes cantidades de información.

Las técnicas de PLN han ayudado de manera muy notable en tareas tales como la identificación del idioma de los textos, la búsqueda y la extracción de información, permitiendo realizar búsquedas mucho más precisas y contextuales aplicando las técnicas de procesado lingüístico disponibles en la actualidad.

En una sociedad como la actual, donde tenemos acceso a gran parte del conocimiento sin movernos de nuestra mesa, desde un simple equipo informático con conexión a internet, cada vez se empieza a demandar con mayor interés la traducción automática de textos en diferentes idiomas. En su nivel más básico, la traducción automática consiste en sustituir las palabras de un idioma por palabras del idioma destino. Con la utilización de corpus lingüísticos se pueden intentar traducciones más complejas, utilizando etiquetados sintácticos, identificación de entidades y conceptos, etc. Técnicas que permiten una traducción mucho más inteligente que la mera sustitución de palabras.

Otra aplicación que está empezando a tomar relevancia es la creación automática de texto libre. La generación de lenguaje natural es el proceso de construcción de un texto en lenguaje natural para comunicar un objetivo específico. Para generar texto en lenguaje natural, se parte de un conocimiento que se quiere transmitir, hay que decidir cómo organizar esa información y por último hay que determinar cómo producir el texto, incluyendo las entradas léxicas y las estructuras sintácticas. La generación de texto libre está empezando a aplicarse en sistemas de pregunta/respuesta.

A lo largo de la historia, el ser humano ha utilizado el lenguaje para transmitir conoci-

mientos, pero también sentimientos y emociones. Para la detección por máquinas de sentimientos dentro del texto libre escrito en lenguaje natural es necesario un análisis semántico que permita una comprensión automatizada de los contenidos, su análisis y su aprovechamiento en forma de nuevo conocimiento o como ayuda a la toma de decisiones.

Dentro del análisis semántico aparecen multitud de problemas que son el destino de muchos esfuerzos de investigación hoy día: a) resolución de anáforas o pronombres; b) desambiguación del sentido de las palabras (polisémicas) en función del contexto en el que se encuentren; c) roles semánticos, ya que el significado de una sentencia no se basa sólo en las palabras que la componen, sino también en el orden, agrupación y relaciones entre ellas.

Relacionado con el análisis de sentimientos aparece el descubrimiento de opiniones. Esta aplicación trata de determinar la actitud del autor con respecto a un tema. La actitud puede ser bien su juicio o evaluación, su estado afectivo o la comunicación emocional que se pretende a la hora de escribir un texto. El crecimiento y la disponibilidad de recursos y web sociales donde se expresan opiniones (blogs, foros, comercio electrónico, catálogos de productos, etc.) hacen surgir nuevas oportunidades y retos para obtener esa información de manera estadística y su posterior aplicación a la toma de decisiones, por ejemplo.

La desambiguación semántica de las palabras es otro de los problemas que están tomando gran interés hoy día. Se trata de identificar qué sentido de todos los posibles es el que toma una palabra (polisémica) en una determinada sentencia. La investigación ha avanzado firmemente en este problema en la última década, utilizándose diversas técnicas: a) métodos basados en diccionarios (por ejemplo WordNet); b) métodos de aprendizaje automático supervisado, en los cuales se entrena un clasificador para cada palabra en un corpus de ejemplos anotados manualmente con todos los sentidos de dicha palabra; c) métodos de aprendizaje no supervisados que agrupan (*cluster*) las ocurrencias de las palabras, induciendo por tanto, los distintos sentidos de las palabras. A día de hoy, los métodos que están dando mejores resultados son los algoritmos de aprendizaje supervisados, donde se está consiguiendo una precisión del 90% en lengua inglesa.

Una aplicación inmediata de la desambiguación semántica son los motores de búsqueda conceptuales. Un problema de los actuales motores de búsqueda por palabras clave es que no identifican el sentido de dichas palabras. Por ejemplo, cuando se le pide a un motor de búsqueda tradicional que busque documentos con la palabra "banco", no va a distinguir entre documentos que hablen de instituciones financieras, de otros documentos que hablen de bancos de atunes, o de documentos que hablen del banco del parque.

Muchas compañías e instituciones en diferentes áreas de conocimiento están invirtiendo en la investigación y el desarrollo de aplicaciones que involucran PLN. Una de las áreas más activas que se pueden mencionar es el campo de la biomedicina. Las aplicaciones PLN se están aplicando actualmente en los dominios de la biología y la biomedicina (no sólo en tecnologías NER para identificar proteínas y nombres de genes, sino también utilizando las colecciones de documentos y las técnicas de interpretación descritas antes para abarcar la inmensa cantidad de literatura existente). Un creciente desarrollo de tecnologías PLN se están llevando a cabo también en el dominio clínico, de particular interés para pacientes, ya que el diagnóstico y calidad de los tratamientos depende fuertemente de la historia del paciente descrita en texto libre, no estructurado, almacenado en los informes médicos [32].

Por otro lado, las compañías de seguros médicos privadas están también aplicando técnicas de PLN donde la evaluación de potenciales clientes es clave [33].

En proyectos de ingeniería del software, típicamente ocupados en el análisis cuantitativo del código fuente, podrían también beneficiarse del análisis del conocimiento acumulado en la forma de información lingüística durante los ciclos de desarrollo del software: listas de correo, repositorios de documentación, comentarios del código fuente, sistemas de seguimiento de errores, sistemas de control de versiones, trazas, etc. Actualmente se están intentando aplicar estas técnicas en las forjas de nueva generación que analizarán e interconectarán toda esta información para mejorar la calidad en la producción del software.

Otras aplicaciones tradicionales que se están beneficiando de las técnicas de PLN son los negocios que tratan con la satisfacción de los clientes. Aquí los avances en clasificación automática de textos y métodos de *clustering* pueden evitar la evaluación manual y continua de conjuntos de encuestas masivas, que muchas veces ni siquiera se realiza por la necesidad inmediata de los resultados que no se puede conseguir en un etiquetado manual.

5. Mejorando los datos: la Web como corpus y otros recursos externos

La investigación y desarrollo en lingüística computacional, algoritmos y modelos de aprendizaje automático, selección de atributos y otras áreas de PLN, está avanzando de forma continuada en los últimos años. Un buen lugar para estar al día de estos avances es el repositorio abierto de artículos del ACL Journal [34] y también otros *proceedings* de conferencias en este área de investigación [35] [36] [37].

Pero en las siguientes secciones nos gustaría centrarnos en nuevas tendencias en el uso de los mecanismos de PLN disponibles hoy día o en un futuro cercano. El uso generalizado de los ordenadores y las redes globales de comunicación ha propiciado que la mayoría de nuestros intercambios en lenguaje natural (email, mensajería instantánea, informes, documentación e incluso ideas y hobbies) se codifiquen y almacenen en formato digital, compartido a través de los ordenadores. La disponibilidad de estos datos lingüísticos masivos hace que podamos especular si la era de los corpus estáticos compilados con gran esfuerzo manual (transcripciones del lenguaje oral, material textual, escaneado y OCR, corrección) está llegando a su fin.

En 2006 (un año después del primer *Web as Corpus Workshop at Corpus Linguistic* WACWCL [38] organizado por el *Special Interest Group of the Association for Computational Linguistic* ACL [39]), el blog de investigación de Google publicó el siguiente anuncio [40]:

En Google Research hemos estado utilizando modelos de palabras y n-gramas en muchos proyectos de investigación, tales como traducción estadística, reconocimiento de voz, corrección ortográfica, detección de entidades, extracción de información y otros. Tales modelos se han obtenido a partir de corpus de entrenamiento que contienen alrededor de un billón de palabras. Hemos aprovechado la potencia del centro de datos de Google y su infraestructura distribuida para procesar corpus de entrenamiento cada vez más grandes. Nos hemos dado cuenta de que los mejores datos son "más datos", y por eso hemos incrementado el tamaño del corpus en un orden de magnitud, y luego otro orden más, y más... resultando un corpus de entrenamiento de un trillón de palabras obtenido de todas las páginas web [...] Y por esto es por lo que hemos decidido compartir este enorme conjunto de datos con todo el mundo.

Ese mismo año, se empezó a distribuir un corpus con 24 GB de *tokens* en lengua inglesa: "Web 1T 5-gram" [41]. En 2009, un año después del cuarto WACWCL [42] se hizo disponible públicamente el "1Web 1T 5-gram, 10 European Languages" [43]. Estos datos complementaron los de 2006 en inglés con las lenguas checa, alemana, francesa, italiana, polaca, portuguesa, española, rumana, sueca. En ese momento, la web-como-(multilenguaje)-corpus se hizo realidad.

Las herramientas y técnicas de PLN podrán beneficiarse claramente de éstos y otros corpus gigantescos que están apareciendo. Aunque siempre es deseable que estos datos estén convenientemente anotados, sin embargo, la tendencia es hacia "*más palabras y menos anotación lingüística*". Las palabras de Sinclair en 1992 siguen siendo dignas de lectura hoy día: "*Según el tamaño del corpus va creciendo a cientos de millones [...] el análisis se debería hacer en tiempo real [...] mantengamos el texto en formato plano y hagamos un análisis cada vez que sea necesario*" [44]. El equipo de Google Research también se ha pronunciado en este sentido [45]:

Elijamos una representación que pueda utilizar aprendizaje no supervisado sobre los datos no etiquetados, la cual es mucho más abundante y disponible que los datos etiquetados. Representemos todos los datos con un modelo no parametrizado más que intentar resumir con un modelo parametrizado, porque en fuentes de datos muy extensas los datos contienen mucho detalle. Para aplicaciones de lenguaje natural, confiemos que el lenguaje humano ha evolucionado las palabras para los conceptos importantes. Veamos en qué medida podemos atar las palabras que ya están allí, más que inventar nuevos conceptos con grupos de palabras. Ahora recojamos algunos datos, y veamos qué podemos hacer.

De esta forma, el procesamiento en tiempo real de datos masivos en formato plano parece ser la tendencia en PLN. Mejorar los datos con recursos externos adicionales en tiempo real, es un enfoque perfectamente válido para algunas de las tareas que tratan con el lenguaje. En el resto de esta sección revisaremos fuentes de conocimiento externas que están empezando a explotarse para mejorar y enriquecer los procesos de PLN:

1) Los WordNets o diccionarios de conceptos mencionados anteriormente, hacen posible mejorar los términos léxicos con la identificación de sus posibles sentidos y sus relaciones conceptuales. Sin embargo, los valiosos pero complejos detalles lingüísticos de los conjuntos de sinónimos de WordNet (en ocasiones, un único y simple término puede tener 20 o más sentidos) hacen difícil utilizarlos para anotación semántica, desambiguación de sentidos y otras tareas más especializadas como búsquedas mejoradas (utilizando sinónimos, antónimos, merónimos, holónimos, hiperónimos e hipónimos) o expansión multilenguaje de las consultas. La integración de recursos externos adicionales, tales como ontologías [46] [47], son ya una realidad en proyectos como MEANING [48], un repositorio de conocimiento léxico multilenguaje, que ya está disponible para utilización con aplicaciones de PLN [49].

2) Los procedimientos de PLN venideros combinarán el conocimiento interno extraído de los documentos (definidos en su más amplio sentido) con la información externa disponible, no sólo a través de bancos de datos especializados o diccionarios semánticos y conceptuales multilenguaje (como se ha mencionado antes), sino, y quizás más importante, a través de APIs disponibles para otras muchas fuentes de conocimiento globales y típicamente no estructuradas. Algunos ejemplos de tales fuentes (principalmente compuestas por contenido proporcionado por individuos en sus propios idiomas) que son buenas candidatas para la integración con sistemas de PLN son:

a) Blogs personales y especializados (agrupados por categorías en servicios de directorios de blogs) normalmente etiquetados con categorías o con estados de ánimo. Yahoo! [50], Google [51] y algunos directorios de blogs [52] han implementado APIs muy potentes para extraer la información que contienen. Existen investigaciones muy interesantes que los utilizan en el análisis de sentimientos y estados de ánimo [53] [54].

b) Los proveedores de noticias online son la alternativa profesional a los blogs. Igual que los proveedores de blogs, éstos pueden proporcionar APIs para obtener noticias categorizadas (por país, por sección, por género, etc.). Aunque el acceso a datos masivos y actualizados de proveedores generales de noticias podría mejorar las herramientas de PLN, las noticias financieras analizadas en tiempo real podrían ser la prioridad actualmente para clientes y operadores profesionales [55].

c) Wikipedia ya ofrece acceso a millones de artículos enciclopédicos en muchos idiomas. Los artículos no están categorizados, pero es relativamente fácil encontrar una misma entrada traducida o adaptada a cualquiera de los lenguajes de la Wikipedia. MediaWiki (el software que soporta la Wikipedia) además de ofrecer descargas completas de la enciclopedia completa, ofrece APIs para interactuar en tiempo real con su contenido. En un futuro cercano esperamos que la Wikipedia y sus usuarios adopten la anotación semántica [56], lo cual facilitará la extracción de entidades de forma no ambigua (tales como lugares, personas u organizaciones) para utilización en PLN.

d) Los repositorios de libros de dominio público también ofrecen información lingüística de forma masiva. Como se describe en [57] la búsqueda en libros de Google es un programa muy ambicioso para permitir explorar en todos los libros del mundo. Además de Google Books, el proyecto Gutenberg [58] ofrece unos 100,000 libros que se pueden descargar, no sólo para leer, sino también y más importante, para utilización como corpus lingüístico.

e) Existen otros recursos muy interesantes para el futuro: redes sociales tales como Facebook, MySpace o Twitter, ya que sus *posts* pueden buscarse de forma pública. Y por supuesto los APIs para acceder a comercios electrónicos como Amazon para obtener descripciones de productos, así como comentarios y revisiones de los clientes.

6. El futuro de las aplicaciones de PLN

Además del tratamiento de datos externos descrito en la sección anterior, hay otras áreas en las cuales el PLN jugará un papel importante:

- Motores de búsqueda semántica y conceptual. Aunque los motores de búsqueda resultan ya extremadamente acertados, usando las tradicionales consultas por palabras clave y ordenando los resultados según su ranking, para ir más allá del paradigma de coincidencia de palabras clave se requieren métodos de PLN, donde los usuarios puedan indicar qué quieren realmente buscar. Esto es, comprender y expandir el significado y las relaciones de las palabras en la cadena de búsqueda. También se necesitan técnicas de PLN para pre-procesar los datos presentados como resultados para estar seguros de que su contenido satisface lo que el usuario deseaba cuando realizó la búsqueda. WordNet y mapeo de ontologías [59], así como la integración de fuentes externas de conocimiento presentadas anteriormente, combinados con los avances en algoritmos de comprensión del lenguaje natural ayudarán a mejorar los motores de búsqueda, así como complementar los mecanismos de ordenación (ranking) de los resultados.
- Además de mejorar los motores de búsqueda con capacidades a nivel semántico y conceptual, los procedimientos de PLN llevarán a nuevas tendencias en visualización de los resultados de búsqueda: resúmenes automáticos, clasificación y *clustering* automático de los documentos por categorías, nubes de palabras mejoradas semánticamente [60], son algunas de las funcionalidades que veremos en un futuro cercano en la página de resultados. Los sistemas realizarán un post-proceso de los documentos cargados y, utilizando técnicas de PLN (métodos de resumen automático y simplificación), serán capaces de extraer las partes relevantes de una consulta. Para un ejemplo de investigación en esta dirección en blogs tenemos [61].
- Un campo donde las técnicas de PLN van a ser aplicadas en el futuro es la ayuda lingüística inmediata, tanto para negocios como para entornos privados. El abrumador y creciente volumen de correos electrónicos requiere el desarrollo de mecanismos de filtrado inteligente, utilizando algo más que lo obvio (asunto, remitente, fecha) que proporcionan los clientes de correo actuales. La comprensión del contenido de los mensajes de correo electrónico y la relación de estos contenidos a otros mensajes previos, requiere motores de PLN avanzados. En la misma línea, la respuesta de correos asistida, o las alertas de predicción de respuestas [62], son buenos

candidatos para integración con PLN.

■ El chino, el inglés, el español y el árabe son lenguas que conjuntamente suman más de 2,000 millones de hablantes en el mundo. Sin embargo, el 94% de las lenguas sólo las hablan el 6% de personas del mundo [63]. Paradójicamente, las lenguas minoritarias y en peligro de extinción podrían beneficiarse de un mundo digitalizado y de aplicar e integrar las tecnologías de PLN disponibles hoy día. Conseguir colecciones de datos lingüísticos es un paso previo que es muy complicado para las lenguas minoritarias. Sin embargo, la traducción automática a/desde lenguas minoritarias se beneficiará de la disponibilidad de datos en formato digital [64]. Muchos contenidos actuales de Wikipedia en distintos idiomas son traducciones manuales de textos introducidos originalmente en inglés o español, haciendo fácil aplicar las técnicas de aprendizaje para desarrollar aplicaciones de traducción automática para estos idiomas. A cambio, la integración de lenguas minoritarias en aplicaciones de PLN permitirá una recopilación más comprensiva de información multilenguaje. Un ejemplo de posibles tendencias de integración para lenguas minoritarias sería Golfiño [65], una solución abierta y gratuita que utiliza técnicas de PLN. Se trata del primer corrector gramatical para la lengua gallega (utilizado en OpenOffice.org), el cual utiliza FreeLing para la fase de etiquetado PoS.

7. Conclusión: el futuro está en movimiento

Finalmente, la expansión y popularización de los dispositivos móviles, así como la necesidad de obtener y generar información relevante requerirá más integración de los sistemas de posicionamiento geoespacial y de las fuentes de conocimiento externas analizadas antes, pero también de contenido multimedia no lingüístico (rodeado habitualmente por texto libre [66]). Es fácil identificar y analizar este contenido multimedia teniendo en cuenta el contexto lingüístico que le rodea, más que confiar en sistemas de identificación de audio/imagen/vídeo, con el fin de incorporarlos en las aplicaciones.

Un buen ejemplo de esta clase de sistemas integrados se puede encontrar en el proyecto LibreGeoSocial [67], una red social abierta y gratuita con interfaz de usuario para dispositivos móviles con realidad aumentada. La integración futura de las técnicas de PLN en este ecosistema, nos parece que es la clave para hacer que la aparición de las aplicaciones de realidad aumentada sea no sólo una realidad de hecho, sino que también suponga una realidad llena de "significado".

Referencias

[1] D. A. Cruse. *Lexical Semantics*. Cambridge [Cambridgeshire]/New York: Cambridge University Press, 1986.

[2] Kenneth Church, Ramesh Patil. "Coping with syntactic ambiguity or how to put the block in the box on the table". *American Journal of Computational Linguistics, 8*:139–149, 1982.

[3] Helen Leckie-Tarry, David Birch (ed.). *Language and Context: A Functional Linguistic Theory of Register*. London/New York: Pinter Publishers, 1995.

[4] George Lakoff, Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 1980.

[5] Stuart Russell, Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2009.

[6] Henry Kuc?era, W. Nelson Francis. *Computational Analysis of Present-Day American English*. Providence, Rhode Island: Brown University Press, 1967.

[7] The Brown Corpus. <http://khnt.hit.uib.no/icame/manuals/brown/index.htm>.

[8] The British National Corpus (BNC). <http://sara.natcorp.ox.ac.uk/>.

[9] The American National Corpus (ANC). <http://americannationalcorpus.org/>.

[10] Corpus de Referencia del Español Actual (CREA). <http://corpus.rae.es/creanet.html>.

[11] The Penn Treebank Project. <http://www.cis.upenn.edu/~treebank/>.

[12] The Reuters Corpora. <http://trec.nist.gov/data/reuters/reuters.html>.

[13] Biomedical corpora. <http://compbio.uchsc.edu/ccp/corpora/obtaining.shtml>.

[14] The Juris Corpus. <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC98T32>.

[15] LFG Pargram Project. <http://www2.parc.com/istl/groups/nltt/pargram/>.

[16] HPSG LinGO Matrix framework. <http://www.delph-in.net/matrix/>.

[17] XTAG Project. <http://www.cis.upenn.edu/~xtag/>.

[18] Christiane Fellbaum (ed). *WordNet: An Electronic Lexical Database*. MIT Press, 1988. <http://wordnet.princeton.edu/>.

[19] EuroWordNet. <http://www.illc.uva.nl/EuroWordNet/>.

[20] Christopher D. Manning, Hinrich Schuetze. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts. The MIT Press, 1999.

[21] Daniel Jurafsky, James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Second Edition). Prentice Hall, 2000.

[22] Steven Bird, Ewan Klein, Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, 2009.

[23] Google Research. <http://research.google.com/about.html>.

[24] Microsoft Research. <http://research.microsoft.com/en-us/groups/nlp/>.

[25] IBM Research. <http://domino.research.ibm.com/comm/research.nsf/pages/r.nlp.html>.

[26] Yahoo! Research. <http://research.yahoo.com/>.

[27] FreeLing. <http://www.lsi.upc.edu/~nlp/freeling/>.

[28] Weka. <http://www.cs.waikato.ac.nz/ml/weka/>.

[29] GATE. <http://gate.ac.uk/>.

[30] NLTK. <http://www.nltk.org/>.

[31] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. An Introduction to Information Retrieval. Cambridge University Press, 2008.

[32] Danielle L. Mowery MS, Henk Harkema PhD, John N. Dowling MS MD, Jonathan L. Lustgarten PhD, Wendy W. Chapman PhD. "Distinguishing Historical from Current Problems in Clinical Reports—Which Textual Features Help?". *Proceedings of the Workshop on BioNLP, 2009*, pp. 10-18.

[33] Barry Glasgow, Alan Mandell, Dan Binney, Lila Ghemri, David Fisher. "MITA: An Information-Extraction Approach to the Analysis of Free-From Text in Life Insurance Applications", *AI Magazine 19(1)* (Spring 1998), pp. 59-72.

[34] Computational Linguistics. <http://www.mitpressjournals.org/loi/coli>. Official Journal of the Association for Computational Linguistics <http://www.aclweb.org/>.

[35] Conference on Computational Linguistics (COLING). <http://www.coling-2010.org/>.

[36] Empirical Methods on Natural Language Processing (EMNLP). <http://conferences.inf.ed.ac.uk/emnlp09/>.

[37] International Conference on Machine Learning (ICML). <http://www.cs.mcgill.ca/~icml2009/>.

[38] Web as Corpus Workshop at Corpus Linguistics. <http://sslmit.unibo.it/~baroni/web_as_corpus_cl05.html>, 2005.

[39] SIGWAC. <http://www.sigwac.org.uk/>.

[40] Google Research Blog. <http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html>.

[41] Thorsten Brants, Alex Franz. *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia, 2006: <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13>.

[42] Stefan Evert, Adam Kilgarriff, Serge Sharoff. *Proceedings of the 4th Web as CorpusWorkshop (WAC-4)*. <http://webascorpus.sourceforge.net/download/WAC4_2008_Proceedings.pdf>.

[43] Thorsten Brants, Alex Franz. *Web 1T 5-gram, 10 European Languages*. Version 1. Linguistic Data Consortium, Philadelphia, 2009: <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13>.

[44] J. Sinclair. "The automatic analysis of corpora". En J. Svartvik (ed.) *Directions in Corpus Linguistics* (*Proceedings of Nobel Symposium 82*). Berlin: Mouton de Gruyter. (pp. 382-384), 1992.

[45] Alon Halevy, Peter Norvig, Fernando Pereira. "The Unreasonable Effectiveness of Data". *IEEE Intelligent Systems, March/April 2009*.

[46] A. Alonge, F. Bertagna, L. Bloksma, S. Climent, W. Peters, H. Rodríguez, A. Roventini, P. Vossen. Top Concept Ontology. "The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology". En Piek Vossen (ed.) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, 1998.

[47] Suggested Upper Merged Ontology (SUMO). <http://www.ontologyportal.org/>.

[48] MEANING. <http://www.lsi.upc.es/~nlp/meaning/>.

[49] Multilingual Central Repositor (MRC). <http://www.lsi.upc.es/~nlp/meaning/downloads.html>.

[50] Yahoo!'s Livejournal API. <http://developer.yahoo.com/>.
[51] Google's Blogger API. <http://code.google.com/intl/en/apis/blogger/>.
[52] BlogCatalog API. <http://www.programmableweb.com/api/blogcatalog>.
[53] Gilad Mishne. "Experiments with Mood Classification in Blog Posts". *Stylistic Analysis of Text for Information Access*, 2005.
[54] Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena. "Large-Scale Sentiment Analysis for News and Blogs". *International Conference on Weblogs and Social Media, ICWSM 2007* (Boulder, Colorado, USA).
[55] M. Costantino, R. G. Morgan, R. J. Collingham, R. Carigliano. "Natural language processing and information extraction: qualitative analysis of financial news articles". *Computational Intelligence for Financial Engineering (CIFEr). Proceedings of the IEEE/IAFE 1997*. pp. 116-122.
[56] Semantic MediaWiki. <http://semantic-mediawiki.org/wiki/Semantic_MediaWiki>.
[57] Luc Vincent. "Google Book Search: Document Understanding on a Massive Scale". *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02, 2007*. pp. 819-823.
[58] Project Gutenberg. <http://www.gutenberg.org/wiki/Main_Page>.
[59] Dario Bonino, Fulvio Corno, Laura Farinetti, Alessio Bosca. "Ontology Driven Semantic Search". *WSEAS Transaction on Information Science and Application 1 (6)* (2004) pp. 1597-1605.
[60] Byron Y-L. Kuo, Thomas Hentrich, Benjamin M. Good, Mark D. Wilkinson. "Tag Clouds for Summarizing Web Search Results". *WWW 2007*, May 8–12, 2007, Banff, Alberta, Canada.
[61] Michel Généreux. "Summarizing a Blog Search Engine Hits". *Workshop on Web Search Result Summarization and Presentation, 2009*, Madrid, Spain.
[62] Mark Dredze, Tova Brooks, Josh Carroll, Joshua Magarick, John Blitzer, Fernando Pereira. "Intelligent Email: Reply and Attachment Prediction". *IUI'08*, January 13-16, 2008, Maspalomas, Gran Canaria, Spain.
[63] M. Paul Lewis (ed). *Ethnologue: Languages of the World*, Sixteenth edition. Dallas, Tex.: SIL International, 2009. <http://www.ethnologue.com/>.
[64] Wikipedia. Asturiano <http://ast.wikipedia.org/wiki/>, Aragonés <http://an.wikipedia.org/wiki/>, Vasco <http://eu.wikipedia.org/wiki/>, Catalán <http://ca.wikipedia.org/wiki/>, Gallego <http://gl.wikipedia.org/wiki/>.
[65] Golfiño. *Corrector gramatical de galego para OpenOffice.org* <http://www.imaxin.com/ficha.asp?IDproxecto=68>.
[66] Hrishikesh Aradhye, George Toderici, Jay Yagnik. "Video2Text: Learning to Annotate Video Content", *ICDM Workshop on Internet Multimedia Mining 2009*.
[67] LibreGeoSocial, GSyC/LibreSoft (URJC). <http://libregeosocial.morfeo-project.org/>.