

SIFR Project: The Semantic Indexing of French Biomedical Data Resources

**Juan Antonio Lossio-Ventura,
Clement Jonquet**
LIRMM, CNRS, Univ. Montpellier 2
Montpellier, France
fName.lName@lirmm.fr

**Mathieu Roche,
Maguelonne Teisseire**
TETIS, Cirad, Irstea, AgroParisTech
Montpellier, France
fName.lName@teledetection.fr

Abstract

The Semantic Indexing of French Biomedical Data Resources project proposes to investigate the scientific and technical challenges in building ontology-based services to leverage biomedical ontologies and terminologies in indexing, mining and retrieval of French biomedical data.

1 Introducción

Hoy en día la gran cantidad de datos disponibles en línea suele componerse de texto no estructurado, por ejemplo reportes clínicos, informes de reportes adversos, historiales clínicos electrónicos (Lossio-Ventura et al., 2013). Regularmente estos textos son escritos usando un lenguaje específico (expresiones y términos) usados por una comunidad. Es por eso existe la necesidad de formalizar e indexar términos o conceptos técnicos. Lo cual implica un gran consumo de tiempo.

Los términos relevantes son útiles para obtener una mayor comprensión de la estructura conceptual de un dominio. Estos pueden ser: (i) términos de una sola palabra (sencillo a extraer), o (ii) términos de varias palabras (difícil). En el ámbito biomédico, hay una gran diferencia entre los recursos existentes (ontologías) en inglés y francés. En Inglés hay cerca de 7 000 000 de términos asociados a 6 000 000 de conceptos, tales como los de UMLS¹ o BioPortal (Noy et al., 2009). Mientras que, en francés sólo hay alrededor de 330 000 términos asociados a 160 000 conceptos (Neveol et al., 2014). Por lo tanto, hay una necesidad de enriquecer terminologías u ontologías en francés. Por lo tanto, nuestro trabajo se compone de dos pasos principales: (i) la extracción de términos biomédicos, y (ii) el en-

riquecimiento de ontologías, con el fin de poblar ontologías con los términos extraídos.

El artículo es organizado como sigue. Primero discutimos sobre la metodología puesta en marcha para este proyecto en la Sección 2. La evaluación de la precisión es presentada en la Sección 3 seguida de las conclusiones en la Sección 4.

2 Metodología

Nuestro trabajo se divide en dos procesos principales: (i) la extracción de términos biomédicos, y (ii) el enriquecimiento de ontologías, explicados a continuación.

2.1 Extracción Automática de Términos Biomédicos

La extracción de términos es una tarea esencial en la adquisición de conocimiento de un dominio. En este trabajo presentamos las medidas creadas para este objetivo. Medidas que se basan en varios criterios como lingüístico, estadístico, grafos y web para mejorar el resultado de extracción de términos biomédicos. Las medidas presentadas a continuación son puestas a disposición de la comunidad, bajo la aplicación llamada BIO-TEX (Lossio-Ventura et al., 2014).

2.1.1 Lingüística

Estas técnicas intentan recuperar términos gracias a la formación de patrones. La idea principal es la construcción de reglas para describir las estructuras de los términos de un dominio mediante el uso de características ortográficas, léxicas o morfo-sintácticas. La idea principal es la construcción de reglas, normalmente de forma manual, que describen las estructuras comunes de términos para ciertos campos. En muchos casos también, diccionarios conteniendo términos técnicos (e.g., prefijos, sufijos y acrónimos específicos) son usados para ayudar a extraer

¹<http://www.nlm.nih.gov/research/umls>

términos (Krauthammer et al., 2004).

2.1.2 Estadística

Las técnicas estadísticas se basan en la evidencia presentada en el corpus a través de la información contextual. Tales enfoques abordan principalmente el reconocimiento de términos generales (Van Eck et al., 2010). La mayoría de medidas se basan en la frecuencia. La mayor parte de trabajos combinan la información lingüística y estadística, tal es el caso de *C-value* (Frantzi et al., 2000) combina la información estadística y lingüística tanto para la extracción de términos de varias palabras como de términos largos y anidados. Es la medida más conocida en la literatura. En el trabajo de (Zhang et al., 2008), demostraron que *C-value* obtiene los mejores resultados comparado a otras medidas. Además del inglés, *C-value* también ha sido aplicado a otros idiomas tales como japonés, serbio, esloveno, polaco, chino (Ji et al., 2007), español (Barrón-Cedeno et al., 2009), árabe. Es por eso, en nuestro primer trabajo (Lossio-Ventura et al., 2013), la modificamos y adaptamos para el francés.

A partir de *C-value*, hemos creados otras medidas, como *F-TFIDF-C*, *F-OCapi*, *C-OKapi*, *C-TFIDF* (Lossio-Ventura et al., 2014), estas medidas obtienen mejores resultados que *C-value*. Finalmente una nueva medida basada en la información lingüística y estadística es *LIDF-value* (Lossio-Ventura et al., 2014) (patrones Lingüísticos, *IDF*, and *C-value* information), que mejora con gran diferencia los resultados obtenidos por las medidas antes citadas.

2.1.3 Grafos

El modelo de grafos es una alternativa al modelo de información, muestra claramente las relaciones entre los nodos gracias a las aristas. Gracias a los algoritmos de centralidad se puede aprovechar los grupos de información en grafos. Existen aplicaciones de grafos para la Recuperación de Información (RI) en el contexto de las redes sociales, de colaboración y sistemas de recomendación (Noh et al., 2009).

Una medida basada en grafos creada para este proceso es *TeRGraph* (Lossio-Ventura et al., 2014) (Terminology Ranking based on Graph information). Esta medida tiene como objetivo mejorar la precisión de los primeros k términos extraídos después de haber aplicado *LIDF-value*. El grafo es construido con la lista de términos obtenidos

con *LIDF-value*, donde los nodos representan los términos relacionados con otros términos gracias a la co-ocurrencia en el corpus.

2.1.4 Web

Diferentes estudios de Web Mining se enfocan en la similitud semántica, relación semántica. Esto significa para cuantificar el grado en el que algunas palabras están relacionadas, teniendo en cuenta no sólo similitud sino también cualquier posible relación semántica entre ellos. La primera medida web creada fue *WebR* (Lossio-Ventura et al., 2014), finalmente la mejora llamada *WAHI* (Lossio-Ventura et al., 2014) (**W**eb **A**ssociation based on **H**its **I**nformation). Nuestra medida basada en la Web tiene por objetivo volver a clasificar la lista obtenida previamente con *TeR-Graph*. Demostramos con esta medida que la precisión de los k primeros términos extraídos superan los resultados de las medidas arriba mencionadas (ver Sección 3).

2.2 Enriquecimiento de Ontologías

El objetivo de este proceso es enriquecer las terminologías u ontologías con los términos nuevos extraídos en el proceso anterior. Los tres grandes pasos a seguir en este proceso son:

- (1) **Determinar si un término es polisémico:** con la ayuda del Meta-Learning, hemos podido predecir con una confianza de 97% si un término es polisémico. Esta contribución será valorizada en la conferencia ECIR 2015.
- (2) **Identificar los posibles significados si el término es polisémico:** es nuestro trabajo actual, con la ayuda de clustering, clustering sobre los grafos tratamos de resolver este problema.
- (3) **Posicionar el término en una ontología.**

3 Experimentaciones

3.1 Datos, protocolo y validación

En nuestros experimentos, hemos usado el corpus estándar GENIA², el cual es compuesto de 2 000 títulos y resúmenes de artículos de revistas que han sido tomadas de la base de datos Medline, contiene más de 400 000 palabras. GENIA corpus contiene expresiones lingüísticas que se refieren a entidades con interés en biología molecular tales como proteínas, genes y células.

²<http://www.nactem.ac.uk/genia/genia-corpus/term-corpus>

3.2 Resultados

Los resultados son evaluados en términos de *precisión* obtenidos sobre los primeros k términos extraídos ($P@k$) para las medidas propuestas y las medidas base (referencia) para la extracción de términos compuestos de varias palabras. En las subsecciones siguientes, limitamos los resultados para la medida basada en grafos con sólo los primeros 8 000 términos extraídos y los resultados para la medida basada en la web con sólo los primeros 1 000 términos.

3.2.1 Resultados lingüísticos y estadísticos

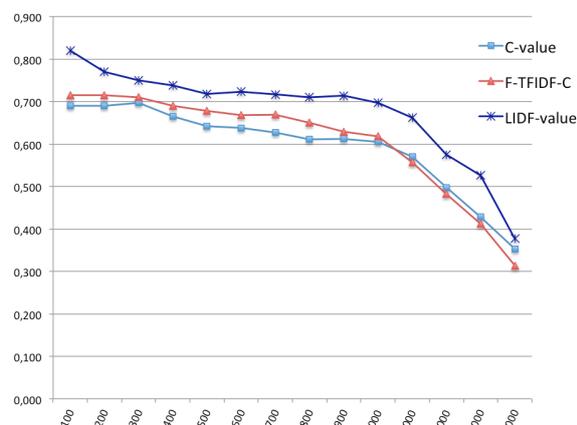


Figure 1: Comparación de la precisión de *LIDF-value* con las mejores medidas de base

3.2.2 Resultados basados en grafos

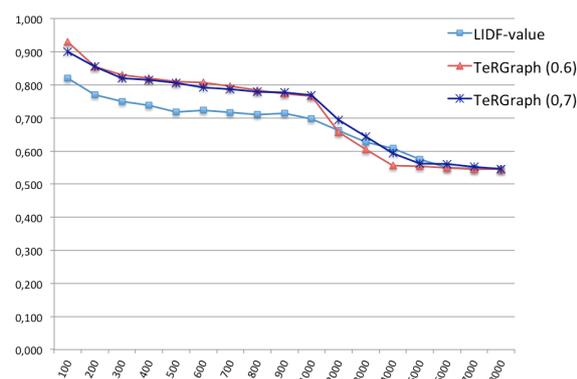


Figure 2: Comparación de la precisión de *TeR-Graph* y *LIDF-value*

3.2.3 Resultados basados en la web

4 Trabajo Futuro

Este artículo presenta la metodología propuesta para el proyecto SIFR. Este proyecto consta de dos

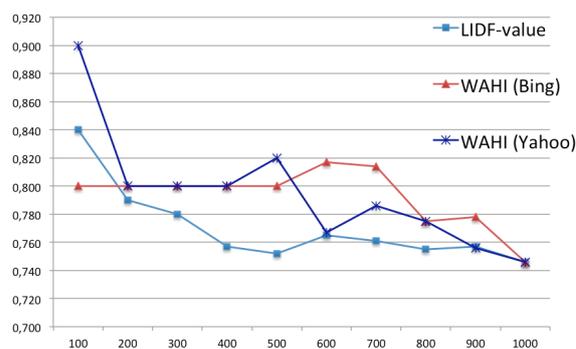


Figure 3: Comparación de la precisión de *WAHI* y *TeRGraph*

grandes procesos.

El primer proceso *Extracción Automática de Términos Biomédicos*, terminado y siendo valorizado en varias publicaciones citadas anteriormente. En este proceso demostramos que las medidas propuestas mejoran la precisión de la extracción automática de términos en comparación a las medidas más populares de extracción de términos.

El segundo proceso *Enriquecimiento de Ontologías*, a la vez dividido en 3 etapas, es nuestra tarea actual, solo la primera etapa ha sido finalizada. En este proceso buscamos encontrar la mejor posición de un término en una ontología.

Como trabajo futuro, pensamos acabar el segundo proceso. Además, planeamos probar estos enfoques generales sobre otros dominios, tales como ecología y agronomía. Finalmente, planeamos aplicar estos enfoques con corpus en español.

Agradecimientos

Este proyecto es apoyado en parte por la Agencia Nacional de Investigación de Francia bajo el programa JCJC, ANR-12-JS02-01001, así como por la Universidad de Montpellier 2, el CNRS y el programa de becas FINCYT, Perú.

References

- Barrón-Cedeno, A., Sierra, G., Drouin, P., Ananiadou, S. 2009. An improved automatic term recognition method for Spanish. *Computational Linguistics, Intelligent Text Processing*, pp. 125-136. Springer.
- Frantzi K., Ananiadou S., Mima, H. 2000. Automatic recognition of multiword terms: the C-value/NC-value Method. *International Journal on Digital Libraries*, (3):115-130.

- Ji, L., Sum, M., Lu, Q., Li, W., Chen, Y. 2007. Chinese Terminology Extraction Using Window-Based Contextual Information. *Proceedings of the 8th International Conference on Computational Linguistics, Intelligent Text Processing (CICLing07)*, pp. 62-74. Springer-Verlag, Mexico City, Mexico.
- Krauthammer, M., Nenadic, G. 2004. Term Identification in the Biomedical Literature. *Journal of Biomedical Informatics*, vol. 37, pp. 512-526. Elsevier Science, San Diego, USA.
- Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire M. 2014. BIOTEX: A system for Biomedical Terminology Extraction, Ranking, and Validation. *Proceedings of the 13th International Semantic Web Conference (ISWC'14)*. Trento, Italy.
- Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire M. 2014. Integration of linguistic and Web information to improve biomedical terminology ranking. *Proceedings of the 18th International Database Engineering and Applications Symposium (IDEAS'14)*, ACM. Porto, Portugal.
- Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire M. 2014. Yet another ranking function to automatic multi-word term extraction. *Proceedings of the 9th International Conference on Natural Language Processing (PolTAL'14)*, Springer LNAI. Warsaw, Poland.
- Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire M. 2014. Biomedical Terminology Extraction: A new combination of Statistical, Web Mining Approaches. *Proceedings of Journées internationales d'Analyse statistique des Données Textuelles (JADT2014)*. Paris, France.
- Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire M. 2013. Combining C-value, Keyword Extraction Methods for Biomedical Terms Extraction. *Proceedings of the Fifth International Symposium on Languages in Biology, Medicine (LBM13)*, pp. 45-49, Tokyo, Japan.
- Neveol, A., Grosjean, J., Darmoni, S., Zweigenbaum, P. 2014. Language Resources for French in the Biomedical Domain. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland
- Noh, TG., Park, SB., Yoon, HG., Lee, SJ., Park, SY. 2009. An Automatic Translation of Tags for Multimedia Contents Using Folksonomy Networks. *Proceedings of the 32nd International ACM SIGIR Conference on Research, Development in Information Retrieval SIGIR '09*, pp. 492-499. Boston, MA, USA, ACM.
- Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M., Chute, C.G., Musen, M. A. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, vol. 37(suppl 2), pp 170-173.
- Van Eck, N.J., Waltman, L., Noyons, E.C.M., Buter, R.K. 2010. Automatic term identification for bibliometric mapping. *Scientometrics*, vol. 82, pp. 581-596.
- Zhang, Z., Iria, J., Brewster, C., Ciravegna, F. 2008. A Comparative Evaluation of Term Recognition Algorithms. *Proceedings of the Sixth International Conference on Language Resources, Evaluation (LREC08)*. Marrakech, Morocco.