aplicabilidad de los resultados a entornos prácticos, como la MO.

- Mejorar el rendimiento de los sistemas de MO mediante la utilización de la estructura sintáctica para extraer la opinión vertida en un texto, con especial atención al tratamiento de las variadas formas de negación, las frases adversativas y la diferenciación entre texto en modo realis (que se refiere eventos o acciones reales) e irrealis (que expresa deseo, potencialidad o condicionalidad).
- Definir modelos de aprendizaje que faciliten la elección de los mejores analizadores, minimizando el coste del proceso de entrenamiento sin perjuicio de la calidad.
- Definir técnicas efectivas que permitan proyectar las herramientas y recursos desarrollados para una lengua, a otra distinta. Ello permitirá, por ejemplo, obtener un analizador sintáctico para un idioma en el que no está disponible un corpus de textos anotados sintácticamente (como es el caso del gallego), a partir de los analizadores obtenidos para otros (como puede ser el español) que sí disponen de tales corpus.
- Definir técnicas efectivas de adaptación de los analizadores a un dominio distinto de aquel para el que fueron concebidos inicialmente, lo que permitirá obtener herramientas para textos no convencionales, como es el caso de los microtextos presentes en los medios web de comunicación social. Ello conlleva también mejorar el rendimiento de los algoritmos de análisis léxico en este contexto, con especial atención al tratamiento de sus peculiaridades léxicas: errores ortográficos, abreviaturas, emoticonos y almohadillas. Todo ello permitirá extraer unidades lingüísticas coherentes que contengan las expresiones de opinión presentes en un enunciado, así como su orientación semántica o polaridad.

4 Resultados alcanzados

Análisis sintáctico: se han realizado desarrollos relevantes en analizadores de dependencias basados en grafos (Gómez Rodríguez, 2016b) y transiciones (Gómez Rodríguez y Fernández-González, 2016). Se ha descrito la relación entre la manera en que funcionan los analizadores basados en transiciones y la forma en que los humanos procesamos el lenguaje (Gómez Rodríguez, 2016a). Se han analizado las dependencias no proyectivas (Ferrer-i-Cancho y Gómez-Rodríguez, 2016a) y se han estudiado las propiedades y distribución estadística de las longitudes de las dependencias (Ferrer-i-Cancho y Gómez-Rodríguez, 2016b; Esteban, Ferrer-i-Cancho, y Gómez-Rodríguez, 2016). Se ha comparado la eficacia de analizadores sintácticos, modelos vectoriales y redes neuronales en tareas de similaridad léxica y analogía (Gamallo, 2017).

Sistemas de MO: se han diseñado e implementado sistemas de minería de opiniones multilingües no supervisados (Vilares, Gómez-Rodríguez, y Alonso, 2017) y supervisados (Vilares, Alonso, y Gómez-Rodríguez, 2017) capaces de proporcionar un análisis de la polaridad de una oración teniendo en cuenta los fenómenos sintácticos que la condicionan (negación, oraciones adversativas, intensificación e irrealis), obteniendo resultados más precisos que los sistemas que se quedan en un nivel léxico. Mediante la aplicación de técnicas de *deep learning* se obtuvo el segundo puesto en las subtareas B y D en la campaña de evaluación SemEval 2016 task 4 (Vilares et al., 2016).

Modelos de aprendizaje: se han diseñado e implementado sendos algoritmos para la predicción del rendimiento en procesos de aprendizaje automático y localización de las instancias para el muestreo (Vilares, Darriba, y Ribadas, 2017).

Recursos lingüísticos: se ha comprobado empíricamente la efectividad de las Universal Dependencies en el procesamiento multilingüe (Vilares, Alonso, y Gómez-Rodríguez, 2016). Se ha creado Galician-TreeGal, un treebank de dependencias universales manualmente revisado para gallego (Garcia, Gómez-Rodríguez, y Alonso, 2016). Se ha creado el corpus EN-ES-CS con tuits en los que se utiliza más de un idioma (Vilares, Alonso, y Gómez-Rodríguez, 2017). Se ha creado el recurso Spanish SentiStrength, cuya eficiencia y utilidad práctica ha sido analizada sobre un conjunto de mensajes de naturaleza política (Vilares, Thelwall, y Alonso, 2015; Vilares y Alonso, 2016).

Normalización de textos: se ha estudiado la robustez de las técnicas basadas en

n-gramas de caracteres para la corrección de palabras en un entorno multilingüe (Vilares et al., 2016a; Vilares et al., 2016b) y se ha experimentado con técnicas de deep learning para la segmentación de palabras (Doval, Gómez-Rodríguez, y Vilares, 2016).

Bibliografía

Carter, S., W. Weerkamp, y M. Tsagkias. 2013. Microblog language identification: overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215.

Doval, Y., C. Gómez-Rodríguez, y J. Vilares. 2016. Segmentación de palabras en español mediante modelos del lenguaje basados en redes neuronales. *Procesamiento del Lenguaje Natural*, 57:75–82.

Esteban, J. L., R. Ferrer-i-Cancho, y C. Gómez-Rodríguez. 2016. The scaling of the minimum sum of edge lengths in uniformly random trees. *Journal of Statistical Mechanics: Theory and Experiment*, (2016):063401.

Ferrer-i-Cancho, R. y C. Gómez-Rodríguez. 2016a. Crossings as a side effect of dependency lengths. *Complexity*, 21(S2):320–328.

Ferrer-i-Cancho, R. y C. Gómez-Rodríguez. 2016b. Liberating language research from dogmas of the 20th century. *Glottometrics*, 33:33–34.

Gamallo, P. Pendiente de publicación. Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. *Language Resources and Evaluation*.

Garcia, M., C. Gómez-Rodríguez, y M. A. Alonso. 2016. Creación de un treebank de dependencias universales mediante recursos existentes para lenguas próximas: el caso del gallego. *Procesamiento del Lenguaje Natural*, 57:33–40.

Gómez Rodríguez, C. 2016a. Natural language processing and the now-or-never bottleneck. *Behavioral and Brain Sciences*, 39:e74.

Gómez Rodríguez, C. 2016b. Restricted non-projectivity: Coverage vs. efficiency. *Computational Linguistics*, 42(4):809–817.

Gómez Rodríguez, C. y D. Fernández-González. 2015. An efficient dynamic oracle for unrestricted non-projective parsing. En *Proceedings of ACL-IJCNLP 2015*, páginas 256–261, Beijing, China.

Vilares, D. y M. A. Alonso. 2016. A review on political analysis and social media. *Procesamiento del Lenguaje Natural*, 56:13–23.

Vilares, D., M. A. Alonso, y C. Gómez-Rodríguez. 2016. One model, two languages: training bilingual parsers with harmonized treebanks. En *Proceedings of ACL 2016*, páginas 425–431, Berlin, Germany.

Vilares, D., M. A. Alonso, y C. Gómez-Rodríguez. 2017. Supervised sentiment analysis in multilingual environments. *Information Processing & Management*, 53(3):595–607.

Vilares, D., Y. Doval, M. A. Alonso, y C. Gómez-Rodríguez. 2016. Exploiting neural activation values for Twitter sentiment classification and quantification. En *Proceedings of SemEval-2016*, páginas 79–84, San Diego, California.

Vilares, D., C. Gómez-Rodríguez, y M. A. Alonso. 2017. Universal, unsupervised (rule-based), uncovered sentiment analysis. *Knowledge-Based Systems*, 118:45–55.

Vilares, D., M. Thelwall, y M. A. Alonso. 2015. The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets. *Journal of Information Science*, 41(6):799–813.

Vilares, J., M. A. Alonso, Y. Doval, y M. Vilares. 2016a. Studying the effect and treatment of misspelled queries in cross-language information retrieval. *Information Processing & Management*, 52(4):646–657.

Vilares, J., M. Vilares, M. A. Alonso, y M. P. Oakes. 2016b. On the feasibility of character n-grams pseudo-translation for cross-language information retrieval tasks. *Computer Speech and Language*, 36(36):136–164.

Vilares, M., V. M. Darriba, y F. J. Ribadas. 2017. Modeling of learning curves with applications to POS tagging. *Computer Speech and Language*, 41:1–28.