DOMINO: Traducción Automática Neuronal, en DOMInio, NON supervisada

Descrición

A tradución máquina (MT) foi unha das aplicacións máis destacadas da intelixencia artificial desde o comezo do campo. A maiores do interese intrínseco dada a dificultade e completitude do problema, a tradución máquina presenta un gran interese nun mundo cada vez máis globalizado, pola súa capacidade para romper a barreira da linguaxe, á vez que constitúe unha salvagarda para a herdanza cultural e a diversidade das linguas faladas do mundo.

Aínda que en 2018 a tradución automática (TA) de calidade seguía sendo un reto para a maioría de pares de idiomas, o desenvolvemento deste campo nos últimos anos fai que este preto de ser unha realidade. A conxunción dentro de NMT (Tradución automática neuronal) da aprendizaxe profunda (Deep Learning), coa clara achega dos embeddings, e das técnicas neuronais conseguiu uns resultados que parecían impensables fai tres anos.

Doutra banda as empresas usuarias e os usuarios particulares familiarizáronse coas vantaxes e limitacións do uso desta tecnoloxía. Mentres as primeiras focalízanse en aumentar a produtividade, combinando as memorias de tradución, as ferramentas de TA e as contornas de postedición; os segundos úsana intensivamente a pesar de que en moitos casos, sobre todo para idiomas con recursos limitados, a calidade que ofrecen non é comparable á tradución profesional. Isto fai que a demanda, tanto profesional como social (axenda dixital), vaia en aumento.
Este proxecto, coordinado polo grupo IXA da UPV/ EHU, propón investigar en técnicas que melloren a estado da arte dos sistemas de TA de aprendizaxe profunda e neuronais. Conta coa colaboración da Fundación Elhuyar, e o CiTIUS.

Obxectivos

De forma máis específica, os obxectivos do proxecto son os seguintes:

- **Mellora da calidade da tradución NMT e obtención de avaliacións fiables** Hai diversas carencias, sobre todo para a fidelidade do texto xerado, que deben ser estudadas e solucionadas: segmentos sen traducir, problemas con terminoloxía, entidades nomeadas, cantidades e adxectivos. Tamén é importante mellorar os tempos de aprendizaxe e execución destes sistemas.
- **Novas achegas para tradución automática para idiomas con poucos recursos** Dentro dos resultados do proxecto TADEEP é de resaltar o alto impacto que obtivo esta liña de investigación, con publicacións nos foros máis importantes da área (ACL, EMNLP, AAAI, ICLR). Profundar nesta liña é unha das claves deste proxecto para conseguir publicacións de impacto.
- **MT adaptado a dominios específicos e transferencia á contorna empresarial** ademais da aplicación do paradigma NMT a outros problemas seq2 seq (corrección gramatical, por exemplo). É a parte máis aplicada do proxecto que se presenta pero que tenta resolver necesidades reais de contorna empresariais e sociais próximos.

INVESTIGADORES

Proxecto de
Universidad del País Vasco

Investigador principal externo
Eneko Agirre

Colaboradores do CiTIUS
Pablo Gamallo Otero

DETALLES

Data de execución:
01/01/2019 - 31/12/2021

Financiado por
Programa Estatal de I+D+i Orientada a los Retos de la Sociedad, Ministerio de Economía y Competitividad, PGC2018-102041-B-I00

PO FEDER Galicia 2014-2020 "*Unha maneira de facer Europa*"

PUBLICACIÓNS

The Impact of Linguistic Knowledge in Different Strategies to Learn Cross-Lingual Distributional Models
European Association for Artificial Intelligence, 2020

A Methodology to Measure the Diachronic Language Distance between Three Languages Based on Perplexity
Journal of Quantitative Linguistics, 2020

Cross-lingual Diachronic Distance: Application to Portuguese and Spanish
Procesamiento del Lenguaje Natural, 2019

Ver todas

PROGRAMAS CIENTÍFICOS

Tecnoloxías da Linguaxe Natural