Un sistema de traducción automática neuronal para todas las lenguas románicas de la península ibérica

Un proyecto de investigación aplicará la traducción automática neuronal al español, el portugués, el catalán, el gallego, el asturiano, el aragonés y el aranés

El hecho de formar parte de la revolución digital puede contribuir a impulsar el uso de las lenguas con menos hablantes

(Zaragoza/Barcelona, jueves, 6 de julio de 2023). En los últimos años, el número y la eficacia de las tecnologías de traducción automática se han disparado. Gracias a la inteligencia artificial (IA), llevamos en nuestro bolsillo potentes herramientas que traducen con facilidad cualquiera de las lenguas mayoritarias. Pero ¿qué pasa con aquellas que tienen menos hablantes y menos recursos? **¿Cómo puede llegar a "entenderlas" una IA?** La respuesta podría estar en el aprendizaje por transferencia y el entrenamiento de sistemas multilingües para las lenguas románicas de la península ibérica.

El proyecto Traducción automática neuronal para las lenguas románicas de la península ibérica (TAN-IBE) explora las técnicas más efectivas para **entrenar sistemas de traducción automática basados en redes neuronales** (un tipo de IA) aplicadas a siete lenguas románicas de la península ibérica: español, portugués, catalán, gallego, asturiano, aragonés y aranés.

Este proyecto, financiado por el Ministerio de Ciencia, Innovación y Universidades, está coordinado por la **Universitat Oberta de Catalunya (UOC), con la participación de las universidades de Zaragoza, Oviedo y Lleida.** El proyecto ha contado con la participación de **Alejandro Pardos**, egresado en **Lenguas Modernas por la Universidad de Zaragoza**, doctorando y **especialista en Lengua Aragonesa.**

Una IA que transfiere conocimiento entre lenguas

Los sistemas de traducción basados en redes neuronales se entrenan a partir de millones de oraciones en una lengua con su traducción en otra lengua. Es lo que se conoce como ***corpus paralelos*, inmensos conjuntos de datos disponibles en dos lenguas**. Una vez que la red neuronal está entrenada, es capaz de traducir con eficacia cualquier texto en esas lenguas. El problema es que, con idiomas como el español o el portugués, es sencillo encontrar esos corpus paralelos, pero con aquellas lenguas que tienen menos material disponible —como el aranés, el aragonés o el asturiano— es complicado tener suficientes datos para entrenar a la inteligencia artificial.

"Lo bueno es que los sistemas neuronales pueden aprender cosas de una lengua a partir de otra que se le parezca", explica Antoni Oliver, investigador del grupo de investigación interuniversitario en Aplicaciones Lingüísticas (GRIAL-UOC), coordinador del proyecto TAN-IBE y profesor de los Estudios de Artes y Humanidades de la UOC. "Por eso escogimos las lenguas románicas. El proceso deberá **ser capaz de aprender por transferencia** utilizando un modelo entre dos lenguas para construir el sistema de traducción entre otras dos. Así, por ejemplo, cuando esté terminada, la herramienta de traducción español-aranés habrá aprendido en parte gracias al sistema español-catalán o al español-portugués", añade.

Unidad de Cultura Científica
660 010 349
comunica@unizar.es

La construcción del modelo de traducción **no es el único objetivo** del proyecto de investigación, que busca, además:

· Compilar **corpus paralelos y monolingües** para las siete lenguas románicas que se incluyen en la propuesta, dedicando un mayor esfuerzo al asturiano, el aragonés y el aranés.

· Explorar **nuevas técnicas** para el entrenamiento de sistemas de traducción automática neuronal. Además del aprendizaje por transferencia, se estudiará la traducción automática multilingüe, la traducción automática autosupervisada y la traducción automática no supervisada.

· Entrenar **sistemas de traducción automática neuronal** entre el español y el resto de las lenguas del proyecto, en ambas direcciones.

· Entrenar **sistemas multilingües** capaces de traducir desde y hacia todas las lenguas del proyecto.

· Crear **guías y *scripts*** que faciliten el entrenamiento de sistemas de traducción automática neuronal en general y, más en concreto, para las lenguas del proyecto.

· Publicar los resultados del proyecto **con licencias libres**. Esto incluye los corpus compilados, los modelos y motores de traducción automática y las guías y *scripts*.

El proyecto consiste, en primer lugar, en recopilar todos los corpus para las lenguas con menos material (asturiano, aragonés y aranés), y, en segundo lugar, en entrenar los sistemas de traducción. El resultado final del proyecto será tanto **la publicación libre de los recursos**, en la medida que sea posible, como la creación de un sistema de traducción automática neuronal libre de uso.

Acuerdos y estudios para impulsar las lenguas minoritarias

La primera parte del proyecto está llevándose a cabo fuera de los laboratorios. Para disponer de los datos necesarios para entrenar los modelos de inteligencia artificial, es necesario recopilar el máximo material posible del **asturiano, el aragonés y el aranés**. Por eso, esta primera fase se centra en lograr acuerdos con gobiernos autonómicos, universidades o editoriales para que nos faciliten el material para crear los corpus paralelos con los que entrenar al sistema neuronal.

En este sentido, el pasado mes de mayo se alcanzó un importante **acuerdo con el gobierno del Principado de Asturias** para la cesión de todo el corpus de textos traducidos del castellano al asturiano que posee la Dirección Xeneral de Política Llingüística. El convenio recoge también que, si el Principado lo requiere, podrá disponer de los desarrollos tecnológicos y lingüísticos del proyecto TAN-IBE para su aprovechamiento en posibles proyectos propios de traducción automática.

En última instancia, con este proyecto se pretende ayudar a **fomentar el uso de las lenguas** con menos recursos y que se publique más en dichas lenguas. "Por ejemplo, todas las leyes podrían publicarse en dos lenguas de forma rápida y eficiente, invirtiendo menos recursos, aunque siempre se necesitaría una revisión humana. Además, las personas que no se atreven a usar estas lenguas porque no se sienten seguras pueden apoyarse en estas herramientas para mejorar sus textos. Por último, las lenguas como el asturiano, el aragonés o el aranés deben formar parte de las tecnologías digitales. Si no, pueden ir desapareciendo y ser olvidadas", concluye Oliver.

Más información: Alejandro Pardos

Contacto para los medios: Carmina Puyod, UCC-Unizar: 660 010 349

Unidad de Cultura Científica
660 010 349
comunica@unizar.es

Unidad de Cultura Científica
660 010 349
comunica@unizar.es