TeresIA

Terminología +
Inteligencia Artificial

ALGORITMOS PARA LA EXTRACCIÓN DE TERMINOLOGÍAS Y SU EVALUACIÓN (INICIAL)

APETE-1.0.0-20241230

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

ÍNDICE

1. INTRODUCCIÓN Y OBJETIVOS

En el marco del proyecto TeresIA, se plantea una serie de tareas para llevar a cabo el desarrollo de este metabuscador de términos y palabras clave de textos en castellano. El Centro de Supercomputación de Barcelona (BSC-CNS) es el responsable de la **coordinación de la tarea 3**. Esta tarea consiste en el desarrollo de sistemas de extracción de términos y palabras clave, así como la evaluación de los mismos. Este entregable recoge los resultados del objetivo **E.3.3. Algoritmos para la extracción de terminologías y su evaluación (inicial)**.

Este entregable cubre el desarrollo de sistemas basados en redes neuronales profundas para la detección de términos relevantes en textos, utilizando datos generados en la tarea E.3.1. Estos sistemas se basan en el estado del arte descrito en la literatura, integrando herramientas que explotan modelos de lenguaje, como KeyBERT, y arquitecturas de entrenamiento que aprovechan la semántica contextual, como AttentionRank. Además, estas técnicas avanzadas se evalúan frente a métodos de aprendizaje automático tradicionales para medir el impacto de los modelos de lenguaje en español en la extracción terminológica, incluyendo estrategias específicas para dominios concretos.

Para el desarrollo de este objetivo, se han implementado y/o adaptado varias librerías que pretenden cubrir los escenarios o casos de uso del proyecto TeresIA, y que se presentan a continuación:

1. La librería KeyCARE para extracción de términos y palabras clave de corpus del dominio biomédico, así como su clasificación y la extracción de relaciones entre ellos (como se explica en el E.3.5). Esta librería integra diferentes módulos del estado de la cuestión, incluyendo métodos basados en Aprendizaje Profundo (Deep Learning), así como métodos estadísticos y basados en grafos. Los métodos implementados han sido escogidos por su adecuación a la tarea, así como por su baja dependencia en datos anotados.
2. Un conjunto de librerías del estado de la cuestión que se han adaptado al español y probado en el dominio jurídico, a saber, PromptRank, AttentionRank y MDERank.

2. LIBRERÍA PARA LA EXTRACCIÓN EN EL DOMINIO BIOMÉDICO: KEYCARE

Este apartado se redacta con el fin de mostrar el proceso de desarrollo de la **herramienta KeyCARE**[1], así como con el fin de ofrecer una justificación razonada

[1] https://github.com/nlp4bia-bsc/KeyCARE

de la misma. KeyCARE (Keyword extraction, term Categorization, and semantic Relation) proporciona una interfaz común para la extracción, categorización y asociación de términos extraídos de un texto. Su principal característica es que implementa métodos no supervisados y few-shot generalmente basados en Inteligencia Artificial que proporcionan alternativas viables en escenarios de pocos datos anotados. Esto ha sido implementado en una librería de Python que incluye los siguientes módulos:

1. **Extracción de palabras clave:** KeyCARE implementa varias técnicas de extracción de términos no supervisadas, las cuales permiten la extracción automática de términos clave de un texto para la expansión de terminologías, tal y como se plantea en este proyecto. Los métodos implementados se basan en diferentes técnicas, incluyendo métodos estadísticos, grafos y modelos de lenguaje, y son los siguientes:

   - **YAKE (Yet Another Keyword Extractor)[2]:** Método no supervisado de extracción de palabras clave utilizando información estadística del texto, como la frecuencia de términos, posición y coocurrencia, priorizando la relevancia dentro del documento analizado.
   - **RAKE (Rapid Automatic Keyword Extraction)[3]:** Método de extracción de términos basado en la puntuación de frases clave a partir de la frecuencia y coocurrencia de palabras. No requiere un corpus externo ni modelos entrenados.
   - **TextRank[4]:** Método de extracción de términos clave inspirado en PageRank que emplea grafos de co-ocurrencias para extraer términos relevantes mediante medidas de importancia basadas en las conexiones entre los términos.
   - **KeyBERT[5]:** Método basado en un modelo de lenguaje BERT que representa los términos en un espacio vectorial para identificar palabras clave más similares semánticamente al contenido del documento, permitiendo un enfoque basado en contexto. Al basarse en modelos de lenguaje, para su implementación se ha usado como modelo base la versión española

[2] Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018). Yake! collection-independent automatic keyword extractor. In Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40 (pp. 806-810). Springer International Publishing.

[3] Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. Text mining: applications and theory, 1-20.

[4] Wongchaisuwat, P. (2019, April). Automatic keyword extraction using textrank. In 2019 IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA) (pp. 377-381). IEEE.

[5] Sammet, J., & Krestel, R. (2023, September). Domain-Specific Keyword Extraction using BERT. In Proceedings of the 4th Conference on Language, Data and Knowledge (pp. 659-665).

de SapBERT[6], un modelo basado en la arquitectura BERT específico para el dominio biomédico.

Estos métodos han sido implementados de manera que permiten flexibilidad en su uso, pudiendo escoger por ejemplo la longitud de los términos extraídos, así como su Part-of-Speech tag (estructura sintáctica). Los cuatro métodos se han evaluado sobre corpus de entidades nombradas (NER) del ámbito médico, a falta de documentos anotados con palabras clave, sobre los cuales se ha alcanzado una sensibilidad del 89%.

2. **Categorización de términos:** KeyCARE permite el entrenamiento y la aplicación de técnicas de clustering y clasificación para separar las palabras clave en categorías predefinidas del dominio biomédico. Esto ha sido implementado tanto de forma no supervisada, en un algoritmo de clustering, como de forma supervisada, en dos clasificadores basados en modelos de lenguaje. Los clasificadores supervisados implementados emplean diferentes arquitecturas de Deep Learning, como es el clasificador de secuencias de texto estándar de Huggingface. El otro clasificador implementado es SetFit[7] (Sentence Transformers Fine-Tuning), el cual es un clasificador few-shot, es decir, que requiere mínimos datos de entrenamiento. Para ambos clasificadores se ha usado SapBERT en español como modelo base.

   Ambos clasificadores han sido entrenados sobre entidades extraídas de corpus NER generados por el grupo NLP4BIA del BSC. Estas entidades constituyen 20 clases semánticas diferentes del dominio biomédico, incluyendo enfermedades, síntomas, procedimientos y fármacos, entre otros. Además, una clase ha sido añadida para descartar aquellos términos clave extraídos incorrectamente por los extractores no supervisados, de manera que el clasificador también actúa a modo de filtro para los extractores. La evaluación de este clasificador sobre términos médicos como enfermedades, síntomas o procedimientos resulta en un f-score de 93%.

3. **Clasificación de relaciones semánticas:** la herramienta también incluye técnicas para extraer relaciones jerárquicas entre dos términos mediante clasificadores basados en Deep Learning. Esto permite interconectar los términos extraídos y puede utilizarse para el enriquecimiento terminológico, entre otras tareas. La evaluación del clasificador de relaciones jerárquicas entre términos biomédicos sobre terminologías estructuradas del dominio (SNOMED-CT) alcanza un f-score del 90%. Este módulo no es relevante para esta tarea y se desarrolla en detalle en el E.3.5.

[6] Gallego, F., López-García, G., Gasco-Sánchez, L., Krallinger, M., & Veredas, F. J. (2024). ClinLinker: Medical Entity Linking of Clinical Concept Mentions in Spanish. *arXiv preprint arXiv:2404.06367.*

[7] Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient few-shot learning without prompts. arXiv preprint arXiv:2209.11055.

Cabe mencionar que todos los módulos han sido implementados con diferentes sistemas alternativos, la mayoría de los cuales basados en Inteligencia Artificial, lo cual dota a esta librería de cierta flexibilidad y adaptabilidad a diferentes escenarios. Además, los tres módulos han sido implementados con al menos una alternativa que no precisa de datos de entrenamiento para funcionar (o que requiere muy pocos datos que se podrían generar manualmente). Esto permite el uso y aplicación de esta librería en dominios y lenguajes específicos en los cuales puede haber una falta de recursos anotados.

3. LIBRERÍAS PARA LA EXTRACCIÓN EN EL DOMINIO JURÍDICO

Para el dominio jurídico se han adaptado tres trabajos del estado de la cuestión que trabajan en dominios heterogéneos en inglés. Las librerías en cuestión son **AttentionRank, MDERank y PromptRank**. Todas ellas utilizan modelos de lenguaje.

3.1 AttentionRank

AttentionRank[8] integra los pesos de autoatención extraídos de un modelo de lenguaje preentrenado (originalmente BERT) con el valor de relevancia de atención cruzada para identificar palabras clave que son importantes en el contexto local de una oración y que, además, tienen una fuerte relevancia con todas las oraciones dentro del documento completo.

La implementación de los autores originales tuvo que ser reimplementada desde cero. El repositorio original no incluye especificaciones de bibliotecas ni versiones. Además, el código original depende de librerías para modelos de lenguaje que ya no se mantienen, así como del componente de identificación de sintagmas nominales, que utiliza la anotación POS de **Stanford CoreNLP** y librerías de terceros. No fue posible reproducir el trabajo original.

Se ha creado un nuevo repositorio[9] para la implementación del método AttentionRank. Este repositorio utiliza la librería **Transformers** de HuggingFace para usar los modelos de lenguaje y **spaCy** para identificar frases nominales. El repositorio detalla las bibliotecas específicas y sus versiones, así como los módulos externos necesarios. Además, permite el uso de modelos basados en **BERT** (como en el trabajo original) y en la arquitectura **RoBERTa** en diferentes idiomas. En este

[8] Ding, H., & Luo, X. (2021, November). Attentionrank: Unsupervised keyphrase extraction using self and cross attentions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1919-1928).

[9] https://github.com/proyectoTeresIA/attentionrank

caso concreto, se ha utilizado el modelo MarIA del BSC que fue entrenado con el corpus de la BNE para el español[10].

La adaptación para los modelos de RoBERTa tuvo que abordar dos problemas específicos relacionados con el tokenizador. El primero es el uso de diferentes *tokens* especiales para delimitar oraciones al inicio y al final, que se emplean para enfocar los mecanismos de atención. Mientras que BERT utiliza los *tokens* `[CLS]` y `[SEP]`, RoBERTa emplea `<s>` y `</s>`. El segundo problema está relacionado con los *tokens* generados, ya que BERT usa un tokenizador **WordPiece**, en el cual las subpalabras se marcan con el prefijo ## (por ejemplo, la palabra *thicknesses* se divide en los *tokens* `thickness` y `##es`). En contraste, los modelos de RoBERTa utilizan **Byte-Pair Encoding (BPE)**, clasificando los *tokens* de manera diferente según si corresponden al inicio de una palabra o si están dentro de ella. Los *tokens* que inician una palabra incluyen el espacio en blanco antes de esta y se marcan con el carácter especial Ġ. Por ejemplo, la palabra *extrapolate* se divide en dos *tokens*: `Ġextrap` y `olate`.

Más allá de las diferencias estudiadas en trabajos previos sobre los beneficios o diferencias entre ambos tipos de tokenizadores, este trabajo tuvo que desarrollar un proceso de alineación entre las palabras clave y sus correspondientes *tokens*. Con **WordPiece** es más sencillo encontrar los *tokens* y recomponer la palabra original, pero **BPE** es sensible a la aparición del espacio en blanco antes del *token*. Si este no aparece, el *token* es diferente y su valor de atención cambia. Este problema se ha solucionado asegurándose de que las oraciones de entrada siempre incluya un espacio en blanco antes de cada palabra.

3.2 MDERank

MDERank[11] se basa en la identificación de palabras clave en la representación de los embeddings de la oración utilizando *tokens* enmascarados, como en el proceso de entrenamiento de BERT. Además, su trabajo propone un nuevo tipo de arquitectura BERT diseñada para ser entrenada como un modelo de lenguaje, pero con el propósito específico de identificar palabras clave.

Para adaptar el método, se ha creado un nuevo repositorio[12], en el cual se han mejorado los requisitos, el código y el proceso de ejecución. Al igual que **AttentionRank**, **MDERank** utilizaba **Stanford CoreNLP** para la identificación de sintagmas nominales, y esto se ha actualizado a **spaCy** (para poderlo usar también en

[10] Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., ... & Villegas, M. (2021). Maria: Spanish language models. *arXiv preprint arXiv:2107.07253.*

[11] Zhang, L., Chen, Q., Wang, W., Deng, C., Zhang, S., Li, B., ... & Cao, X. (2021). MDERank: A masked document embedding rank approach for unsupervised keyphrase extraction. *arXiv preprint arXiv:2110.06651.*

[12] https://github.com/proyectoTeresIA/mderank

español). Finalmente, el método ahora puede soportar modelos de **RoBERTa**, teniendo en cuenta los problemas mencionados en **AttentionRank** del tokenizador.

3.3 PromptRank

PromptRank[13] es un modelo no supervisado que ofrece el beneficio de emplear un *prompt* para obtener términos relevantes utilizando el modelo de lenguaje previamente entrenado T5 (arquitectura codificador-decodificador).

El *prompt* no es más que una frase que proporciona contexto al decodificador acerca de la tarea a realizar. En este escenario, en el trabajo consiste en la obtención de términos relevantes, el *prompt* sería para el codificador: **"**`El párrafo: [párrafo]`**"**, y la entrada del decodificador sería **"**`Este párrafo se refiere principalmente a [candidato a término relevante]`**"**.

Para la clasificación de cuáles serán los términos relevantes seleccionados, PromptRank calcula la probabilidad de generar el candidato que estamos valorando a la salida del decodificador dada la secuencia anterior. Después, los ordena en una clasificación y escoge los candidatos que han obtenido mejores resultados. Los candidatos son extraídos del texto con expresiones regulares que buscan sintagmas nominales. Posteriormente se valora uno a uno la posibilidad de generarlo a la salida del decodificador.

PromptRank, además de emplear la probabilidad de que el candidato sea generado en la salida del decodificador, también emplea la ubicación del candidato en el documento y su longitud como otros indicadores para su categorización. Cuanto más próximo esté un aspirante al inicio, más alta será la calificación que obtiene para convertirse en una frase clave. Esto se debe a que en varias investigaciones se ha evidenciado que las expresiones que se presentan al comienzo de los documentos suelen tener más relevancia que las que hallamos hacia el final del mismo. Además, la longitud del candidato (número de palabras que contiene) puede suponer una penalización para el propio candidato. Tanto la penalización dependiendo de la longitud y la posición en el texto son modificadas por hiperparámetros.

Como el modelo original T5 era solo para español, se ha hecho una adaptación[14] para utilizar mT5 (modelo multilingüe de T5) y se ha usado spacy para extraer sintagmas nominales.

[13] Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., & Bai, X. (2023). PromptRank: Unsupervised keyphrase extraction using prompt. *arXiv preprint arXiv:2305.04490*.

[14] https://github.com/oeg-upm/PromptRankLib

4. CONCLUSIÓN

Con respecto a la librería diseñada para documentos del dominio biomédico, KeyCARE cumple con los requisitos impuestos y alcanza resultados prometedores. Aunque esta librería haya sido diseñada específicamente para su aplicación en textos del dominio biomédico, su uso puede extenderse a otros dominios o especialidades, dado que se basa en métodos mínimamente dependientes en datos anotados. Además, es importante destacar que el rendimiento de la herramienta podría optimizarse mediante el entrenamiento de modelos específicos para cada área en su caso. De la misma forma ocurre en el dominio jurídico, las herramientas reportadas, AttentionRank, MDERank y PromptRank, son no supervisadas, no se han adaptado a un dominio ni se han entrenado con un corpus anotado. Se utilizan modelos de lenguaje de base generalistas para una representación del idioma heterogéneo. Utilizar modelos de lenguaje específicos del dominio jurídico podría aumentar su rendimiento en cuanto a sus resultados.

Los próximos pasos del PT.3 consisten en evaluar la herramienta utilizando el conjunto de datos anotados por expertos (resultado del PT.3: E.3.2), los cuales proporcionarán una evaluación más fiable que la actual. Esto permitirá determinar su escalabilidad y validar su efectividad en los diferentes escenarios presentados en el proyecto.

TerêsIA

Terminología +
Inteligencia Artificial