Actas de las 7ma Jornadas de Intercambio y Reflexión acerca de la Investigación en Bibliotecología
17, 18 y 19 de abril de 2024. Departamento de Bibliotecología, FaHCE-UNLP. ISSN 1853-5631

Relevamiento de tecnologías de Inteligencia Artificial para el Procesamiento de Lenguaje Natural (PLN)Reconocimiento de Entidades Nombradas (REN) para la identificación de entidades geográficas en artículos científicos con fines de georeferenciación.

Gustavo G. Archuby
Facultad de Humanidades y Ciencias de la Educación Universidad Nacional de La Plata
La Plata, Buenos Aires, Argentina
gustavoa@fahce.unlp.edu.ar

Resumen: Teniendo como base el proyecto de investigación "La representación del territorio en el discurso científico. Aproximación bibliométrica a los estudios de ciencia local desde la delimitación geográfica de los objetos de estudio", el propósito de este trabajo es llevar a cabo un análisis de las diversas herramientas de Inteligencia Artificial (IA) disponibles en el ámbito del software libre, centrándose específicamente en aquellas dedicadas al procesamiento del lenguaje natural (PLN)(«Procesamiento de lenguajes naturales», 2024) y, más particularmente, en el reconocimiento de entidades nombradas (NER)(«Reconocimiento de entidades nombradas», 2023). La evaluación se extiende a la curva de aprendizaje y los requisitos de recursos para su entrenamiento.

Simultáneamente, se busca identificar herramientas utilitarias que faciliten el proceso de entrenamiento, como aquellas que simplifiquen el marcado de entidades. Además, se exploran herramientas que permitan agregar funcionalidades al resultado obtenido, como la ubicación de entidades geográficas reconocidas automáticamente.

Palabras claves: inteligencia artificial, software libre, lenguaje natural, entidades, georeferenciación

Introducción

La presente investigación se centra en una aproximación heurística para la evaluación y comparación de herramientas de procesamiento del lenguaje natural (NLP-NER). A diferencia de la aplicación de métricas rigurosas utilizadas en la evaluación de software, este tipo de enfoque se fundamenta en el estudio de trabajos previos, en la experimentación utilizando algunos casos de prueba al que se suma en gran medida la experiencia y conocimientos del autor en el campo. Este enfoque busca proporcionar, desde una perspectiva más holística y práctica, la eficacia y el rendimiento de las herramientas.

En este trabajo se realiza un análisis de una selección de herramientas de IA disponibles en el ámbito del software libre, centrándose específicamente en aquellas dedicadas al PLN y, más particularmente, en el REN. La evaluación se extiende a la curva de aprendizaje y los requisitos de recursos necesarios para su entrenamiento(«Aprendizaje automático», 2024). Simultáneamente, se busca identificar herramientas utilitarias que faciliten el proceso de entrenamiento, como aquellas que simplifiquen el marcado de entidades. Además, se exploran herramientas que permitan agregar funcionalidades al resultado obtenido, como la ubicación de entidades geográficas reconocidas automáticamente.

El proceso de entrenamiento de estos modelos implica la definición de nuevas entidades, cuando sea necesario, el entrenamiento propiamente dicho con artículos de revistas científicas y la evaluación del mismo una vez finalizado el entrenamiento(«Aprendizaje automático», 2024). La información extraída puede utilizarse, por ejemplo, para la georeferenciación de determinadas entidades encontradas, lo que permitiría la identificación automatizada de ubicaciones geográficas nombradas en los textos.

El entrenamiento de estos modelos para el PLN en artículos científicos se presenta como una contribución en el ámbito de la bibliometría y la cienciometría. Este enfoque permite, por ejemplo, distinguir entre la ubicación geográfica donde se lleva a cabo el trabajo y los lugares nombrados por filiación de los autores de forma automática.

En el contexto del proyecto “La representación del territorio en el discurso científico. Aproximación bibliométrica a los estudios de ciencia local desde la delimitación geográfica de los objetos de estudio” que tiene por objetivo:

> ”...contribuir al conocimiento de los modos y usos de la representación de lo geográfico en el discurso científico según prácticas disciplinares, contextos de producción y circuitos de circulación de las publicaciones científicas, que favorezcan a una mayor comprensión del alcance conceptual de "lo local" en los temas y problemas de investigación…”

La necesidad de encontrar entidades que representen lugares geográficos dentro de artículos científicos de manera automática para luego utilizar esta información y así realizar

diferentes análisis bibliométricos, es la motivación de esta ponencia, la posibilidad de poder reconocer entidades y sus tipos automáticamente (instituciones, regiones geográficas, personas, países, organizaciones, etc.), permite abrir un abanico de posibilidades para el procesamiento de documentos diversos y la posibilidad de dar acceso a los mismos. Además si se tiene en cuenta que estas herramientas pueden ser entrenadas en contextos específicos, podemos pensar, que en entornos especializados, pueda utilizarse para el reconocimiento de entidades propias del ámbito, como ya se ha hecho en medicina , biología, análisis financiero, legal, redes sociales, etc.

Herramientas de PLN

El PLN ha experimentado un notable auge en los últimos 25 años, impulsado por innovadores métodos y tecnologías emergentes. Entre ellas, destaca el REN, que involucra la identificación de elementos específicos dentro de un texto, como nombres de personas, empresas u organizaciones(Goldberg, 2016).

La evolución del PLN comenzó con aproximaciones basadas en reglas lingüísticas y análisis morfosintácticos, posteriormente surgieron modelos estadísticos como Conditional Random Fields (CRF) y Maximum Entropy2. Actualmente, las redes neuronales y el aprendizaje profundo constituyen pilares centrales en el desarrollo de técnicas de PLN.

Metodología

Como dijimos anteriormente, este trabajo se centra en una aproximación heurística para la evaluación y comparación de herramientas de NLP-NER. A diferencia de la aplicación de métricas rigurosas utilizadas en la evaluación de software(Jalal et al., 2020), este tipo de enfoque se fundamenta en el estudio de trabajos previos, en la experimentación utilizando algunos casos de prueba al que se suma en gran medida la experiencia y conocimientos del autor en el campo.

Herramientas propiamente dichas

De acuerdo a la información relevada se decidió seleccionar las siguientes herramientas de software libre para el Reconocimiento de Entidades Nombradas:

- GATE (General Architecture for Text Engineering)(*GATE.ac.uk - index.html*, s. f.): es un entorno de desarrollo y un framework que permite la creación, anotación o etiquetado de corpus y evaluación de aplicaciones generadas conectándose a una base de datos de textos. GATE es una herramienta de software libre y cuenta con una amplia comunidad de usuarios y una documentación clara.
- NLTK (*NLTK :: Natural Language Toolkit*, s. f.): es una biblioteca de Python que ofrece herramientas para el procesamiento del lenguaje natural, incluyendo el reconocimiento de entidades nombradas. NLTK es una herramienta de software libre y cuenta con una amplia comunidad de usuarios y una documentación clara.
- Stanford coreNLP(Software - The Stanford Natural Language Processing Group, s. f.): es una herramienta de procesamiento de lenguaje natural de código abierto que permite a los usuarios obtener anotaciones lingüísticas para el texto, incluyendo límites de token y oración, partes del discurso, entidades nombradas, valores numéricos y de tiempo,

análisis de dependencia y de constituyentes, correferencia, atribuciones de citas y relaciones.

- OpenNLP(*Apache OpenNLP*, s. f.): es una biblioteca de Java para el procesamiento del lenguaje natural que incluye herramientas para el reconocimiento de entidades nombradas. OpenNLP es una herramienta de software libre y cuenta con una amplia comunidad de usuarios y una documentación clara.
- SpaCy(Hidalgo, 2024): es una biblioteca de Python para el procesamiento del lenguaje natural que incluye herramientas para el reconocimiento de entidades nombradas. SpaCy es una herramienta de software libre y cuenta con una amplia comunidad de usuarios y una documentación clara.
- SparkNLP(*Spark NLP - Getting Started*, 2021): es una biblioteca de código abierto basada en Apache Spark, diseñada para el procesamiento del lenguaje natural en Python, Java y Scala. Ofrece capacidades avanzadas para el análisis de texto y la creación de pipelines de procesamiento de lenguaje natural, permitiendo escalar y procesar grandes volúmenes de datos de manera distribuida.

Parámetros de análisis

A continuación se detallan los parámetros que se usarán para evaluar las herramientas seleccionadas.

- Comunidad de desarrollo: la comunidad de desarrollo, en el ámbito del software libre, es fundamental ya que de ella dependen la corrección de errores, la actualización a nuevas versiones y la continuidad del proyecto en el tiempo, dicha comunidad incluye desarrolladores, colaboradores y usuarios.
- Precisión y cobertura: la precisión se refiere a la exactitud con la que la herramienta identifica las entidades nombradas en el texto, mientras que la cobertura se refiere a la cantidad de entidades que la herramienta es capaz de identificar. Es importante encontrar un equilibrio entre precisión y cobertura dependiendo de las necesidades del proyecto.
- Adaptabilidad y personalización: la capacidad de la herramienta para ser adaptada y personalizada para casos de uso específicos o dominios de aplicación. Esto puede incluir la capacidad de entrenar modelos personalizados con datos etiquetados propios.
- Definición de nuevas entidades: es la capacidad de la herramienta para definir nuevas entidades, propias de un ámbito determinado, como por ejemplo la medicina o la ingeniería o, como en este caso, artículos científicos.
- Interpretación de contextos complejos: evaluar la capacidad de la herramienta para comprender y manejar contextos complejos en los textos, como ambigüedades o referencias cruzadas, lo cual es crucial para un NER preciso y efectivo en situaciones donde el contexto es fundamental.
- Disponibilidad de recursos y modelos pre-entrenados: la disponibilidad de modelos pre-entrenados y recursos lingüísticos que pueden acelerar el desarrollo y la implementación de aplicaciones de NER.

- Soporte multilingüe: la capacidad de la herramienta para reconocer entidades en múltiples idiomas, lo cual es crucial para aplicaciones que operan en entornos multilingüe.
- Simplicidad de aprendizaje: una herramienta fácil de aprender puede facilitar la implementación y el uso en proyectos a gran escala, ya que reduce el tiempo necesario para entrenar al personal y poner en marcha el sistema.
- Actualizaciones y mantenimiento: la frecuencia y la calidad de las actualizaciones de la herramienta, así como el soporte continuo y el mantenimiento por parte de la comunidad de desarrollo. Este parámetro, en el ámbito del software libre, está íntimamente relacionado con el parámetro "Comunidad de desarrollo".
- Facilidad de integración: la facilidad con la que la herramienta puede integrarse en sistemas existentes y trabajar con otras tecnologías y herramientas de procesamiento de lenguaje natural.
- Escalabilidad y rendimiento: la capacidad de la herramienta para manejar grandes volúmenes de texto de manera eficiente y mantener un buen rendimiento a medida que aumenta la escala de los datos de entrada.
- Uso de recursos computacionales: la eficiencia en el uso de recursos computacionales, como memoria y procesamiento, especialmente importante para aplicaciones que deben ejecutarse en entornos con recursos limitados.
- Facilidad de despliegue: la facilidad con la que la herramienta puede ser implementada y desplegada en diferentes entornos, incluyendo servidores locales o en la nube.
- Compatibilidad con diferentes formatos de entrada y salida: la capacidad de la herramienta para manejar diferentes formatos de texto de entrada y proporcionar resultados en formatos compatibles con otras herramientas y sistemas.
- Transparencia y explicabilidad: la capacidad de la herramienta para proporcionar explicaciones claras y transparentes sobre cómo se realizan las predicciones y cómo se identifican las entidades, lo que es importante para la confianza del usuario y la interpretación de los resultados.
- Lenguaje de programación: el lenguaje de programación en el que está desarrollada la herramienta.

Resultados

Los valores en la grilla representan un ranking en función de la información recopilada de cada una de las herramientas seleccionadas, es decir, cuales, según la información relevada, están mejor posicionada en cada uno de los ítems. Salvo el caso de los lenguajes de programación donde se registró en que lenguaje de programación está desarrollada la herramienta.

Tabla 1: Parámetros de evaluación vs Softwares evaluados

Softwares/ Parámetros	GATE	NLTK	Stanford coreNLP	OpenNLP	SpaCy	SparkNLP
Comunidad de desarrollo	5	1	3	4	1	6
Precisión y cobertura	2	2	1	3	1	3
Adaptabilidad y personalización	2	2	1	3	1	2
Definición de nuevas entidades	3	3	2	3	1	3
Interpretación de contextos complejos	1	2	1	4	1	3
Disponibilidad de recursos y modelos pre-entrenados	2	2	2	3	1	2
Soporte multilingüe	1	1	1	1	1	1
Simplicidad de aprendizaje	3	2	3	3	1	4
Actualizaciones y mantenimiento	2	2	2	3	1	2
Facilidad de integración	1	2	1	1	1	2
Escalabilidad y rendimiento	2	5	4	4	3	1
Uso de recursos computacionales	1	3	3	3	2	1

Facilidad de despliegue	3	3	2	3	1	3
Lenguaje de programación	Java, Python y otros	Python	Java	Java	Python	Scala
Compatibilidad con diferentes formatos de entrada y salida	1	1	1	1	1	1
Transparencia y explicabilidad	1	2	1	1	1	2

Conclusión

Elección

En función de lo estudiado la elección fue SpaCy, es la herramienta que más se destaca, por su comunidad de desarrollo, por su facilidad de aprendizaje, por la disponibilidad de modelos pre entrenados, de herramientas para definición de entidades o para entrenamiento, está desarrollada en Python que es un lenguaje más simple de utilizar, por lo menos a mi criterio.

Pruebas

Se instaló SpaCy en una máquina con sistema operativo GNU-Linux distribución Mint se realizaron pruebas básicas sobre análisis gramatical, para chequear su funcionamiento, para lo que fue necesario utilizar modelos pre-entrenados, tanto en inglés como en español.(Pinto et al., 2016)

Luego se realizaron pruebas para reconocer entidades dentro de un texto, hubo errores pero en general se comportó bien, es decir en la mayoría de los casos encontró las entidades, y sí, hubo más errores en la distinción del tipo de entidad, de todas maneras esto puede mejorarse re-entrenando el modelo mediante el marcado de entidades de forma correcta.

Luego se realizaron pruebas con artículos científicos, aquí al darle textos para analizar de un tipo particular (artículos de revistas científicas) hubo entidades que no están definidas y por tanto no fueron reconocidas, por ejemplo, las citas bibliográficas, que, además, son entidades complejas ya que están constituídas de otras entidades, como autores, títulos, fechas, etc.

Para que la herramienta pueda reconocer nuevas entidades hay que definirlas, seleccionar un corpus de documentos y marcar en estos las nuevas entidades, si bien este proceso puede ser manual es muy engorroso es por esto que se analizaron dos herramientas para el marcado de entidades en textos para spacy:

- NER Text Annotator (https://tecoholic.github.io/ner-annotator/) (Ner-Annotator/Docs at Main · Tecoholic/Ner-Annotator, s. f.)y
- Spacy-annotator (https://github.com/ieriii/spacy-annotator) (Alemani, 2020/2024)

Estas dos herramientas son de código abierto y software libre, lo que permite su descarga e instalación sin inconvenientes

Por último la georeferenciación, en este punto decidimos utilizar la API (del inglés Application Program Interface) de google primero para geodecodificar, es decir para convertir esas entidades en ubicaciones, en este punto es importante tratar el tema de la ambigüedad(Singh, 2017), ya que puede haber más una entidad geográfica con el mismo nombre, para resolver este problema se puede:

- **Contexto Adicional** proporcionar un contexto adicional cuando es posible (es decir proporcionar información, por ejemplo en que continente, región, país o provincia se encuentra.
- **Geocodificación Diferenciada:** en determinados casos es posible una geocodificación diferenciada para cada una de las entidades geográficas con el mismo nombre. Esto implica utilizar información adicional, como códigos postales o descripciones detalladas, para asignar las coordenadas correctas a cada lugar.
- **Validación Manual:** Ante la ambigüedad de la geocodificación, puede ser necesario realizar una validación manual para confirmar las coordenadas asignadas a cada entidad geográfica. Esto garantiza la precisión y evita confusiones al visualizar los lugares en un mapa.

A modo de cierre

Estas herramientas pueden tener fines diversos en el ámbito de la bibliotecología y en particular de la bibliometría, por dar un ejemplo, en los estudios de impacto se podrían utilizar NER para

- identificar entidades clave dentro de los artículos citados, como autores, instituciones, términos científicos específicos, etc. Esto permitirá una comprensión más detallada de las conexiones y relaciones entre las publicaciones.
- análisis de redes de citación para identificar entidades en los citantes, como autores, instituciones o temas específicos. Esto facilitará la creación de redes de citación más precisas y detalladas, lo que ayudará a visualizar y comprender mejor las interacciones entre las publicaciones

- evaluación del impacto por entidades se podría utilizar NER para cuantificar el impacto de entidades específicas en las publicaciones periódicas. Esto permitirá identificar qué entidades tienen un mayor impacto en la comunidad científica.

Bibliografía y recursos utilizados

Alemani, E. (2024). *Ieriii/spacy-annotator* [Python]. https://github.com/ieriii/spacy-annotator (Obra original publicada en 2020)

Apache OpenNLP. (s. f.). Recuperado 1 de marzo de 2024, de https://opennlp.apache.org/

Aprendizaje automático. (2024). En *Wikipedia, la enciclopedia libre*. https://es.wikipedia.org/w/index.php?title=Aprendizaje_autom%C3%A1tico&oldid=158513673

GATE.ac.uk—Index.html. (s. f.). Recuperado 1 de marzo de 2024, de https://gate.ac.uk/

Goldberg, Y. (2016). A Primer on Neural Network Models for Natural Language Processing. *Journal of Artificial Intelligence Research*, *57*, 345-420. https://doi.org/10.1613/jair.4992

Hidalgo, R. C. (2024, enero 24). spaCy: Una herramienta para realizar Procesamiento del Lenguaje Natural de manera amigable [Billet]. *Amontonamos las palabras: Blog de la Biblioteca de El Colegio de México*. https://bdcv.hypotheses.org/5507

Jalal, M., Mays, K. K., Guo, L., & Betke, M. (2020). *Performance Comparison of Crowdworkers and NLP Tools on Named-Entity Recognition and Sentiment Analysis of Political Tweets* (arXiv:2002.04181). arXiv. https://doi.org/10.48550/arXiv.2002.04181

Ner-annotator/docs at main · tecoholic/ner-annotator. (s. f.). GitHub. Recuperado 1 de marzo de 2024, de https://github.com/tecoholic/ner-annotator/tree/main/docs

NLTK :: Natural Language Toolkit. (s. f.). Recuperado 1 de marzo de 2024, de https://www.nltk.org/

Pinto, A., Gonçalo Oliveira, H., & Oliveira Alves, A. (2016). Comparing the Performance of Different NLP Toolkits in Formal and Social Media Text. *DROPS IDN/v2/Document/10.4230/OASIcs.SLATE.2016.3*. 5th Symposium on Languages, Applications and Technologies (SLATE'16) (2016). https://doi.org/10.4230/OASIcs.SLATE.2016.3

Procesamiento de lenguajes naturales. (2024). En *Wikipedia, la enciclopedia libre*. https://es.wikipedia.org/w/index.php?title=Procesamiento_de_lenguajes_naturales&oldid=157612573

Reconocimiento de entidades nombradas. (2023). En *Wikipedia, la enciclopedia libre*. https://es.wikipedia.org/w/index.php?title=Reconocimiento_de_entidades_nombradas&oldid=149453089

Singh, S. K. (2017). Evaluating two freely available geocoding tools for geographical inconsistencies and geocoding errors. *Open Geospatial Data, Software and Standards*, *2*(1),11. https://doi.org/10.1186/s40965017-0026-3

Software—The Stanford Natural Language Processing Group. (s. f.). Recuperado 1 de marzo

de 2024, de https://nlp.stanford.edu/software/

Spark NLP - Getting Started. (2021, marzo 20). https://sparknlp.org/docs/en/quickstart