Advanced techniques for domain adaptation in Statistical Machine Translation

Doctorando: Mara Chinea Rios
Directores: Francisco Casacuberta y Germán Sanchis Trilles

Resumen

La Traducción Automática Estadística es un sup-campo de la lingüística computacional que investiga como emplear los ordenadores en el proceso de traducción de un texto de un lenguaje humano a otro. La traducción automática estadística es el enfoque más popular que se emplea para construir estos sistemas de traducción automáticos. La calidad de dichos sistemas depende en gran medida de los ejemplos de traducción que se emplean durante los procesos de entrenamiento y adaptación de los modelos. Los conjuntos de datos empleados son obtenidos a partir de una gran variedad de fuentes y en muchos casos puede que no tengamos a mano los datos más adecuados para un dominio específico. Dado este problema de carencia de datos, la idea principal para solucionarlo es encontrar aquellos conjuntos de datos más adecuados para entrenar o adaptar un sistema de traducción para un dominio o tarea específico. En este sentido, esta tesis propone un conjunto de técnicas de selección de datos que identifican los datos bilingües más relevantes para una tarea extraídos de un gran conjunto de datos. Algunas de estas técnicas aprovechan las ventajas que presenta la representación vectorial del texto en un espacio continuo.

Como primer paso en esta tesis, las técnicas de selección de datos son aplicadas para mejorar la calidad de la traducción de los sistemas de traducción automática estadísticos bajo el paradigma basado en frases. Estas técnicas se basan en el concepto de representación continua de las palabras o las oraciones en un espacio vectorial. Las técnicas desarrolladas fueron aplicadas a la tarea de aumentar el tamaño de un conjunto de entrenamiento pequeño que pertenece al dominio de la tarea. Los resultados experimentales presentados para esta tarea demuestran que es posible lograr un aumento de la calidad de la traducción y al mismo tiempo una reducción significativa en el tamaño del conjunto de entrenamiento. Otra tarea dentro de este paradigma fue seleccionar los mejores conjuntos de desarrollo que se emplean durante el proceso de ajuste de pesos del modelo log-lineal. Enfocándonos en este problema, en esta trabajo se presentan diferentes métodos para la selección de los conjuntos de desarrollo, prestando especial atención a aquellos casos que sólo tenemos disponible el conjunto de oraciones a traducir. Los resultados experimentales demuestran que las técnicas utilizadas son efectivas para diferentes lenguajes y dominios. Además, los experimentos llevados a cabo en un entorno real son muy positivos y demuestran la efectividad de los métodos.

El paradigma de Traducción Automática Neuronal también fue aplicado en esta tesis. Dentro de este paradigma, investigamos la aplicación que pueden tener las técnicas de selección de datos anteriormente validadas en el paradigma basado en frases. El trabajo realizado se centró en la utilización de dos tareas diferentes de adaptación del sistema. Por un lado, investigamos cómo aumentar la calidad de traducción del sistema, aumentando el tamaño del conjunto de entrenamiento. Los conjuntos de entrenamiento que se emplearon se construyeron

concatenando el conjunto de entrenamiento de dominio y el sub-conjunto obtenido por un método de selección. Por otro lado, el método de selección de datos se empleó como una estrategia eficiente para crear un conjunto de datos sintéticos. Estos conjuntos sintéticos fueron empleados para adaptar un sistema de traducción automática neuronal general al dominio que deseábamos. Los experimentos se realizaron para diferentes dominios y los resultados de traducción obtenidos son convincentes para ambas tareas.

Además de la tarea de selección de datos, se prestó atención al proceso de optimización de los pesos del modelo log-lineal. Con este propósito, se estudió un método de optimización que pretende aumentar la calidad de la traducción de un sistema para un dominio específico, todo lo referente a este tema se puede encontrar en un Apendice de esta tesis.

Finalmente, cabe señalar que las técnicas desarrolladas y presentadas a lo largo de esta tesis pueden implementarse fácilmente dentro de un escenario de traducción real; donde el sistema de traducción esta diseñado para resolver un problema real existente.