Procesamiento del Lenguaje Natural, Revista nº 44, marzo de 2010, pp 155-156 recibido 08-01-10 revisado 10-02-10 aceptado 04-03-10

Sistemas de clasificación de preguntas basados en corpus para la búsqueda de respuestas

Corpus-based question classification in question answering systems

David Tomás
Depto. de Lenguajes y Sistemas Informáticos - Universidad de Alicante
Carretera San Vicente del Raspeig s/n 03690, Alicante, España
dtomas@dlsi.ua.es

Resumen: Esta tesis se centra en el desarrollo de sistemas automáticos de clasificación de preguntas fácilmente adaptables a diferentes idiomas y dominios de trabajo. Estos sistemas se basan en técnicas de aprendizaje automático sobre corpus, siguiendo un enfoque estadístico del tratamiento del lenguaje humano. El objetivo es evitar en gran medida el uso de herramientas y recursos lingüísticos más allá de los propios corpus de aprendizaje, obteniendo sistemas que destacan por su flexibilidad y sus escasos requerimientos.
Palabras clave: Clasificación de preguntas, búsqueda de respuestas, aprendizaje supervisado, aprendizaje semisupervisado, aprendizaje mínimamente supervisado

Abstract: This thesis is focused on the development of question classification systems that are easily adaptable to different languages and domains. These systems are based on machine learning techniques and corpus, following a statistical approach to human language. The goal is to almost avoiding the need for linguistic tools and resources, obtaining flexible systems with few requirements.
Keywords: Question classification, question answering, supervised learning, semi-supervised learning, minimally-supervised learning

1. Introducción

Los sistemas de *búsqueda de respuestas* (BR) o *question answering* tienen como finalidad encontrar respuestas concretas a necesidades precisas de información formuladas por los usuarios mediante lenguaje natural. En los sistemas de BR, un primer paso para poder devolver la respuesta solicitada por el usuario es analizar la pregunta y comprenderla, saber sobre qué se nos está preguntando.

La *clasificación de preguntas* se ha desmarcado como una tarea en sí misma dentro del mundo del procesamiento del lenguaje natural y de la BR. Su objetivo es identificar de forma automática qué se nos está preguntando, categorizando las preguntas en diferentes clases semánticas en función del tipo de respuesta esperada. Así, ante preguntas como "¿Quién es el presidente de los Estados Unidos?" o "¿Dónde está la Torre Eiffel?", un sistema de clasificación de preguntas detectaría que se está preguntando por una *persona* o un *lugar* respectivamente.

En esta tesis se han desarrollado una serie de aproximaciones al desarrollo de sistemas de clasificación de preguntas flexibles, entendiendo por flexibilidad la capacidad del sistema para adaptarse de forma sencilla a diferentes idiomas y dominios. Para ello se han empleado técnicas de aprendizaje automático sobre corpus, permitiendo a estos sistemas mejorar a través de la experiencia sin necesidad de conocimiento humano.

2. Aportaciones

En este trabajo se han desarrollado tres aproximaciones a la tarea de clasificación de preguntas, buscando en cada una de ellas reducir, con respecto a la anterior aproximación, la necesidad de recursos para la construcción del clasificador.

Clasificación supervisada basada en n-gramas. En esta primera aproximación el clasificador aprende de forma automática a partir de información obtenida estrictamente de un corpus de entrenamiento. No se requie-

ISSN 1135-5948

re ningún otro tipo de herramienta o recurso lingüístico. El objetivo es establecer un sistema de referencia para aquellas situaciones en las que únicamente se dispone de un corpus para el aprendizaje. Llevar a cabo esta aproximación requiere la ejecución de diversas tareas:

- Determinar el algoritmo de aprendizaje más apropiado para la tarea de clasificación de preguntas.
- Analizar diferentes características de aprendizaje a nivel de palabra obtenidas exclusivamente de los datos de entrenamiento.
- Desarrollar corpus en diferentes idiomas para la tarea de clasificación multilingüe.
- Desarrollar corpus en diferentes dominios para la tarea de clasificación en dominio abierto y restringido.
- Evaluar y comparar los algoritmos y características anteriores sobre los corpus desarrollados.
- Estudiar diferentes técnicas de selección de características.

Clasificación semisupervisada empleando textos no etiquetados. En esta segunda aproximación se enriquece el modelo básico definido anteriormente, completando la información extraída del conjunto de entrenamiento mediante información semántica externa. Esta información se obtiene a partir de texto no etiquetado adquirido de forma automática de la Web. De esta forma se consigue mejorar la capacidad del sistema empleando datos no etiquetados, dando lugar a una aproximación semisupervisada a la clasificación de preguntas. Las tareas realizadas en esta aproximación son:

- Incorporar información semántica partiendo de texto no etiquetado, empleando *métodos kernel* y análisis de la semántica latente (LSA).
- Comparar esta aproximación con otros sistemas que incorporan información semántica proveniente de recursos lingüísticos complejos.
- Evaluar el sistema sobre diferentes idiomas.

Clasificación mínimamente supervisada sobre taxonomías refinadas. En esta tercera aproximación afrontamos el problema de la clasificación de preguntas sobre taxonomías refinadas en ausencia de datos de entrenamiento. A partir de un pequeño conjunto de semillas iniciales definidas por el usuario para cada clase, el sistema aprende a discriminar de forma automática entre ellas a partir de información adquirida de forma automática de la Web. De esta forma se obtiene una aproximación mínimamente supervisada a la clasificación sobre taxonomías refinadas, que tradicionalmente requeriría de grandes corpus de entrenamiento para ofrecer una cobertura adecuada al problema. Las tareas realizadas en esta fase son:

- Definir un modelo para la adquisición automática de muestras de entrenamiento, evitando así la necesidad de grandes conjuntos de datos para el aprendizaje.
- Desarrollar un algoritmo que permite aprovechar estas muestras para la construcción del clasificador.
- Desarrollar un conjunto de datos de evaluación sobre una taxonomía refinada que nos permita medir el rendimiento del sistema.
- Evaluar el sistema sobre diferentes idiomas.
- Comparar esta aproximación con los sistemas de aprendizaje empleados habitualmente en esta área, valorando las ventajas aportadas.

Información adicional

Tesis doctoral en Informática realizada en la Universidad de Alicante por David Tomás Díaz, bajo la dirección del doctor José Luis Vicedo González. La defensa tuvo lugar el día 21 de julio de 2009 ante un tribunal formado por los doctores Manuel Palomar (Univ. de Alicante), Patricio Martínez (Univ. de Alicante), Horacio Rodríguez (Univ. Politécnica de Cataluña), Paolo Rosso (Univ. Politécnica de Valencia) y Günter Neumann (DFKI). La calificación obtenida fue Sobresaliente *Cum Laude* por unanimidad, con mención de Doctor Europeo.[1]

[1] El texto completo de la tesis está disponible en `http://gplsi.dlsi.ua.es/gplsi09/lib/exe/fetch.php?media=tesis_david.pdf`.