Question Aswering aplicado a la Web Semántica. Predicción de la respuesta esperada.

Matias Oyarzun and Sandra Roger

email: matias.oyarzun@est.fi.uncoma.edu.ar
roger@fi.uncoma.edu.ar

Grupo de Investigación en Lenguajes e Inteligencia Artificial
Departamento de Teoría de la Computación - Facultad de Informática
UNIVERSIDAD NACIONAL DEL COMAHUE

Resumen

El Procesamiento de Lenguaje Natural es uno de los campos más desafiantes que se tiene en la actualidad. Dentro de éste, una de las áreas que surgen naturalmente es aquella que incluye los Sistemas de Búsqueda de Respuestas (*Question Answering - QA*), cuyo objetivo consiste en dar respuestas correctas o concretas automáticamente a preguntas formuladas por el ser humano en lenguaje natural, evitando traer documentos u otros tipos de fuentes de información extensa.

Al momento de realizar este tipo de sistemas, se hacen presente múltiples dificultades. Esto se debe a que el lenguaje natural es ambiguo y por lo tanto puede ser interpretado de diversas formas. Se hace principalmente evidente durante la interpretación de preguntas, pues basta con malinterpretar el tipo de la misma para generar resultados erróneos. Aquí surge la importancia de poder determinar correctamente lo esperado por la pregunta para poder procesarla, lo que se denomina como *Question/Answer Classification*. Gracias a la evolución de la *Web Semántica*, gran parte de la información disponible en la web se encuentran en forma de bases de conocimientos (*Knowledge Bases - KB*) para ser utilizados, lo que permite minimizar la posibilidad de existencia de ambigüedades, facilitando así el trabajo necesario para el desarrollo de aplicaciones que hagan uso de los datos tal y como es el caso.

Así, el objetivo principal de este plan es la investigación y desarrollo de soluciones basadas en tecnologías de sistemas QA que permitan reducir la búsqueda de información para extraer las respuestas, sobre tecnologías de la Web Semántica a través de herramientas de Lenguaje Natural, lo que contribuye al desarrollo de agentes inteligentes inmersos en la Web. Para esto, se busca en una primera etapa poder realizar una clasificación correcta del tipo de pregunta, lo que permitirá optimizar a las siguientes etapas que abarcan el proceso de búsqueda, pues reduce el espacio de búsqueda e incluso filtrar cualquier tipo de respuesta que no sea apropiada. Una vez que este proceso de clasificación de preguntas se encuentre hecho, se procederá a afrontar el problema de localizar, extraer y presentar al usuario aquella información que desea conocer, mediante una respuesta concreta y de la forma más amigable posible.

Palabras Clave: Sistemas de Búsqueda de Respuestas, Question Answering, Predicción de la Respuesta Esperada, Generalización de Texto, Web Semántica, Procesamiento de Lenguaje Natural.

Contexto

Este trabajo está parcialmente financiado por la UNCo, en el marco del nuevo proyecto de in-

vestigación *Tecnologías Semánticas para el desarrollo de Agentes Inteligentes*. Como así también, lo financia parcialmente el Consejo Interuniversitario Nacional (CIN) con una Beca de Estímulo a las Vocaciones Científicas 2021. El proyecto de investigación tiene una duración de cuatro años y ha comenzado en 2022 y se desarrolla en forma colaborativa con docentes-investigadores de la UNS.

1. Introducción

El rápido aumento de la información y la popularidad del uso de la web se debe a que las personas comienzan a almacenar datos y poner los mismos a disposición del público. Ésto, provoca que el desarrollo de aplicaciones que acceden a estas grandes cantidades de información que se modifican en tiempo real, de manera constante y que posiblemente se encuentren en diferentes formatos, presenten problemas al momento de la exploración de éstos y hace que la búsqueda de información sea una tarea compleja y costosa en términos de tiempo. Pues, muchas de estas aplicaciones ocurren bajo restricciones de tiempo críticas y en intensa interacción con el usuario. Esta dificultad ha motivado el desarrollo de nuevas herramientas de investigación adaptadas, como los sistemas de QA. Por ello, la representación conceptual de los dominios para la generación y extracción de información y conocimiento, es central en la toma de decisiones. De esta manera, los sistemas de QA contribuyen a que el usuario sea capaz de formular una pregunta en lenguaje natural y obtenga una respuesta concreta en lugar de un conjunto de documentos considerados relevantes, como es el caso de los motores de búsqueda.

A su vez, la Web Semántica es un ambiente ideal para el desarrollo de este tipo de agentes, pues ésta busca crear una web de conocimiento en la cual la semántica del contenido es explicitada, permitiendo novedosas aplicaciones que combinan datos heterogéneos para, entre otros objetivos, mejorar la experiencia de los usuarios de acuerdo a sus necesidades. Por ello, lograr que este valioso conocimiento semántico sea accesible y utilizables por los usuarios finales es de principal importancia en los sistemas de QA sobre KB.

Bajo este aspecto, el objetivo general que persigue el proyecto de investigación es el de generar conocimiento especializado en el área de agentes inteligentes y en lo referente a la representación y el uso del conocimiento en sistemas computacionales basados en la web, es decir lo que se ha llamado Web Semántica. Para ello, es necesario profundizar en el estudio de técnicas de representación de conocimiento y razonamiento, tecnologías del lenguaje natural, metodologías de modelado conceptual y mecanismos para la interoperabilidad de aplicaciones, tanto a nivel de procesos como de datos. Se pretende aplicar estos conceptos como soporte para comunidades de desarrollo de ontologías, entre otros.

De esta manera, se busca en una primera instancia la investigación en el área del *Question/Answer Classification* de un sistema QA, pues desempeña un rol importante al momento de determinar las expectativas del usuario. Su objetivo es identificar el tipo de pregunta y, basándose en el mismo, extraer la respuesta esperada de los datos [1]. Ésto permitiría a los sistemas QA identificar de manera más precisa una respuesta adecuada a la pregunta.

El desarrollo del plan de trabajo se realizará en el marco del proyecto de investigación "Tecnologías Semánticas para el desarrollo de Agentes Inteligentes". En dicho proyecto de investigación se desarrolla una línea de investigación que explora sobre temas afines tanto al análisis y desarrollo de técnicas y herramientas útiles tanto el Procesamiento en Lenguaje Natural como la Generación del lenguaje Natural con el objetivo de dar soporte a los agentes inteligentes en estudio y en la definición de metodologías basadas en técnicas del Lenguaje Natural para el modelado de herramientas de búsquedas de respuestas semánticas. Particularmente, se ha escogido experimentar sobre herramientas de QA que den soporte a la búsqueda de información en el ámbito de la Web Semántica.

2. Línea de Investigación y Desarrollo

El proyecto de investigación *Tecnologías Semánticas para el desarrollo de Agentes Inteligentes* tiene como objetivo general, generar conocimiento especializado en el área de agentes inteligentes que accedan, procesen y recuperen información mediados por tecnologías semánticas.

En este sentido, se desarrolla una línea de investigación que explora sobre técnicas de representación de conocimiento y razonamiento, tecnologías del lenguaje natural, metodologías para la interoperabilidad e integración de datos, y generar nuevos principios, metodologías formales y herramientas basadas en la gestión semántica de los datos. Particularmente, se ha escogido experimentar sobre herramientas de QA que den soporte a la búsqueda de información en el ámbito de la Web Semántica.

El proceso de QA consta de una etapa de análisis de la pregunta, recuperación de los datos relevante de fuentes de conocimiento y la extracción de la información concreta y correcta como respuesta.

El análisis de la pregunta es fundamental. En este sentido, continuando con [2] nos concentramos en una primera etapa en la predicción de la respuesta esperada a partir de la pregunta de entrada. En este sentido, se ha realizado un análisis de las diferentes metodologías y estudio de herramientas disponibles.

Dentro de esta subtarea de QA focalizada en la predicción del tipo de respuesta, existen distintas competencias, una de ella es el desafío denominado SMART *SeMantic Answer Type and Relation Prediction Task*, de la cual se han realizado hasta el momento dos instancias de tales competencias: año 2021[1] y 2020[2] [3].

La predicción de relaciones para la pregunta es una tarea difícil: algunas relaciones están alejadas semánticamente, a veces los tokens que deciden las relaciones están distribuidas a lo largo de la pregunta, algunas relaciones están implícitas en el texto, entre otras.

Es posible una clasificación de tipo de respuesta granular con ontologías de Web Semántica populares como DBpedia (∼760 clases) y Wikidata (∼50K clases). En esta competencia se cuenta con dos tareas principales e independientes: 1) predicción del tipo de respuesta y 2) predicción de un conjunto de relaciones usadas para la identificación de la respuesta correcta.

Se está desarrollando un módulo para la clasificación del tipo de respuesta utilizando aprendizaje automático. La competencia dispone de varios corpus que se pueden utilizar para clasificar la categoría (Boolean, Literal, Resource) y el tipo de respuesta para cada una de las diferentes ontologías que proponen.

Asimismo, se pretende diseñar y desarrollar un módulo para la segunda tarea de predicción de relaciones usando tanto la ontología de DBpedia como la de Wikidata. Al igual que en la tarea uno, se provee de corpus para trabajar.

3. Resultados Obtenidos y Trabajos Futuros

En una primera instancia se ha realizado el relevamiento y análisis de las diferentes estrategias y características empleadas en los sistemas de búsquedas de respuestas semánticos, entre otras herramientas consideradas de utilidad. A partir de este análisis, nos encontramos en la fase de diseño e implementación de los módulos correspondientes a nuestro sistema de búsqueda de respuesta semántico, tales como aquellos para la clasificación de la pregunta y los encargados de la búsqueda y análisis de las posibles respuestas. Esto es un primer paso a nuestro objetivo de implementar nuestro primer prototipo dentro del marco del proyecto de investigación.

La finalidad de nuestra propuesta es la implementación de un Sistema de Búsqueda de Respuesta aplicado sobre la Web Semántica, que nos brinde una respuesta precisa a una pregunta planteada en lenguaje natural, en lugar de una lista de enlaces a documentos como lo hacen los motores de búsquedas tradicionales. Además, como mencionamos en [2], se pretende ampliar

[1] https://smart-task.github.io/2021/
[2] https://smart-task.github.io/2020/

el conocimiento en el área de Ciencias de la Computación, y por sobre todo en los campos de la Web Semántica y la Ontología.

4. Formación de Recursos Humanos

Durante la realización de esta investigación se espera lograr, como mínimo, la culminación de 2 tesis de grado dirigidas y/o codirigidas por los integrantes del proyecto. Uno de los autores de este trabajo posee una Beca de Estímulo a las Vocaciones Científicas (CIN) 2021.

Finalmente, es constante la búsqueda hacia la consolidación como investigadores de los miembros más recientes del grupo.

Referencias

[1] Riyanka Manna, Dipankar Das, and Alexander Gelbukh. Question classification in a question answering system on cooking. In *Mexican International Conference on Artificial Intelligence*, pages 103–108. Springer, 2020.

[2] Matías Oyarzun and Sandra Roger. Tecnologías de sistemas de QA aplicadas a la Web Semántica. In *XXIII Workshop de Investigadores en Ciencias de la Computación (WICC 2021, Chilecito, La Rioja)*, 2021.

[3] Nandana Mihindukulasooriya, Mohnish Dubey, Alfio Gliozzo, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, and Ricardo Usbeck. SeMantic AnsweR Type prediction task (SMART) at ISWC 2020 Semantic Web Challenge. *CoRR/arXiv*, abs/2012.00555, 2020.