TASS 2018: Workshop on Semantic Analysis at SEPLN, septiembre 2018, págs. 77-82

SINAI en TASS 2018 Task 3. Clasificando acciones y conceptos con UMLS en MedLine.

SINAI in TASS 2018 Task 3. Classifying actions and concepts with UMLS on MedLine

Pilar López-Úbeda, Manuel C. Díaz-Galiano,
María Teresa Martín-Valdivia, L. Alfonso Ureña-López
Departamento de informática, Centro de Estudios Avanzados en TIC (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, España
{plubeda,mcdiaz,maite,laurena}@ujaen.es

Resumen: Este artículo describe la primera participación del grupo SINAI en *Task 3. eHealth Knowledge Discovery* del Taller de Análisis Semántico en la SEPLN. Nuestro objetivo ha sido desarrollar un sistema de detección de entidades médicas en español utilizando técnicas de Procesamiento del Lenguaje Natural para finalmente clasificar términos en *acciones* o *conceptos*. Para ello, hemos utilizando la ontología biomédica UMLS además de diversos filtros para conseguir un sistema más eficaz. Los resultados obtenidos han sido satisfactorios aunque se debe mejorar la exhaustividad en cuanto a la clasificación.
Palabras clave: UMLS, Reconocedor de Entidades, Clasificación, Análisis Morfológico.

Abstract: This article describes the first participation of the SINAI group in the Task 3. eHealth Knowledge Discovery Workshop at SEPLN. Our goal has been to develop a system for detecting medical entities in Spanish using Natural Language Processing techniques to finally classify terms into *actions* or *concepts*. For this, we have used the biomedical ontology UMLS as well as several filters to make the system more efficient. The results obtained have been satisfactory although the classification needs to be improved.
Keywords: UMLS, Entity Recognition, Classification, Morphological Analysis.

1 Introducción

Hoy día, la necesidad de tener un buen sistema informático para la extracción de información en informes médicos está teniendo gran importancia, pues los datos contenidos en dichos documentos son relevantes. Aunque existen diversas herramientas y estudios que realizan detección de conceptos biomédicos (Aronson, 2001; Krauthammer y Nenadic, 2004; Osborne et al., 2007; Wright et al., 1999; Allones, Martínez, y Taboada, 2014), la mayoría son para documentos en inglés, por lo que se hace necesario el diseño y desarrollo de nuevas y potentes herramientas de procesamiento de la información en español que aprovechen los avances de las tecnologías relacionadas con la información para poder acceder y analizar estos datos.

La clasificación de entidades es una técnica englobada dentro del Procesamiento del Lenguaje Natural (PLN) que sirve para asignar un tópico o categoría de forma automática a cualquier entidad detectada en un texto.

La nueva tarea llamada *Task 3: eHealth Knowledge Discovery* que nos propone el Taller de Análisis Semántico de la Sociedad Española para el Procesamiento del Lenguaje Natural (TASS) no ha sido incluida en años anteriores (Martínez-Cámara et al., 2017), por lo que es un nuevo reto a seguir. Está inspirada en tareas como *Semeval-2017 Task 10: ScienceIE* (Gonzalez-Hernandez et al., 2017) y líneas de investigación como *Teleologies* (Giunchiglia y Fumagalli, 2017), ambas no centradas específicamente en el área de la salud. eHealth-KD propone modelar el lenguaje humano en un escenario en el que los documentos electrónicos de salud españoles puedan ser legibles por máquina desde un punto de vista semántico.

ISSN 1613-0073

Los documentos usados para esta tarea se han obtenido de MedlinePlus[1] y han sido tratados manualmente para esta tarea, tal y como se describe en el artículo resumen del TASS (Martínez-Cámara et al., 2018).

Este artículo explica nuestra participación en la tarea en las siguientes tres secciones: en la Sección 2 se describe el funcionamiento del sistema, a continuación, en la Sección 3 se muestra los resultados obtenidos y para finalizar, exponemos las conclusiones y trabajos futuros en la Sección 4.

2 Descripción del sistema

Para nuestra primera participación en esta tarea, hemos diseñado un sistema sencillo que realiza un análisis morfológico para detectar «acciones» y «conceptos», tal y como se muestra en la Figura 1. Posteriormente, utiliza una base de conocimiento biomédica para reconocer entidades médicas y finalmente aplica varios filtros para eliminar falsos positivos.

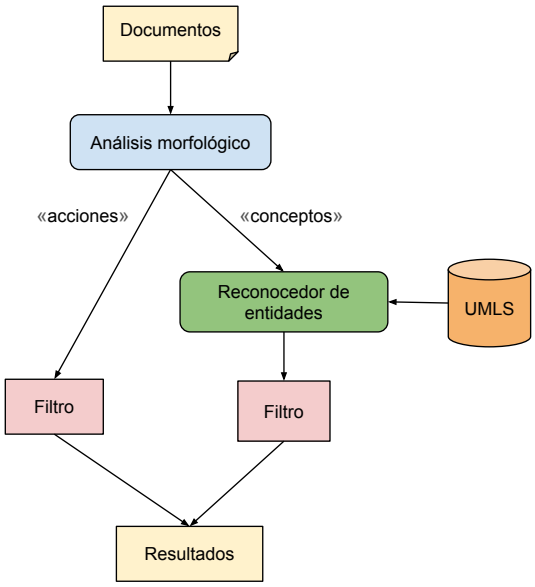

Figura 1: Arquitectura del sistema.

En las siguientes subsecciones se describen con mas detalle cada parte del sistema.

2.1 Análisis morfológico

Para el desarrollo del sistema utilizamos el analizador sintáctico incluido en la herramienta CoreNLP desarrollada por la Universidad de Stanford para el español (Manning et al., 2014).

Con esta herramienta realizamos una separación entre verbos y no verbos. Siendo los verbos etiquetados como posible «acción» y todo lo demás será analizado posteriormente para encontrar posibles «conceptos». Esta separación nos permite realizar un análisis distinto para cada tipo de entidad que queremos detectar.

Detección de acciones

Al realizar un estudio exhaustivo de los datos de test ofrecidos por la organización, se ha comprobado que las acciones compuestas por perífrasis verbales, sólo tienen marcado el verbo de la acción principal, obviándose el verbo auxiliar. Entre ellos, vemos diferentes conjugaciones de los verbos "poder", "haber", "soler", etc. Observamos que la mayoría son perífrasis de infinitivo, como por ejemplo: "pueden sufrir" o "debe hablar".

Tras ello y para llegar a realizar un sistema más eficaz, hemos obviado los siguientes verbos y sus respectivas conjugaciones: **estar**, **deber**, **poder**, **soler**.

2.2 Reconocedor de conceptos con UMLS

En esta sección, presentamos la herramienta creada para conseguir detectar «conceptos» en el texto. Esta herramienta reconoce términos biomédicos basado en *Unified Medical Languaje System* (UMLS) (Bodenreider, 2004). UMLS es un proyecto de la *National Library of Medicine* (NLM) que agrupa actualmente más de 3,46 millones de conceptos sobre dominio médico, de los cuales hemos obtenido un total de 846.044 conceptos en español.

El sistema utiliza PLN tanto en el texto de entrada como en texto de los conceptos del diccionario UMLS y busca los conceptos con mayor porcentaje de términos contenidos en el texto de entrada. El sistema realiza los pasos siguientes:

1. Normalizar: realizamos una normalización en el texto quitando acentos y signos de puntuación, utilizando para ello la Forma de Normalización de Descomposición Canónica (NFD - *Normalization Form canonical Decomposition*)(Atkin, 2005).
2. Separar términos: el separador utilizado para esta tarea es *WordPunctTokenizer*

[1]https://medlineplus.gov/

de Python *Natural Language Toolkit* (NLTK)[2].

3. Extraer raíces: el algoritmo usado para obtener la raíz de las palabras en español es SnowballStemmer[3].

El identificador de entidades médicas automático devuelve 10 resultados, cada uno de ellos compuesto por: el concepto detectado en la frase, el código UMLS asociado y el porcentaje de concordancia con el texto. Dicho porcentaje de concordancia se calcula en función del número de palabras del concepto UMLS que se pueden encontrar en la frase. En la Figura 2 podemos ver un ejemplo de los conceptos detectados para la frase "síntomas de las mujeres". Tanto concepto "síntoma" como el concepto "Mujeres" se encuentran completos dentro de la frase, por lo que el sistema devuelve el 100 % de acuerdo para ambos. Sin embargo, para el resultado 3 ("Salud de las Mujeres") que está compuesto por 4 palabras, sólo 3 de ellas se encuentran en la frase de búsqueda por lo que el sistema devuelve una coincidencia del 75 %.

```
Frase buscada:
    síntomas de las mujeres
Resultado 1
-----------
Concepto: síntoma
Código UMLS: C0043210
Coincidencia: 100%

Resultado 2
-----------
Concepto: Mujeres
Código UMLS: C1457887
Coincidencia: 100%

Resultado 3:
-----------
Concepto: Salud de las Mujeres
Código UMLS: C0080339
Coincidencia: 75%
```

Figura 2: Ejemplo de resultados obtenidos con el reconocedor automático.

[2]https://www.nltk.org/

[3]https://snowballstem.org/

2.3 Términos multipalabra

Posteriormente, intentamos localizar *conceptos* multipalabra utilizando la lista de etiquetado gramatical obtenida en la Sección 2.1 y el reconocedor de la Sección 2.2. Para ello buscamos sentencias que comiencen y acaben en sustantivo o adjetivo sin que esa frase contenga un verbo en su interior. De esta manera, intentaremos identificar conceptos en frases como: "personas con enfermedad" o "adultos de mayor edad", pero no buscaremos frases como las siguientes: "niños tienen enfermedades" o "paciente no toma aspirina".

Tras analizar las frases convenientes en nuestro sistema, calculamos una nueva métrica basada en número de palabras y número de *stop-word* contenidas tanto en la frase de entrada como en el concepto devuelto por el sistema basado en UMLS.

$$M = (TR - SWR)/(TI - SWI) \qquad (1)$$

donde:

- TR es el número de términos reconocidos por el sistema de detección.
- SWR es el número de *stop-words* recocidas por el sistema de detección.
- TI es el número de términos introducidos para buscar dentro del sistema.
- SWI es el número de *stop-words* introducidas para buscar dentro del sistema.

Finalmente, se comprueba: (1) que el concepto detectado contiene más de una palabra y (2) que la métrica M (definida en la fórmula 1) sea menor estricta que el porcentaje de acierto que devuelve el sistema, si es así, la frase introducida será válida para etiquetarla como *concepto*.

En la Figura 3 se muestran algunos ejemplos del calculo de la métrica M para finalmente concluir si es un concepto válido o no lo es. Por ejemplo, para la frase "vías urinarias" el primer resultado no es satisfactorio puesto que el concepto devuelto contiene un único término, en cambio, el resultado 2, si es un concepto válido puesto que es multipalabra y además satisface que el valor de coincidencia 1 y es igual a (2-0)/(2-0) de la fórmula 1. Por otro lado, el último ejemplo, el concepto "sin dolor

genitourinario" cumple la condición de ser multipalabra pero no satisface que 0.66 sea mayor o igual que (3-1)/(3-1).

```
Frase buscada: vías urinarias
Resultado 1:
-----------
Concepto: Urinarios
Código UMLS: C0184219
Coincidencia: 100%
M = 0.5

Resultado 2:
-----------
Concepto: Vías Urinarias
Código UMLS: C0042027
Coincidencia: 100%
M = 1

Frase buscada: análisis de orina
Resultado 1:
-----------
Concepto: análisis de orina
Código UMLS: C0042014
Coincidencia: 100%
M = 1

Frase buscada: urgencia sin dolor
Resultado 1:
-----------
Concepto: sin dolor genitourinario
Código UMLS: C0423702
Coincidencia: 66%
M = 1
```

Figura 3: Ejemplos de valores obtenidos en la detección de multipalabras.

En la Figura 4 mostramos el resultado en las distintas subtareas después de haber utilizando el reconocedor multipalabra.

3 Resultados

La tarea 3 del TASS está compuesta por varias subtareas[4]:

1. Subtarea A: Identificación de frases clave.
2. Subtarea B: Clasificación de frases clave.
3. Subtarea C: Determinación de relaciones semánticas.

La organización proponía poder participar en los siguientes escenario por subtareas: A-B-C, B-C y C. Nuestro grupo SINAI ha participado en dos de los escenarios propuestos: A-B-C y B-C. Para ambos escenarios se implementaron todas las demás funcionalidades explicadas en la Sección 2 además de utilizar el *baseline* desarrollado por la organización[5].

Los resultados obtenidos se muestran en la Tabla 1.

Round	*Precision*	*Recall*	*Micro F1 score*
A-B-C	0,84	0,62	0,71
B-C	0,90	0,54	0,67

Tabla 1: Resultados obtenidos para cada escenario.

La medidas de evaluación propuestas por la organización son: precisión, cobertura y F1. Con la precisión, podemos observar que alcanzamos clasificando en torno al 85 % y 90 %, aunque decaemos en la exhaustividad por lo que en la medida Micro F1 nos penaliza estos errores.

La media del F1 final ha sido de 0,461 ya que no hemos participado en el escenario C y con ello, no hemos podido ponderar esa parte obteniendo finalmente este bajo resultado.

4 Conclusiones y trabajos futuros

En este trabajo, el grupo SINAI presenta su primera participación a la tarea 3 del TASS. En nuestra aproximación hemos realizado un sencillo análisis del texto proporcionado, adaptando los resultados de diversas herramientas a los resultados esperados. Como novedad, hemos creado una métrica que ayuda a nuestro sistema a seleccionar aquellos conceptos de UMLS más importantes en el texto. Dicha métrica es sencilla de implementar y rápida en su ejecución, por lo que permite tener un detector de entidades biomédicas con una eficiencia moderada y un gran rendimiento.

En el ranking final expuesto por los organizadores, nuestro grupo ha quedado en segunda posición aunque no hemos llegado a participar en el escenario C y por ello se obtuvieron resultados no muy altos.

Para próximas ediciones pretendemos finalizar un nuevo sistema, el cual nos

[4]http://www.sepln.org/workshops/tass/2018/task-3/

[5]https://github.com/TASS18-Task3/data

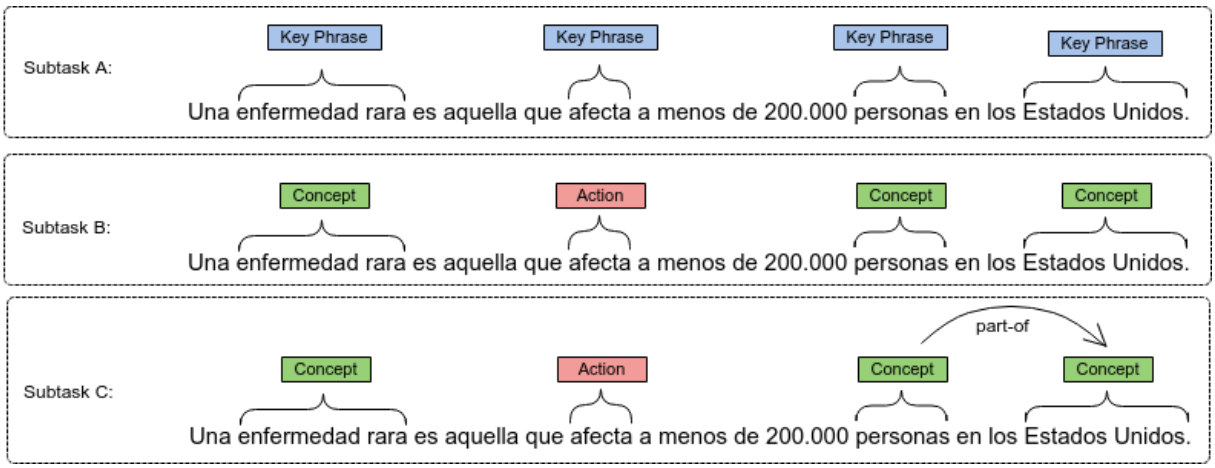

Figura 4: Ejemplo frase anotada para el escenario A-B-C

permita participar en el escenario C utilizando las relaciones entre conceptos existentes en UMLS. Sin embargo, nuestro principal objetivo consiste en mejorar el reconocedor automático de entidades biomédicas en español, incluyendo nuevas métricas que nos permitan mejorar la precisión en la detección de multipalabras. Pretendemos entrenar un sistemas de aprendizaje automático para que nos permita combinar dichas métricas de la mejor forma posible.

Agradecimientos

Este trabajo está parcialmente subvencionado por el Fondo Europeo de Desarrollo Regional (FEDER) y el proyecto REDES (TIN2015-65136-C2-1-R) del Gobierno de España.

Bibliografía

Allones, Jose Luís, Diego Martínez, y Maria Taboada. 2014. Automated mapping of clinical terms into snomed-ct. an application to codify procedures in pathology. *J. Medical Systems*, 38(10):134.

Aronson, Alan R. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. En *Proceedings of AMIA*, página 17. American Medical Informatics Association.

Atkin, Steven Edward. 2005. Meta normalization for text, Abril 19. US Patent 6,883,007.

Bodenreider, Olivier. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Giunchiglia, Fausto y Mattia Fumagalli. 2017. Teleologies: Objects, actions and functions. En *International Conference on Conceptual Modeling*, páginas 520–534. Springer.

Gonzalez-Hernandez, G, A Sarker, K O'Connor, y G Savova. 2017. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearbook of medical informatics*, 26(01):214–227.

Krauthammer, Michael y Goran Nenadic. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526.

Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, y David McClosky. 2014. The stanford CoreNLP natural language processing toolkit. En *Proceedings of 52nd annual meeting of ACL: system demonstrations*, páginas 55–60.

Martínez-Cámara, Eugenio, Yudivián Almeida-Cruz, Manuel C. Díaz-Galiano, Suilan Estévez-Velarde, Miguel Á. García-Cumbreras, Manuel García-Vega, Yoan Gutiérrez, Arturo Montejo-Ráez, Andrés Montoyo, Rafael Muñoz, Alejandro Piad-Morffis, y Julio Villena-Román. 2018. Overview of TASS

2018: Opinions, health and emotions. En Eugenio Martínez-Cámara Yudivián Almeida Cruz Manuel C. Díaz-Galiano Suilan Estévez Velarde Miguel Á. García-Cumbreras Manuel García-Vega Yoan Gutiérrez Vázquez Arturo Montejo Ráez André Montoyo Guijarro Rafael Muñoz Guillena Alejandro Piad Morffis, y Julio Villena-Román, editores, *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volumen 2172 de *CEUR Workshop Proceedings*, Sevilla, Spain, September. CEUR-WS.

Martínez-Cámara, Eugenio, Manuel C. Díaz-Galiano, Miguel Á. García-Cumbreras, Manuel García-Vega, y Julio Villena-Román. 2017. Overview of tass 2017. En Julio Villena Román M. Ángel' García Cumbreras Eugenio Martínez-Cámara M. Carlos Díaz Galiano, y Manuel García Vega, editores, *Proceedings of TASS 2017: Workshop on Semantic Analysis at SEPLN (TASS 2017)*, volumen 1896 de *CEUR Workshop Proceedings*, Murcia, Spain, September. CEUR-WS.

Osborne, John D, Simon Lin, Lihua Julie Zhu, y Warren A Kibbe. 2007. Mining biomedical data using metamap transfer (mmtx) and the unified medical language system (umls). *Gene Function Analysis*, páginas 153–169.

Wright, Lawrence W, Holly K Grossetta Nardini, Alan R Aronson, y Thomas C Rindflesch. 1999. Hierarchical concept indexing of full-text documents in the unified medical language system information sources map. *Journal of the Association for Information Science and Technology*, 50(6):514.