Noviembre, 2018.

Sistemas pregunta respuesta

Daniela Salgado*, Georgette Femerling, Karen López, Oscar Infante

Abstract

Los sistemas pregunta respuesta buscan entender preguntas para poder responderlas de manera precisa, basándose en un set de datos. Estos sistemas usan diversas herramientas de la programación neurolingüística (NLP por sus siglas en inglés) para lograr su objetivo; algunas de estas herramientas son el part-of-speech tagging, parseo semántico o la extracción de tripletas de relación de dominio abierto. El principal objetivo de este trabajo es desarrollar un Sistema Web Pregunta-Respuesta que responda las preguntas ¿qué es? y ¿para qué sirve? una entidad biológica de *Salmonella typhimurium*, la cual seleccionará el usuario. El procesamiento del conjunto de datos para lograr responder las preguntas previamente establecidas consistió de tres etapas: 1) Procesamiento de artículos, 2) Simplificación de tripletas y 3) Clasificación de tripletas. En cada etapa se hizo uso de distintas herramientas de NLP.

Keywords

BioNLP — openIE — QA systems

***Autor correspondiente**: dsalgado@lcg.unam.mx

Contenido

1. Introducción

En los últimos años, la cantidad de información ha aumentado de sobre manera y los datos que están disponibles se desestructuran cada vez más. Hoy en día, los conjuntos de datos de tan grandes dimensiones son muy complejos de analizar con las técnicas tradicionales de procesamiento de datos. Para contender con este problema, han surgido nuevas tecnologías para procesar los datos masivos como la programación neurolingüística (NLP por sus siglas en inglés). NLP busca entender cómo las personas organizan sus pensamientos, sus sentimientos y su lenguaje para poder producir los resultados que hacen. Los sistemas pregunta-respuesta es un área de investigación de la programación neurolingüística. Estos sistemas buscan entender preguntas para poder responderlas de manera precisa, basándose en un set de datos. Las respuestas de los sistemas pregunta-respuesta se basan en fragmentos extraídos de un set de datos, por lo mismo, su contenido debe ser informativo y relacionado con el tema de las preguntas.

Del mismo modo, los fragmentos candidatos, que potencialmente contienen las respuestas, deben ser analizados para extraer la respuesta solicitada. Las herramientas utilizadas para la selección de los fragmentos informativos son técnicas de NLP como el part-of-speech tagging, parseo semántico o la extracción de tripletas de relación de dominio abierto.

Las tripletas de relación representan un sujeto, una relación y un objeto al cual se relaciona. Estas relaciones son extraídas a base por medio de la extracción de información abierta (OpenIE por sus siglas en inglés) a partir de un set de datos. La generación de estas tripletas puede tener varios usos,

por ejemplo, los sujetos extraídos se pueden utilizar directamente para hacer búsquedas estructuradas.

Otra herramienta muy utilizada en NLP es la frecuencia de término – frecuencia inversa de documento (TF-IDF por sus siglas en inglés). La frecuencia de término calcula la frecuencia de cada palabra de cada documento de un conjunto de datos y la frecuencia inversa de documento se usa para calcular el peso de los términos raros en todos los documentos del set de datos. Las palabras que aparecen raramente en el cuerpo tienen una puntuación IDF alta. En conjunto, el TF-IDF determina qué tan relevante es una palabra dentro de una colección de documentos.

El principal objetivo de este trabajo es desarrollar un Sistema Web Pregunta-Respuesta que responda las preguntas ¿qué es? y ¿para qué sirve? una entidad biológica de *Salmonella typhimurium*, la cual seleccionará el usuario. Las preguntas se responderán a partir de un set de datos de aproximadamente 560 artículos.

2. Método y Materiales

El procesamiento del conjunto de datos para lograr responder las preguntas ¿qué es? y ¿para qué sirve? en función de los artículos sobre *Salmonella typhimurium* consistió de tres etapas: 1) Procesamiento de artículos, 2) Simplificación de tripletas y 3) Clasificación de tripletas. Los métodos utilizados en cada etapa se describirán a continuación.

A partir de un set de datos inicial compuesto de aproximadamente 560 artículos sobre *Salmonella typhimurium*, mediante OpenIE se extrajeron las tripletas de relación de dominio abierto, que representan una entidad biológica de *Salmonella typhimurium*, una relación y el objeto relacionado. Notamos que cada vez que corríamos OpenIE con los artículos resultaban distintas tripletas con nueva información, por esta razón, openIE se corrió dos veces con el mismo set de datos inicial, concluyendo así con la etapa del procesamiento de los artículos (Fig. 1A). Es importante resaltar que los artículos no se procesaron con openIE más de dos veces porque observamos que también se incrementaba la información redundante.

Al analizar las tripletas resultantes observamos que no todas las tripletas eran candidatas para responder las preguntas de interés, entonces procedimos a iniciar con la etapa de simplificación de tripletas (Fig. 1B). Primero se seleccionaron las tripletas que únicamente contenían información sobre un conjunto de factores sigma, factores de transcripción, unidades de transcripción y genes de *Salmonella typhimurium*, sin embargo, estas tripletas eran bastante redundantes entre ellas.

Para hacer nuestro sistema más preciso, eliminamos las tripletas redundantes primero formando conjuntos de oración que compartieran el mismo sujeto y la misma relación, después, utilizando la librería scikit-learn de Python, se calculó un score por oración usando el TF-IDF y se tomó la oración con el score más alto. A partir del último set de tripletas se realizó el mismo proceso, pero ahora formando conjuntos de oración que compartieran solo el sujeto. El set de datos que mostró tener el menor número de oraciones con información repetida fue el que se obtuvo mediante la agrupación de tripletas que compartieran las primeras dos secciones, o sea, el sujeto y la relación.

Una vez que obtuvimos el menor número de tripletas repetitivas continuamos con la tercera etapa: Clasificación de tripletas (Fig. 1C). Para cada uno de los archivos anteriores se extrajeron los lemmas de los verbos de cada tripleta, utilizando la librería Spacy de Python. Basándonos en la información de la tripleta que proporcionaba el uso de cada verbo, clasificamos los verbos en dos grupos: verbos que podrían responder a la pregunta de ¿para qué funciona la entidad? y verbos que podrían responder a la pregunta de ¿qué es dicha entidad? A partir de los verbos seleccionados para cada pregunta, procedimos a clasificar las oraciones según su verbo. Al separar las tripletas nos dimos cuenta que varias de éstas contenían información referente a experimentos, la cual era poco informativa en cuanto a las entidades, entonces procedimos a identificar las palabras características de este tipo de tripletas y a eliminar estas oraciones de nuestros datos.

Hasta el momento sabíamos que nuestro conjunto de tripletas contenía información sobre ciertos factores sigma, factores de transcripción, unidades de

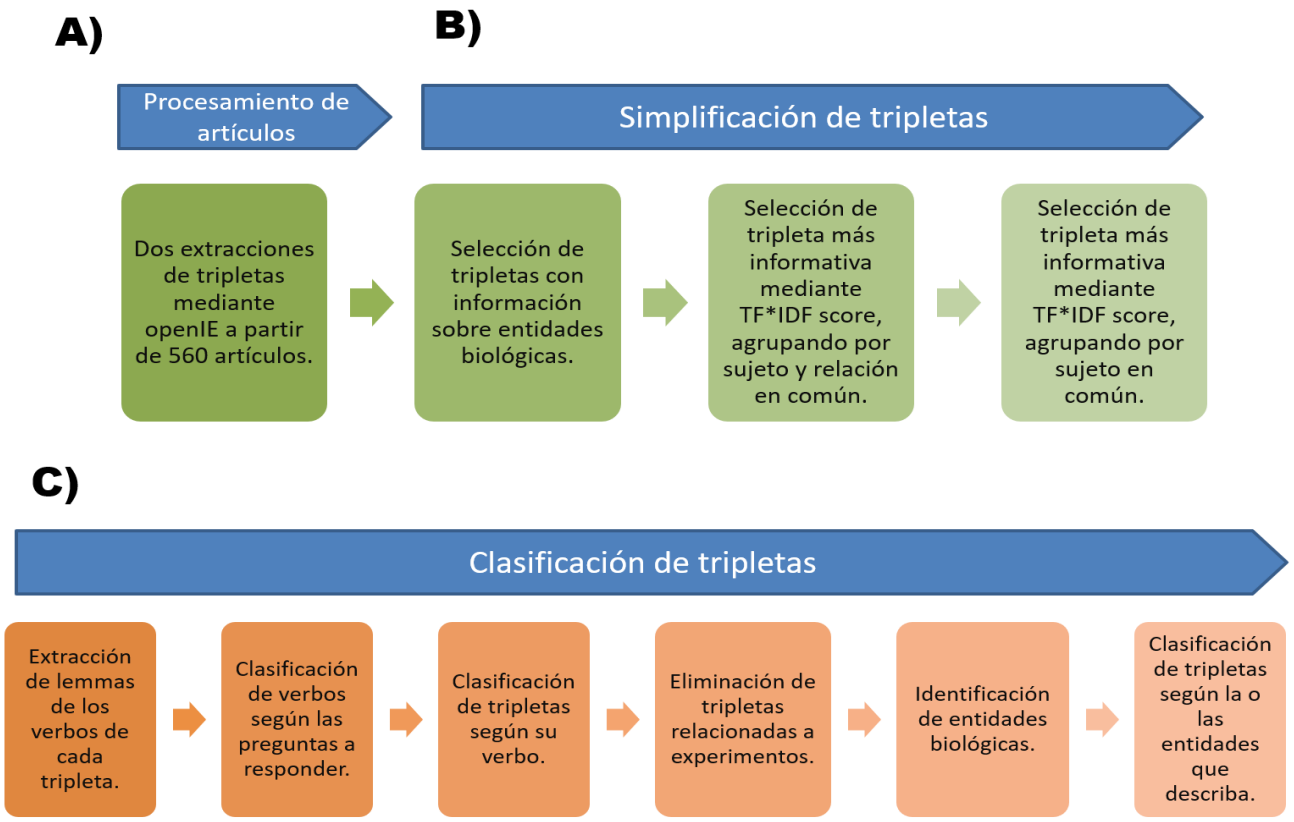

Figura 1. Método para responder las preguntas.

transcripción y genes de *Salmonella typhimurium,* sin embargo, desconocíamos las entidades específicas presentes ya que nuestro set de datos inicial había pasado por varios filtros, entonces, procedimos a extraer las entidades biológicas de *Salmonella typhimurium* presentes en nuestro set de datos.

Para esta parte proceso ya contábamos con las tripletas sin parcial redundancia, las entidades biológicas de *Salmonella typhimurium* presentes en nuestro set de datos y las tripletas clasificadas según su verbo. A partir de esta información, para cada entidad formamos conjuntos de oraciones que aportan información de la entidad y seleccionamos las dos oraciones más informativas mediante la suma de sus TF-IDF. Al seleccionar las dos oraciones más informativas esperábamos que una tripleta respondiera a la pregunta de ¿para qué funciona la entidad? y otra a la pregunta de ¿qué es dicha entidad?, sin embargo, no fue así ya que había oraciones bastante informativas, pero tan concisas que no alcanzaban un buen score de TF*IDF. Dados los resultados anteriores, optamos por formar conjuntos de oraciones que aportan información de la entidad para cada entidad.

3. Resultados y Discusión

Finalmente, el método utilizado para responder las preguntas fue clasificar las tripletas por verbo y por entidad, como mencionamos anteriormente. Así, se generaron dos archivos, uno que contiene la función de 184 entidades y otro que contenía la descripción de 199 entidades. En total el sistema pregunta-respuesta responde mínimo una de las dos preguntas para 257 entidades.

Nuestro sistema contiene información sobre 257 entidades biológicas de *Salmonella typhimurium* que comprenden factores sigma, factores de transcripción, unidades de transcripción y genes. Sin embargo, una limitación del método es que no todas las entidades

contienen información para responder ambas preguntas.

Según el método que utilizamos para responder las preguntas era fundamental la clasificación de las tripletas según su verbo. Esto representó uno de los mayores retos debido a que el significado de algunos verbos depende del contexto. Por lo anterior, una de las principales limitaciones es que el sistema no es muy consistente ya que a veces despliega respuestas intercambiadas, es decir muestra información sobre la función en las respuestas correspondientes a la pregunta ¿Qué es? y viceversa.

Una etapa importante del procesamiento de las tripletas fue identificar verbos y palabras relacionadas a la realización de experimentos, ya que observamos que eran poco informativas. Identificando dichos tokens logramos eliminar alrededor del 60% de este tipo de tripletas, sin embargo, nuestro sistema contiene algunas respuestas que describen la ejecución de algún experimento y que no aportan información relevante para describir la entidad biológica.

Una posible propuesta para un futuro trabajo relacionado a este tema sería generar conjuntos de lemmas que nos permitan precisar mejor las respuestas, o bien extraer únicamente la información relevante de los artículos antes de procesarlos con openIE.

En conclusión, los sistemas pregunta respuesta son implementaciones tecnológicas bastante útiles en diversos campos, desde el área biología hasta el área de negocios. Creemos que las herramientas actuales de la programación neurolingüística son indispensables y suficientes para el desarrollo de este tipo de sistemas, sin embargo, nuevas técnicas de implementación de estas herramientas son necesarias.

4. Referencias

Eckert, F. and Neves, M. Semantic role labeling tools for biomedical question answering: a study of selected tools on the BioASQ datasets.

Tf-idf weighting. (2009). Nlp.stanford.edu. Retrieved from https://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html

What Is NLP? | Neuro Linguistic Programming | NLP Academy. (2018). Nlpacademy.co.uk. Retrieved from https://www.nlpacademy.co.uk/what_is_nlp/

5. Material suplementario

[1] Todos los scripts y los archivos generados en este trabajo pueden consultarse en el github de la siguiente liga:

https://github.com/oinfante98/BioNLP/tree/master/QA_System