La enseñanza del Procesamiento del Lenguaje Natural en facultades de Informática y Filología

Miguel A. Alonso Pardo[1], **Margarita Alonso Ramos** [2], **Carlos Gómez Rodríguez**[1], **Jorge Graña Gil**[1], **Jesús Vilares Ferro**[1]

[1]Facultad de Informática y [2]Facultad de Filología

Grupo de Lengua y Sociedad de la Información, Universidade da Coruña

{alonso, lxalonso, cgomezr, grana, jvilares} @ udc.es

Introducción

El Procesamiento del Lenguaje Natural (PLN), también en ocasiones referido como Lingüística Computacional [Jurafsky y Martin, 2009; Mitkov, 2005] es la disciplina encargada del diseño e implementación de los elementos software necesarios para el tratamiento computacional del lenguaje natural, entendiendo como tal todo lenguaje humano, en contraposición a los lenguajes formales propios del ámbito lógico, matemático o computacional. El objetivo último que se persigue, y que todavía está muy lejos de conseguirse, es el de la comprensión y la producción o generación del lenguaje humano por parte del ordenador.

El PLN tiene un carácter marcadamente interdisciplinar, ya que su tratamiento implica combinar conceptos lingüísticos –en particular aquellos relacionados con el análisis morfológico, el análisis sintáctico y el análisis semántico– con conceptos informáticos referidos al diseño de algoritmos eficientes sobre estructuras de datos complejas. Por ello, el PLN se imparte tanto en titulaciones universitarias de las facultades y escuelas de Informática como en las de las facultades de Filología. Aunque en ambos casos los contenidos básicos son muy cercanos, su presentación varía en gran medida según los conocimientos y experiencias previas de los alumnos en cada ámbito [Moore y otros, 2007]. En este contexto, este artículo relata la experiencia de un grupo de profesores en la impartición de materias sobre PLN en las facultades de Informática y Filología de la Universidade da Coruña (UDC).

La enseñanza del PLN en una facultad de Informática

En la titulación de Ingeniería Informática de la UDC, la enseñanza de la disciplina del PLN se ha encomendado a la asignatura *Lenguajes Naturales*, optativa de segundo ciclo con una carga lectiva de 4 créditos ECTS. En su diseño se han seguido las directrices del *Computing Curricula* [IEEE y ACM, 2008], que establece el PLN como unidad IS7 del área de Sistemas Inteligentes (IS). Se ha pretendido dar a la asignatura una orientación práctica, estructurando el temario de manera que la primera mitad de la asignatura se centra en los aspectos teóricos del PLN, estudiando los fenómenos lingüísticos a nivel léxico, morfológico, sintáctico y semántico así como los formalismos computacionales para su representación; mientras que la segunda mitad se dedica a sus aspectos prácticos y aplicativos mediante la realización de algoritmos y programas de ordenador que transforman los fundamentos teóricos en aplicaciones reales como son la Recuperación de Información [Manning y otros, 2008], la Extracción de Información [Moens, 2006] y la Búsqueda de Respuestas [Pasca, 2003].

Debido al carácter práctico dado a la materia, la evaluación de los alumnos se realiza mediante mecanismos distintos a los exámenes [Brown, 2003]. Por una parte se enfatiza el ejercicio de competencias más próximas a las necesidades del mundo laboral mediante la realización de un proyecto para el desarrollo de un sistema de NLP, tal y como se propuso en [Alonso y otros, 2010]. Por otra parte, y de cara a una evaluación más eficaz que no tenga en cuenta únicamente el resultado final, cada grupo de prácticas debe elaborar un diario de trabajo a modo de portfolio [Olén y otros, 2006] documentando el trabajo realizado a lo largo de todo el proceso de desarrollo: documentación previa acerca de las técnicas a emplear, decisiones de diseño e implementación y su justificación, ensayos –tanto exitosos como fallidos–, análisis de sus resultados, etc.

La enseñanza del PLN en una facultad de Filología

La enseñanza del PLN en la Facultad de Filología de la UDC se articula mediante la asignatura *Lingüística e Informática*, optativa de segundo ciclo de carácter no específico, por lo que puede ser cursada en las titulaciones de Filología Hispánica, Filología Gallega y Filología Inglesa. Su carga lectiva es de 4 créditos ECTS. El temario diseñado comienza por una introducción al PLN en general y a los corpus electrónicos en particular, como una manera de introducir gradualmente a los alumnos en el ámbito de las tecnologías de la lengua. Posteriormente se presentan los niveles de procesamiento de la lengua, empezando por el Léxico, la Morfología, la Sintaxis y la Semántica, haciendo hincapié en cómo el conocimiento de la lengua que poseen los alumnos se puede formalizar y explicitar de tal modo que sea utilizable por un ordenador. Sin embargo, a diferencia de la asignatura impartida en Informática, en Filología se le da más peso al tema del léxico puesto que se enseña la idea de que una entrada lexicográfica recorre todos los niveles lingüísticos: desde el semántico hasta el

fonético. Finalmente, se muestran brevemente algunas aplicaciones del PLN (Recuperación de la Información, Resumen Automático [Mani, 2001] y Aprendizaje de Lenguas Asistido por Ordenador [Cal y otros, 2005]), dedicando especial atención a la Traducción Automática (TA) [Hutchins y Somers, 1995] porque hace uso de todos los conocimientos lingüísticos estudiados en el curso. La TA, además de ser la pionera en el campo de la Lingüística Computacional, sigue siendo de actualidad para los estudiantes de Filología, que pueden evaluar distintos sistemas que funcionan libremente en la web.

La evaluación se realiza por medio de actividades prácticas correspondientes a cada tema que se van exponiendo oralmente y que al final de curso se deben entregar por escrito [Olén y otros, 2006]. Aunque la parte práctica de la asignatura no tiene un componente de implementación tan acusado como en el caso de la titulación de Informática, se incluye también una pequeña práctica de programación con el lenguaje Prolog [Gazdar y Mellish, 1989]. Esta incursión en el ámbito de la programación resulta especialmente positiva para los estudiantes pues supone un importante factor de motivación que ellos mismos aprecian.

Discusión

Las dos asignaturas coinciden en gran parte de los epígrafes del temario, aunque difieren en el enfoque, un reflejo de la tensión existente entre las aportaciones de lingüistas e informáticos al PLN, ya que mientras los primeros están más interesados en definir una teoría completa de la lengua, los segundos están más interesados en la definición e implementación de algoritmos con una complejidad computacional tratable que permitan resolver problemas prácticos de uso de la lengua de forma eficaz, aunque ello suponga utilizar formalismos con una limitada capacidad de descripción de los fenómenos lingüísticos.

A este respecto resulta curioso que si bien los alumnos de Informática han estudiado en la titulación las aportaciones de N. Chomsky a la Teoría de Lenguajes Formales, sin embargo desconocen que este es un lingüista, y que los formalismos que han estudiado para su aplicación en compilación, englobados en la denominada Jerarquía de Chomsky [Hopcroft y Ullman, 1979], fueron diseñados inicialmente para servir de base formal en la descripción de la sintaxis de los lenguajes naturales. Este caso ejemplifica el hecho de que la interdisciplinariedad del PLN no viene dada solo porque sea preciso utilizar conocimiento lingüístico en el desarrollo de soluciones informáticas en este ámbito, sino también porque investigadores del ámbito humanístico, han realizado contribuciones relevantes a la Teoría de la Computación.

Otro aspecto que puede resultar paradójico es que mientras que la mayoría de los libros de texto de PLN de orientación informática siguen considerando las gramáticas independientes del contexto como punto de partida inevitable para la realización del análisis sintáctico en sistemas prácticos, en el caso de los lingüistas estos han abandonando dicho formalismo hace décadas. Sin embargo,

en los últimos años parece estar produciéndose también un cambio de paradigma en la parte informática del PLN con el auge de los analizadores basados en dependencias [Gómez Rodríguez, 2010; Kübler y otros, 2009]. Para explicar los fundamentos de estos analizadores, debemos abandonar las estructuras de constituyentes, que son las que los alumnos han utilizado en su formación preuniversitaria para la realización del análisis sintáctico de oraciones, para pasar a adoptar estructuras lingüísticas basadas en dependencias, como las sugeridas por la Teoría Sentido-Texto [Mel'čuk, 2009].

Conclusiones

Hemos mostrado la experiencia de un grupo de profesores en la impartición de la disciplina de PLN en titulaciones del ámbito tecnológico y humanístico. La experiencia es positiva, con una tasa de aprobados sobre presentados cercana al 100% en ambas asignaturas.

En el caso de los estudiantes de Filología, cursar la asignatura les ayuda a comprender el carácter aplicado de la lingüística y sus crecientes posibilidades en un mundo con un constante crecimiento en el número de "nativos digitales", fruto de una sociedad de la información que demanda cada vez más aplicaciones avanzadas en las que la lengua, en tanto que medio de expresión básico, tiene un rol principal y donde, consecuentemente, humanidades y tecnología tienden a converger. Por su parte, en el caso de los alumnos de Informática, la docencia en PLN fomenta su contacto con otras disciplinas, aspecto muy importante si tenemos en cuenta que la informática es una disciplina de carácter eminentemente aplicado, y de la que podríamos afirmar que da soporte a las demás ramas del saber, por dispares que sean, ya que actualmente una gran parte de los avances en el conocimiento requieren de soluciones informáticas, bien para su simulación, bien como herramienta de apoyo a su desarrollo, bien directamente para su implantación. Ello hace que un gran número de ingenieros informáticos, en el ejercicio de su profesión, trabajen en equipos interdisciplinares, aspecto que sin embargo no es suficientemente fomentado en los planes de estudio.

Referencias

Alonso, M. A., Fernández, M., Gómez-Rodríguez, C, Graña, J., Molinero, M.A., Vilares, J. (2010). Evaluación sin exámenes. Conclusiones de 10 años de experiencia en una asignatura optativa. En *La innovación educativa en el contexto actual de la educación superior. A innovación educativa no contexto actual da educación superior* (pp. 515-518). Vigo: Vicerreitoría de Formación e Innovación Educativa, Universidade de Vigo.

Brown, S. (2003). Aplicaciones prácticas de una evaluación práctica. En *Evaluar en la Universidad: Problemas y nuevos enfoques* (pp. 117-128). Madrid: Narcea.

Cal, M., Núñez, P., Palacios, I. M. (Eds.). (2005). *Nuevas tecnologías en Lingüística, Traducción y Enseñanza de lenguas.* Santiago de Compostela: Universidad de Santiago de Compostela, Servizo de Publicacións e Intercambio Científico..

Gazdar, G., Mellish C. (1989). *Natural language processing in Prolog*. Wokingham: Addison- Wesley.

Gómez Rodríguez, C. (2010). *Parsing schemata for practical text analysis*. London: Imperial College Press.

Hopcroft, J. E., Ullman, J. D. (1979). *Introduction to Automata Theory, Languages and Computation*. Reading (MA): Addison-Wesley.

Hutchins, W. J., Somers, H. L. (1995). *Introducción a la traducción automática.* Madrid: Visor.

IEEE y ACM. (2008). *Computer Science Curriculum 2008: An Interim Revision of CS 2001*, Report from the Interim Review Task Force. New York: IEEE Computer Society and Association for Computing Machinery.

Jurafsky, D., Martin, J. H. (2009). *Speech and Language Processing. 2nd Edition.* Upper Saddle River (NJ): Pearson Education.

Kübler, S., McDonald, R., Nivre, J. (2009). *Dependency Parsing*. San Rafael (CA): Morgan & Claypool Publishers.

Mani, I. (2001). *Automatic summarization*. Amsterdam: John Benjamins.

Manning, C. D., Raghavan, P., Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Mel'čuk, I. A. (2009). Dependency in natural language. En A. Polguère, I. A. Mel'čuk (Eds.), *Dependency in Linguistic Description* (pp. 1-110). Amsterdam: John Benjamins.

Mitkov, R. (Ed.). (2005). *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

Moens, M.-F. (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context*, Dordrecht. Springer.

Moore, S., Walsh, G., Risquez, A. (2007). *Teaching at College and University. Efective strategies and key principles.* London: Open University Press McGraw-Hill Education.

Olén, M. T., Giné, N., Imbernón, F. (2006). *La carpeta de aprendizaje del alumnado universitario*. Barcelona: Octaedro-ICE.

Pasca, M. (2003). *Open-domain question answering from large text collections*. Stanford: CSLI Publications.