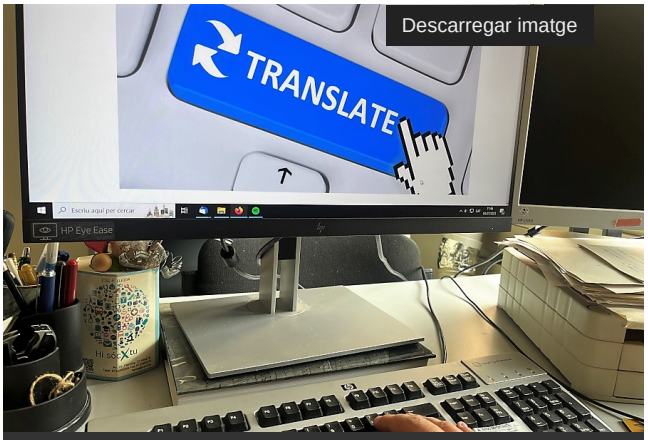
dimecres, 05 de juliol de 2023

La UdL, en un projecte de traducció automàtica neuronal

Inclourà totes les llengües romàniques de la península Ibèrica

La Universitat de Lleida (UdL) participa en un projecte de recerca, liderat per la Universitat Oberta de Catalunya (UOC), per desenvolupar un sistema basat en la Intel·ligència Artificial (IA) que tradueixi de forma automàtica totes les llengües romàniques de la península Ibèrica: castellà, català, portuguès, gallec, asturià, aragonès i aranès / occità. Amb la col·laboració de les universitats d'Oviedo i Saragossa, l'equip explora les tècniques més efectives per *entrenar* sistemes de traducció basats en xarxes neuronals [https://ca.wikipedia.org/wiki/Xarxa_neuronal_artificial], un model matemàtic de processament de dades que imita les connexions del sistema nerviós animal.



La IA ajuda a explorar similituds entre idiomes / Foto: UdL

La traducció automàtica neuronal [https://es.wikipedia.org/wiki/Traducci%C3%B3n_autom%C3%A1tica_neuronal] treballa amb corpus [https://es.wikipedia.org/wiki/Corpus_ling%C3%BC%C3%ADstico] paral·lels, és a dir, conjunts de segments o oracions en una llengua amb els seus equivalents de traducció en una altra. Aquests sistemes no es desenvolupen, sinó que s'*entrenen*; és a dir, aprenen a traduir a partir de textos en la llengua de partida i en la d'arribada. Per a fer-ho, necessiten com a mínim entre 5 i 10 milions d'oracions. Com que aquests corpus no estan disponibles per a tots els parells de llengües, les investigadores i els investigadors se centren en l'aprenentatge per transferència (*transfer learning*). Es tracta d'aprofitar el coneixement d'un parell de llengües amb molts recursos i transferir-lo a altres que en tenen menys. Per exemple, per a entrenar un sistema castellà - aranès, que presenta molt pocs recursos, es pot utilitzar el coneixement d'un altre parell com el castellà - català, que disposa de grans corpus paral·lels.

Una altra tècnica que estan explorant és l'*entrenament* de sistemes multilingües per explotar les similituds entre idiomes. En un sistema com aquest, els parells de llengües amb menys recursos, com per exemple l'espanyol - aranès, s'aprofiten del coneixement après per altres parells, com l'espanyol - portuguès o l'espanyol - català. Els sistemes *entrenats* d'aquesta manera són fins i tot capaços de traduir entre parells de llengües per als quals no existeixen oracions paral·leles en el corpus d'entrenament, com podria ser el parell asturià - aranès.

La primera part del projecte es porta a terme fora dels laboratoris. Per disposar de les dades necessàries per entrenar els models d'IA, cal recopilar tot el material que sigui possible de l'asturià, l'aragonès i l'aranès /occità. En aquesta darrera llengua és on col·labora la professora del departament de Filologia i Comunicació de la UdL, Mar Font Martí [https://estudiscatalans.udl.cat/ca/pla-formatiu/professorat/detall/index.html?enc=NDc5ODMzMzY=].

El projecte *TAN-IBE: traducció automàtica neuronal per a les llengües romàniques de la península Ibèrica*, amb una durada de tres anys, compta amb finançament del Ministeri de Ciència i Innovació mitjançant el programa Projectes de generació del coneixement 2021. "Volem ajudar a fomentar l'ús de les llengües amb menys

recursos i incrementar-ne les publicacions", destaca el coordinador del projecte i professor de la UOC Antoni Oliver. "Les llengües com l'asturià, l'aragonès o l'aranès han de formar part de les tecnologies digitals. Si no, poden anar desapareixent i ser oblidades", afegeix.

Text: Comunicació UOC / Premsa UdL