View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repositorio Institucional de la Universidad de Alicante

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 189-192 recibido 14-04-14 revisado 02-07-14 aceptado 02-07-14

imaxin|software: PLN aplicada a la mejora de la comunicación multilingüe de empresas e instituciones

imaxin|software: NLP applied to enhance multilingual communications for public organisms and companies

José Ramom Pichel, Diego Vázquez, Luz Castro, Antonio Fernández
imaxin|software
Rua Salgueirinhos de abaixo N11 L6, Santiago de Compostela
e-mail: {jramompichel,diegovazquez,luzcastro,afernandez}@imaxin.com

Resumen: imaxin|software es una empresa creada en 1997 por cuatro titulados en ingeniería informática cuyo objetivo ha sido el de desarrollar videojuegos multimedia educativos y procesamiento del lenguaje natural multilingüe. 17 años más tarde, hemos desarrollado recursos, herramientas y aplicaciones multilingües de referencia para diferentes lenguas: Portugués (Galicia, Portugal, Brasil, etc.), Español (España, Argentina, México, etc.), Inglés, Catalán y Francés. En este artículo haremos una descripción de aquellos principales hitos en relación a la incorporación de estas tecnologías PLN al sector industrial e institucional.
Palabras clave: Big Data, Recursos lingüísticos, Análisis de Sentimientos, Minería de Opiniones, Traducción automática, Servicios online con herramientas PLN de código abierto, Aprendizaje de idiomas asistidos por ordenador.

Abstract: imaxin|software is a company created in 1997 by four computer engineers with the aim of developing educational multimedia games and natural language processing tools. After 17 years **imaxin**|software has developed resources, tools and applications for different languages, specially for Portuguese (Galiza, Portugal, Brazil, etc.), Spanish (Spain, Argentina, México, etc.), English, Catalan, French. In this article we will describe the main highlights of this technological and human challenge.
Keywords: Big Data, Language Resources, Sentiment Analysis, Opinion Mining, Machine Translation, Online services using Open-source NLP tools, Computer Aided Language Learning.

1 Introducción

imaxin|software es una empresa dedicada al desarrollo de servicios y soluciones avanzadas de software y multimedia desde el año 1997, especializada en ingeniería lingüística y videojuegos multimedia educativos y formativos (Serious Games, Gamification) (Pichel et al., 2013).

imaxin|software que inicialmente estaba constituída por cuatro socios-trabajadores, tuvo en plantilla hasta veintiseis personas en el año 2010. Las ventas de productos y servicios se han repartido entre público y privado en porcentajes aproximadas de 60 %-40 % variando de año en año entre un sector y otro. Nos centraremos en la primera línea de desarrollos, **imaxin**|software es desde el año 2000 proveedor de tecnología lingüística para Microsoft. Además, podemos destacar entre los principales desarrollos en PLN los sistemas de corrección ortográfica, gramatical, estilística; sumarizadores de textos, sistemas de opinion mining, pesquisa semántica, sistemas de codificación médica de historias clínicas, detección automática de entidades (NER), así como la plataforma líder europea en traducción automática de código abierto: Opentrad (con sus de los motores de traducción Apertium y Matxin) – www.opentrad.com.

ISSN 1135-5948
© 2014 Sociedad Española para el Procesamiento del Lenguaje Natural

2 Principales proyectos PLN aplicados a las necesidades empresariales

2.1 Construcción y uso de recursos linguísticos

- **Corrección ortográfica en red Galgo.NET (2001)**:

  El corrector imaxin Galgo.NET es uno de los primeros correctores desarrollados específicamente para la correción ortográfica multilingüe simultánea (gallego y español) para redacciones de periódicos. **imaxin**|software ha desarrollado toda esta tecnología incluyendo un tries propio para compresión de diccionarios (Malvar y Pichel, 2010).

- **Corrección de lenguaje sexista (Exeria)**:

  Hemos desarrollado un corrector para OpenOffice.org que mejora los textos en gallego ofreciendo textos con lenguaje no sexista. **imaxin**|software ha desarrollado toda la tecnología de corrección estilística e integración en Openoffice.org.

- **Coruxa Biomedical Text Mining**:

  Extractor y codificador automático de información médica relevante mediante el uso del PLN. Financiado por la Dirección Xeral de I+D+i (Xunta de Galicia). Investigador principal: **imaxin**|software, USC-GE, IXA Taldea, Doctor QSolutions. Transferencia al sector industrial: servicio de codificación SNOMED-CT para historias clínicas. **imaxin**|software ha desarrollado toda la tecnología de codificación. IXA Taldea ha desarrollado un anonimizador de historias clínicas y la USC-GE el procesamiento de ontologías.

2.2 Optimizadores semánticos de búsquedas

- **Optimizador de búsquedas en bibliotecas mediante ontología (2008)**:

  El objetivo del módulo Optimizador es expandir las búsquedas efectuadas por los/as usuarios/as en los sistemas de consulta bibliográfica del CSBG (Centro Superior Bibliográfico de Galicia) mediante el uso de ontologías construída adhoc a partir de un corpus construído de bibliografía. Está integrado con el software de gestión de bibliotecas de código abierto Koha. **imaxin**|software ha desarrollado toda la tecnología.

2.3 Servicios online mediante el uso de herramientas de PLN de código abierto

- **Traductor de documentos online (www.opentrad.com)**:

  Existe un servicio en línea de e-commerce de Opentrad para traducir documentos entre diferentes lenguas y manteniendo en todo momento el formato original. **imaxin**|software ha desarrollado toda la tecnología web.

- **Traductor de documentos en la Aplicateca de Telefónica**:

  Desde el año 2012 está instalado el traductor de documentos Opentrad especial para PYMES en la Tienda Cloud Aplicateca de Telefónica. **imaxin**|software ha desarrollado toda la tecnología de integración en Aplicateca.

2.4 Traducción automática de código abierto

- **Opentrad: plataforma de servicios de traducción de código abierto (2004-2014)**

  Opentrad (Alegría et al., 2006) es la plataforma de traducción automática en código abierto pionera en el mercado español (www.opentrad.com). Este proyecto se inició en el año 2004, siendo el resultado de diferentes proyectos de I+D+i (PROFIT y Avanza del Ministerio de Industria) desarrollados por un consorcio formado por Universidades y Empresas (Transducens-UA, Eleka, Elhuyar, IXA Taldea, TALP (UPC), **imaxin**|software y SLI-Universidade de Vigo). Como resultado de este proyecto se constituyó una spin-off especialista en uno de los motores del proyecto (Apertium), Prompsit Language Technologies.

  Opentrad está formada por dos ingenios de traducción de código abierto (Apertium y Matxin). Opentrad mejora la comunicación multilingüe, permite publicar información en diferentes idiomas, reduce costes y tiempos de revisión humana, permitiendo incluso la mejora de

los tiempos en la localización de versiones multilingües de aplicaciones empresariales.

Opentrad está o estuvo implantado en administraciones, empresas y portales de Internet traduciendo millones de palabras diariamente (Ministerio de Administraciones Públicas, Xunta de Galicia, Universidades Públicas Gallegas, La Voz de Galicia, Faro de Vigo, Instituto Cervantes, Kutxa, Eroski, etc.)

La mejoría continua de los de los ingenios de traducción (Apertium y Matxin) permite ofrecer sobre todo, una mejor calidad entre lenguas próximas (Español-Francés Español-Portugués, Español-Portugués do Brasil, Español-Catalán, Español-Gallego, etc.) que otros traductores automáticos.

imaxin|software ha desarrollado todas las tecnologías para integrar los prototipos de Opentrad en cliente final y la mejora lingüísticas de los recursos de traducción automática especificamente para los pares español-galego, español-portugués y español-inglés (Pichel et al., 2009).

2.5 Análisis de sentimientos y minería de opinión para un seguimiento de marca inteligente y análisis Big Data

En este campo hemos desarrollado en el año 2009 un prototipo inicial (Coati) de análisis de sentimientos. En la actualidad, hemos trasladado esta experiencia a un proyecto más ambicioso relacionado con el Análisis de sentimientos y minería de opinión relacionado con el Big Data. Explicaremos cada uno de ellos en detalle:

- **Coati Opinion mining (2009)**:

  En este proyecto hemos investigado como extraer automáticamente de blogs opiniones y tendencias interesantes para el ámbito empresarial y la administración pública (2009) mediante el uso de técnicas de Opinion Mining. **imaxin**|software ha desarrollado toda la tecnología del crawler y el corpus de entrenamiento del Opinion Mining basado en support vector machine (Malvar y Pichel, 2011). Pendiente de evaluación.

- **CELTIC: Conocimiento Estratégico Liderado por Tecnologías para la Inteligencia Competitiva (FEDER-INNTERCONECTA)**:

  El proyecto, actualmente en desarrollo, está orientado al campo de la vigilancia tecnológica y el Social Media Marketing mediante el uso de PLN y el procesamiento en Big Data.

  Este proyecto ha sido financiado mediante los fondos tecnológicos europeos para regiones objetivo 1 de la Unión Europea. Estos fondos conocidos como FEDER INNTERCONECTA, son proyectos Integrados de desarrollo experimental altamente competitivos, con carácter estratégico, de gran dimensión y que tienen como objetivo el desarrollo de tecnologías nuevas en áreas tecnológicas de futuro con proyección económica y comercial a nivel internacional, suponiendo a la vez un avance tecnológico e industrial relevante para las autonomías destinatarias de las ayudas, como es el caso de Galicia.

  imaxin|software consiguió en el año 2012 este proyecto con un consorcio formado polas siguientes empresas y Universidades: Indra, Elogia, SaecData, Gradiant, USC-PRONATL (USC), Computational Architecture Group (USC).

  El objetivo del proyecto es el desarrollo de tecnologías capacitadoras que faciliten al tejido empresarial la toma de decisiones estratégicas en tiempo casi-real, a partir del conocimiento tanto del medio científico-tecnológico como de los impactos económicos presentes y futuros. O lo que es el mismo, el desarrollo de tecnologías capacitadoras para la Inteligencia Competitiva en las organizaciones.

  Las tecnologías a desarrollar durante el proyecto cubren el proceso completo de la Inteligencia Competitiva, en sus respectivas fases: agregación de información, análisis de la información extrayendo de ella el conocimiento necesario, y la distribución mediante mecanismos de visualización e iteracción avanzados para facilitar la toma de decisiones estratégicas.

  El ámbito de aplicación es el Social Media Marketing y la Vigilancia tecnológi-

ca. En el primero, la competitividad actual genera la necesidad de disponer de sistemas de monitorización inteligente y en tiempo real de redes sociales y análisis del impacto de los productos de una marca determinada en el consumidor (Gamallo, Garcia, y Pichel, 2013). Esto puede ser posible mediante la integración de tecnologías avanzadas de procesamiento del lenguaje natural y tecnologías semánticas.

En el campo de la Vigilancia tecnológica los los desarrollos a realizar en este proyecto permitirán el acceso y gestión en tiempo real de los conocimientos científicos y técnicos a las empresas, así como la información más relevante sobre su contexto, junto a la comprensión a tiempo del significado e implicaciones de los cambios y novedades.

imaxin|software ha desarrollado en colaboración con Indra, la USC-GE y USC-CA todos los desarrollos de PLN integrados en Big Data. Todas los desarrollos están pendientes de evaluación al final del proyecto.

2.6 Aprendizaje de Lenguas asistido por Ordenador (Juegos y Lexicografía)

Por último hemos desarrollado el "Portal das palabras" en el año 2013, una web educativa que pone en valor el diccionario de la Real Academia Galega mediante juegos relacionados con las palabras para un mejor aprendizaje del gallego por el público en general y sectores más distantes de la lengua como el mundo empresarial.

Con el Portal de las Palabras no solo podemos mejorar nuestra competencia en idioma gallego sino que también aprenderemos jugando. Incluye también el diccionario de la RAG con búsquedas de lemas y sinónimos, videos explicativos y guías didácticas para la lengua.

imaxin|software ha desarrollado toda la tecnología PLN y web en este proyecto.

3 Conclusiones

Este artículo pretende mostrar por un lado un mosaico de tecnologías PLN (productos, servicios y proyectos de I+D) de más de 17 años de una pequeña empresa, y por otro la importancia que para este fin ha tenido la colaboración y transferencia con los organismos públicos de investigación (Universidades y Centros Tecnológicos).

Bibliografía

Alegría, I., I. Arantzabal, M Forcada, X. Gómez-Guinovart, L. Padró, J. R. Pichel, y J. Waliño. 2006. OpenTrad: Traducción automática de código abierto para las lenguas del estado español. *Procesamiento del Lenguaje Natural*, 37:357–358.

Gamallo, P., M. Garcia, y J. R. Pichel. 2013. A method to lexical normalisation of tweets. En *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural. Workshop on Sentiment Analysis at SEPLN*, páginas 81–85.

Malvar, P. y J. R. Pichel. 2010. Obtaining computational resources for languages with scarce resources from closely related computationally-developed languages. the galician and portuguese case. En *Internacional de Lingüística de Corpus (CILC10)*, páginas 529–536.

Malvar, P. y J. R. Pichel. 2011. Métodos semiautomáticos de generación de recursos de opinion mining para el gallego a partir del portugués y el español. *Novática: Revista de la Asociación de Técnicos de Informática*, 214:61–64.

Pichel, J. R., P. Malvar, O. Senra, P. Gamallo, y A. García. 2009. Carvalho: English-galician SMT system from english-portuguese parallel corpus. *Procesamiento del Lenguaje Natural*, 43:379–381.

Pichel, J. R., D. Vázquez, L. Castro, y A. Fernández. 2013. 16 anos desenvolvendo aplicações no campo do processamento da linguagem natural multilingue. *Linguamática*, 5(1):13–20.