MULTI: MULtilingual Transfer learning for the Inclusion of vulnerable social groups

Description

Hasta hace poco, el desarrollo de sistemas de TA se limitaba necesariamente a pares de idiomas para los que existían grandes corpus paralelos que podían utilizarse para entrenar modelos de traducción. En 2018, dos enfoques concurrentes [Artetxe, 2018] y [Lample,2018] demostraron que la TA no supervisada (utilizando solo corpus monolingües) era posible.

El proyecto CEF MT4All participado por la UPV-EHU y el BSC, se basa en la metodología propuesta por [Artexte, 2018] para crear motores de TA para varios escenarios con pocos recursos. Trabajos más recientes como [Vergés et al. 2020] han demostrado que el uso de modelos multilingües es beneficioso, ya que generaliza mejor al compartir parámetros entre todas las lenguas implicadas, especialmente si las lenguas pertenecen a la misma familia lingüística.

Al mismo tiempo, entrenar modelos multilingües de TA desde cero suele requerir grandes corpus paralelos y puede no ser factible en escenarios de traducción con pocos recursos. Los modelos multilingües, como XLM y XLM-RoBERTa [Conneau, 2020], que combinan objetivos de entrenamiento no supervisados (datos monolingües) y supervisados (datos paralelos), funcionan especialmente bien.En [Kharitonova, de Gibert, Armengol, Rodríguez y Melero, 2021], reutilizamos esta idea inicializando el codificador con un XLM Roberta preentrenado, pero a diferencia de los enfoques anteriores, sólo inicializamos el codificador, para instanciar un decodificador menos profundo, por razones de eficiencia.

Un enfoque alternativo, aún inexplorado, es, en lugar de adaptar un modelo multilingüe, reciclar los pesos de un modelo monolingüe (inglés) como BART (un autocodificador con eliminación de ruido) y sustituir la capa de embeddings por una capa entrenada en los datos de la lengua con pocos recursos. Este novedoso enfoque tiene la ventaja de la modularidad (permite añadir nuevas lenguas sin reentrenar desde cero) y la simplicidad.

Además, como ventaja extra, el método es capaz de producir un modelo lingüístico completo para la lengua de destino, que puede utilizarse potencialmente de forma competitiva en otras tareas monolingües.En este proyecto queremos explorar estos y otros métodos de aprendizaje por transferencia, enriquecidos con técnicas semisupervisadas siempre que haya datos paralelos disponibles.

Barcelona Supercomputing Center - Centro Nacional de Supercomputación