

# Reconocimiento de Entidades Biomédicas para el Español mediante la Combinación de Word Embeddings

## *Recognition of Biomedical Entities in Spanish by Combining Word Embeddings*

Pilar López-Úbeda

Sinai Group

Universidad de Jaén

Campus Las Lagunillas s/n. E-23071

plubeda@ujaen.es

**Resumen:** La identificación y extracción de entidades son fundamentales para mejorar el acceso y la integración de la información procedente de datos no estructurados como la literatura biomédica escrita en español. Para la tarea de reconocimiento de entidades nombradas presentamos una red neuronal compuesta por una BiLSTM con una capa secuencial CRF donde se combinan diferentes word embeddings como entrada a la red. Para llevar a cabo esta tarea, estudiamos los diferentes tipos de word embeddings existentes para el español y además generamos unos nuevos relacionados con el dominio biomédico. Finalmente, esta arquitectura es puesta en marcha en competiciones de congresos nacionales e internacionales para saber cómo funciona y preparar futuras mejoras.

**Palabras clave:** Reconocimiento de entidades nombradas, dominio biomédico, word embeddings, deep learning, Procesamiento del Lenguaje Natural

**Abstract:** Identifying and extracting entities is essential to improve access and integration of information from unstructured data such as biomedical literature written in Spanish. For the task of Named Entity Recognition we present a neural network composed of BiLSTM with a CRF sequential layer where different word embeddings are combined as input to the network. To carry out this task, we study different word embeddings and we generate new ones related to the biomedical domain. Finally, this architecture is applied in national and international workshops in order to know how it works and to plan future improvements.

**Keywords:** Named Entities Recognition, biomedical domain, word embeddings, deep learning, Natural Language Processing

### **1 Justificación de la investigación**

El reconocimiento de entidades nombradas (NER, por sus siglas en inglés *Named Entity Recognition*) -a veces denominado extracción o identificación de entidades- es la tarea de identificar y categorizar la información clave (entidades) en el texto. Una entidad puede ser cualquier palabra o serie de palabras que se refiera consistentemente a la misma cosa. La tarea de NER forma parte del Procesamiento del Lenguaje Natural (PLN) que es la rama de la Inteligencia Artificial (IA) que estudia el lenguaje natural.

En el dominio biomédico, la IA permite generar conocimiento y mejorar los servicios

sanitarios a partir del tratamiento de información no estructurada. Esta información supone una parte importante de los datos que se recogen día a día en la práctica clínica. La IA junto con el PLN nos permite aprovechar mejor la información relacionada con la salud y dar respuesta a los nuevos retos de registro, estructuración y exploración de la información.

Con esta tesis se pretende resolver la tarea de NER aplicada a diferentes sub-dominios relacionados con la biomedicina. Estos sub-dominios pueden ser de diferente índole: farmacología, cáncer, genes, efectos secundarios o incluso áreas más novedosas relacionadas

con el COVID-19. Estas entidades pueden ser de gran importancia para el análisis y estudio de notas clínicas puesto que ofrecen palabras clave e información relevante. La tarea de NER es un proceso que consiste en dos pasos:

1. Detectar una entidad nombrada.
2. Categorizar la entidad.

El primer paso consiste en detectar una palabra o secuencia de palabras que forman la entidad, por ejemplo, “cáncer de mama”. El segundo paso requiere la creación de categorías predefinidas para asignar a las entidades. En este caso podemos definir categorías como: enfermedades, procedimientos quirúrgicos, síntomas, medicamentos, etc.

El objetivo principal de este trabajo se centra en el estudio, análisis y desarrollo de diferentes técnicas y herramientas del PLN para la extracción de entidades en el dominio biomédico y aplicado al español. Específicamente se enfoca en el estudio y aplicabilidad de diversas representaciones de palabras.

El principal inconveniente que nos encontramos al querer abordar esta tarea es la carencia de corpus anotados y disponibles en español. Para hacer frente a este problema, nosotros participamos en distintos workshops y competiciones relacionados con NER aplicado al dominio biomédico y en español. Estos workshops además de proporcionar corpus, animan a la comunidad científica a que desarrollen sistemas y poder compararlos entre ellos. Además, ofrecen medidas de evaluación para equiparar sistemas.

## 2 Trabajo relacionado

Para dar una visión general, en esta sección se proporciona una breve descripción de los trabajos relacionados con la tarea de NER existentes hasta el momento.

En el ámbito biomédico, los sistemas de NER identifican entidades a partir de informes clínicos de pacientes. Se han desarrollado varios módulos de NER utilizando sistemas basados en el PLN como por ejemplo, MedLEE (Friedman, 1997), MetaMap (Aronson y Lang, 2010) and cTAKES (Savova et al., 2010)). La mayoría de estos, son sistemas basados en reglas que se apoyan en amplios vocabularios médicos.

Existe una gran cantidad de ontologías en este ámbito y muchas de ellas estrechamente

relacionadas. Las ontologías más comúnmente utilizadas por la comunidad científica de NER son: SNOMED-CT (Stearns et al., 2001; Patrick, Wang, y Budd, 2007; Soriano y Castro, 2017), UMLS (Bodenreider, 2004; Huang et al., 2005; López-Ubeda et al., 2018) y CIE-10 (Névéal et al., 2017; López-Úbeda et al., 2020). Aunque existen otras muchas terminologías y diccionarios más centrados en un sub-dominio concreto, como por ejemplo DrugBank (Law et al., 2014), ChEBI (Hastings et al., 2012) o PharmGKB (Hernandez-Boussard et al., 2007), relacionados con el ámbito de la farmacología, sustancias o medicamentos.

Se han desarrollado varios sistemas de NER para el dominio biomédico (Campos, Matos, y Oliveira, 2012), utilizando diferentes enfoques y técnicas que generalmente pueden clasificarse como:

- Basado en reglas.
- Basado en diccionarios.
- Basado en aprendizaje automático o *Machine Learning* (ML).

Cada enfoque anterior cumple diferentes requisitos, dependiendo de las características lingüísticas de las entidades que se quiera identificar. Por un lado, los enfoques basados en reglas consisten en (normalmente de forma manual) desarrollar reglas que describen estructuras de nomenclatura comunes para ciertas clases de término, buscando indicios ortográficos o léxicos (Lowe y Sayle, 2015; López-Ubeda et al., 2019). Por otro lado, los métodos basados en diccionarios utilizan recursos terminológicos existentes para localizar ocurrencias de término dentro del texto (López-Úbeda et al., 2020; López-Ubeda et al., 2018; López-Úbeda et al., 2018). Finalmente, las técnicas basadas en ML aprenden y clasifican de forma automática, mejorando con el paso del tiempo (Úbeda et al., 2019; Gorinski et al., 2019; Yadav y Bethard, 2019).

Los estudios más actuales con respecto al NER en el dominio médico se han centrado en los métodos basados en ML. Los enfoques basados en ML más tradicionales como Conditional Random Fields (CRF) (Lafferty, McCallum, y Pereira, 2001) y Support Vector Machines (SVM) (Vapnik, 2013) han sido un punto clave para comenzar a identificar entidades de forma automatizada. Seguidamente, los algoritmos de aprendizaje profundo o

*deep learning* han ido teniendo gran importancia en esta tarea también. Concretamente, el modelo Bidirectional Long Short-Term Memory con una capa CRF (BiLSTM-CRF) muestra resultados prometedores en cuanto a la extracción de conceptos (Luo et al., 2018; Huang, Xu, y Yu, 2015; Lample et al., 2016).

Las redes neuronales toman como entrada un vector para representar cada palabra. Estos vectores, también conocidos como *word embeddings* (WE), ofrecen representaciones similares para palabras con significados semejantes. Concretamente, los WE son una técnica en la que cada palabra es representada como vector numérico en un espacio vectorial predefinido (Kusner et al., 2015).

### 3 Descripción de la investigación propuesta

Esta tesis se encuentra en el proceso de implementación, experimentación y evaluación de los sistemas propuestos. El presente año ha estado marcado por diferentes hitos y tareas propuestas por talleres y competiciones existentes en congresos tanto nacionales como internacionales en el área de NER y dominio biomédico.

Para llevar a cabo la implementación de los sistemas, nosotros proponemos una metodología basada en *deep learning* y redes neuronales para la identificación de conceptos médicos. Específicamente, usaremos el modelo propuesto por (Huang et al., 2005) que consiste en una red neuronal BiLSTM con una capa final de CRF. Esta arquitectura es mostrada en la Figura 1 y descrita en detalle a continuación:

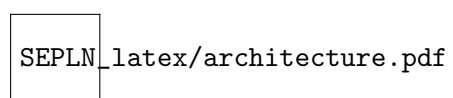


Figura 1: Enfoque propuesto basado en una red neural BiLSTM-CRF que utiliza una combinación de *word embeddings* como capa de entrada.

- **Combinando WE:** La entrada a la red será un vector formado por la concatenación de varios vectores. Hasta el momento se han probado varios tipos de WE:
  - WE clásicos: son representaciones estáticas a nivel de palabra, lo que significa que cada palabra distinta

recibe exactamente un vector precalculado. Estos embeddings con frecuencia son entrenados usando el algoritmo fastText<sup>1</sup> y GloVe (Pennington, Socher, y Manning, 2014).

- WE contextuales: son unos novedosos WE que capturan información semántica y sintáctica que va más allá de los embeddings clásicos. Se basan en un modelo de lenguaje a nivel de carácter y su uso es particularmente beneficioso para la tarea de NER donde se aborda un problema de etiquetado secuencial (Akbik, Blythe, y Vollgraf, 2018; Akbik, Bergmann, y Vollgraf, 2019).
- WE pre-entrenados basados en *Transformers*: utilizan un mecanismo de atención usado en enfoques de *transfer learning*. Este tipo de embeddings son comúnmente pre-entrenados sobre corpus de gran tamaño para capturar similitudes sintácticas y semánticas entre las palabras. Para nuestros experimentos, nosotros utilizaremos aquellos WE disponibles para español como BETO (Devlin et al., 2018) o XLMRoBERTa (Conneau et al., 2019).
- WE específicos para medicina: dado que los demás embeddings suelen estar entrenados sobre dominios genéricos, nosotros proponemos generar unos WE entrenados sobre un corpus de medicina y en español. Primero nosotros utilizamos MeSpEN (Villegas et al., 2018) y seguidamente, añadimos información extra de Webs como Mayo Clinic<sup>2</sup>, World Health Organization<sup>3</sup> y WebMD<sup>4</sup> usando el algoritmo FastText.

- **BiLSTM:** La combinación de diferentes WE son la entrada a una capa BiLSTM. La red BiLSTM se compone por dos LSTMs. Una LSTM hacia adelante que calcula una representación de la secuencia de izquierda a derecha en cada palabra, y otra LSTM hacia atrás que calcula

<sup>1</sup><https://fasttext.cc>

<sup>2</sup><https://www.mayoclinic.org/es-es>

<sup>3</sup><https://www.who.int/es>

<sup>4</sup><https://www.webmd.com/>

una representación de la misma secuencia en sentido contrario.

- **CRF:** La salida de la red BiLSTM es introducida en una capa CRF utilizada para predecir las etiquetas para la palabra. Estas etiquetas atienden a las posibles clases que se encuentran etiquetadas en el corpus.

Como podemos ver en la Figura 1, las anotaciones para cada palabra se codificaron utilizando el esquema de etiquetado BIO. Así pues, cada palabra de una frase fue etiquetada con B (inicio de una entidad), I (dentro de la entidad) y O (no entidad). Este esquema es el más popular en la tarea de NER.

#### 4 Experimentos propuestos

Respecto a la experimentación y evaluación del sistema nosotros usaremos el modelo anteriormente propuesto en diferentes competiciones y congresos para ver su funcionamiento. Algunos de los resultados obtenidos se muestran a continuación.

##### 4.1 EMNLP - PharmaCoNER 2019

PharmaCoNER (Gonzalez-Agirre et al., 2019) fue una de las tareas propuestas en el congreso EMNLP. Esta tarea consistía en el reconocimiento de medicamentos y entidades químicas en un corpus de casos clínicos en español. La tarea de NER estaba compuesta por varias sub-tareas consecutivas: (1) detectar la entidad y (2) etiquetar la entidad según unas categorías preferidas.

Tabla 1: Resultados en PharmaCoNER usando el enfoque BiLSTM-CRF. WE: word embeddings.

	P (%)	R (%)	F1 (%)
WE clásicos	86,53	83,46	84,96
WE contextuales	88,74	85,22	86,95
WE médicos	88,94	86,24	87,08
Combinando WE	91,41	90,14	90,77

La Tabla 1 muestra los resultados obtenidos en la tarea de detección y clasificación de una entidad usando la métrica *micro-averaged* para *precision* (P), *recall* (R) y F1.

Como podemos ver, primero llevamos a cabo un experimento usando algunos de los

WE de forma individual: WE clásicos, contextuales y los WE generados por nosotros mismos (WE médicos). Como podemos ver, el uso de cada uno de ellos mejora el resultado anterior. Posteriormente, proponemos una combinación de WE para representar las palabras del corpus. Nuestro mejor sistema propone la combinación de los tres tipos de incrustaciones que hemos visto por separado y juntos alcanzan el 91,41 % de *precision*, 90,14 % de *recall* y 90,77 % de F1. Este resultado es comparado con el anteriormente obtenido y propuesto en PharmaCoNER por nuestro grupo de investigación (López Úbeda et al., 2019) donde se utilizó un sistema basado en CRF obteniendo un 88,51 % de *precision*, 69,81 de *recall* y 78,06 % de F1. Combinando las diferentes representaciones de palabras conseguimos mejorar en un 12,71 % de F1.

##### 4.2 IberLEF - eHealth-KD 2020

eHealth-KD (Piad-Morffis et al., 2020) es otra de las competiciones usada para aplicar nuestro enfoque de BiLSTM-CRF usando una combinación de embeddings. En esta tarea, dada una lista de documentos relacionados con la medicina y escritos en español, el objetivo de la sub-tarea de NER es identificar todas las entidades y sus categorías.

La Tabla 1 muestra los resultados obtenidos en la tarea de NER en eHealth-KD usando las métricas *precision* (P), *recall* (R) y F1. Como podemos apreciar en la tabla, el uso de una combinación de varios WE mejora siempre el anterior resultado. El mejor resultado fue obtenido concatenando cuatro tipos de word embeddings (WE médicos, contextuales, BETO y XLMRoBERTa). Usando una combinación de todos ellos, obtuvimos un 84,46 % en *precision*, un 80,67 % de *recall* y 82,52 % de F1. Cabe destacar que nuestra metodología obtuvo el mejor resultado de entre todos los participantes en la tarea de NER en eHealth-KD.

Tras analizar los resultados, podemos concluir que la incorporación de cada tipo de word embeddings añade la información necesaria para que la red neuronal BiLSTM-CRF asigne la etiqueta correcta a cada palabra. Dado que cada WE utilizado se genera con un método e información diferente, cada palabra estará representada por aspectos de conocimiento basados en el corpus de formación y su combinación mejora potencialmente

Tabla 2: Resultados en eHealth-KD usando el enfoque BiLSTM-CRF. WE: word embeddings.

	P (%)	R (%)	F1 (%)
WE médicos + contextuales	83,03	80,94	81,97
WE médicos + contextuales + BETO	84,32	79,32	81,74
WE médicos + contextuales + BETO + XLMRoBERTa	84,46	80,67	82,52

el rendimiento final.

## 5 Elementos de investigación para discusión

La extracción e identificación de entidades biomédicas es un tema de interés en el PLN, nuestra intención en este trabajo es discutir las siguientes aspectos para seguir profundizando en el estudio:

- ¿Es necesario la creación de recursos para el NER médicas en español? ¿Cómo contribuyen a la consecución del problema?
- ¿Es necesario generar recursos lingüísticos para cada tarea que se desea abordar?
- ¿Qué otros recursos ayudan a la consecución del NER dentro del campo de la medicina?
- ¿Qué sistemas y algoritmos son los principalmente usados en el NER en el dominio biomédico?
- ¿Es posible otro tipo de representación de palabras en las redes neuronales? ¿Cómo mejoraría su rendimiento?

## Agradecimientos

Este trabajo está parcialmente subvencionado por el Fondo Europeo de Desarrollo Regional (FEDER), el proyecto LIVING-LANG (RTI2018-094653-B-C21) del Gobierno de España.

## Bibliografía

- Akbik, A., T. Bergmann, y R. Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, páginas 724–728.
- Akbik, A., D. Blythe, y R. Vollgraf. 2018. Contextual string embeddings for sequence labeling. En *Proceedings of the 27th International Conference on Computational Linguistics*, páginas 1638–1649.
- Aronson, A. R. y F.-M. Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.
- Bodenreider, O. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl\_1):D267–D270.
- Campos, D., S. Matos, y J. L. Oliveira. 2012. Biomedical named entity recognition: a survey of machine-learning tools. *Theory and Applications for Advanced Text Mining*, páginas 175–195.
- Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, y V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Friedman, C. 1997. Towards a comprehensive medical language processing system: methods and issues. En *Proceedings of the AMIA annual fall symposium*, página 595. American Medical Informatics Association.
- Gonzalez-Agirre, A., M. Marimon, A. Intxaurre, O. Rabal, M. Villegas, y M. Krallinger. 2019. PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. En *Proceedings of The 5th Workshop on*

- BioNLP Open Shared Tasks*, páginas 1–10, Hong Kong, China, Noviembre. Association for Computational Linguistics.
- Gorinski, P. J., H. Wu, C. Grover, R. Tobin, C. Talbot, H. Whalley, C. Sudlow, W. Whiteley, y B. Alex. 2019. Named entity recognition for electronic health records: A comparison of rule-based and machine learning approaches. *arXiv preprint arXiv:1903.03985*.
- Hastings, J., P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, y others. 2012. The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research*, 41(D1):D456–D463.
- Hernandez-Boussard, T., M. Whirl-Carrillo, J. M. Hebert, L. Gong, R. Owen, M. Gong, W. Gor, F. Liu, C. Truong, R. Whalley, y others. 2007. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic acids research*, 36(suppl.1):D913–D918.
- Huang, Y., H. J. Lowe, D. Klein, y R. J. Cucina. 2005. Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the umls specialist lexicon. *Journal of the American Medical Informatics Association*, 12(3):275–285.
- Huang, Z., W. Xu, y K. Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Kusner, M., Y. Sun, N. Kolkin, y K. Weinberger. 2015. From word embeddings to document distances. En *International conference on machine learning*, páginas 957–966.
- Lafferty, J., A. McCallum, y F. C. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
- Lample, G., M. Ballesteros, S. Subramanian, K. Kawakami, y C. Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.
- Law, V., C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, y others. 2014. Drugbank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(D1):D1091–D1097.
- López-Ubeda, P., M. C. Díaz-Galiano, M. T. Martín-Valdivia, y L. A. Urena-López. 2018. Sinai en tass 2018 task 3. clasificando acciones y conceptos con umls en medline. *Proceedings of TASS*, 2172.
- López-Ubeda, P., M. C. Díaz-Galiano, L. A. U. López, y M. T. M. Valdivia. 2019. Anonymization of clinical reports in spanish: a hybrid method based on machine learning and rules. En *IberLEF@ SEPLN*, páginas 687–695.
- López-Úbeda, P., M. C. Díaz-Galiano, M. T. Martín-Valdivia, y S. M. J. Zafra. 2018. Sinai at diann-ibereval 2018. annotating disabilities in multi-language systems with umls. En *IberEval@ SEPLN*, páginas 37–43.
- López-Úbeda, P., M. C. Díaz-Galiano, A. Montejo-Ráez, M.-T. Martín-Valdivia, y L. A. Ureña-López. 2020. An integrated approach to biomedical term identification systems. *Applied Sciences*, 10(5):1726.
- López Úbeda, P., M. C. Díaz Galiano, L. A. Urena Lopez, y M. Martin. 2019. Using snomed to recognize and index chemical and drug mentions. En *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, páginas 115–120, Hong Kong, China, Noviembre. Association for Computational Linguistics.
- Lowe, D. M. y R. A. Sayle. 2015. Leadmine: a grammar and dictionary driven approach to entity recognition. *Journal of cheminformatics*, 7(1):1–9.
- Luo, L., Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, y J. Wang. 2018. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381–1388.
- Névéol, A., R. N. Anderson, K. B. Cohen, C. Grouin, T. Lavergne, G. Rey, A. Robert, C. Rondet, y P. Zweigenbaum. 2017. Clef ehealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in english and french. En *CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*, página 17.

- Patrick, J., Y. Wang, y P. Budd. 2007. An automated system for conversion of clinical notes into snomed clinical terminology. En *Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68*, páginas 219–226. Australian Computer Society, Inc.
- Pennington, J., R. Socher, y C. D. Manning. 2014. Glove: Global vectors for word representation. En *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, páginas 1532–1543.
- Piad-Morffis, A., Y. Gutiérrez, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz, y A. Montoyo. 2020. Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2020. En *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*.
- Savova, G. K., J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, y C. G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.
- Soriano, I. M. y J. Castro. 2017. Dner clinical (named entity recognition) from free clinical text to snomed-ct concept. *WSEAS Transactions on Computers*, 16:83–91.
- Stearns, M. Q., C. Price, K. A. Spackman, y A. Y. Wang. 2001. Snomed clinical terms: overview of the development process and project status. En *Proceedings of the AMIA Symposium*, página 662. American Medical Informatics Association.
- Úbeda, P. L., M. C. D. Galiano, M. T. Martín-Valdivia, y L. A. U. Lopez. 2019. Using machine learning and deep learning methods to find mentions of adverse drug reactions in social media. En *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, páginas 102–106.
- Vapnik, V. 2013. *The nature of statistical learning theory*. Springer science & business media.
- Villegas, M., A. Intxaurreondo, A. Gonzalez-Agirre, M. Marimon, y M. Krallinger. 2018. The MeSpEN resource for English-Spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. *LREC MultilingualBIO: Multilingual Biomedical Text Processing (Malero M, Krallinger M, Gonzalez-Agirre A, eds.)*.
- Yadav, V. y S. Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.