El Confidencial

Estos ingenieros vascos han creado un traductor para hacer sombra a Google

Un equipo de investigadores de la Universidad del País Vasco ha desarrollado un sistema que aprende a traducir entre dos idiomas sin intervención humana y con resultados prometedores

Los investigadores de la UPV Eneko Agirre (a la derecha) y Mikel Artetxe (a la izquierda), coautores del estudio

Autor
Lucía Caballero
Contacta al autor
@Lulucille_
Tiempo de lectura7 min

22.01.2018 – 05:00 H.

Desde el punto de vista lingüístico, las palabras ‘gato’, ‘perro’ o ‘manzana’ son diferentes, aunque las dos primeras tengan cierta relación por su significado. Las redes neuronales artificiales, sin embargo, entienden los vocablos como vectores numéricos que capturan su contenido semántico y que conforman un mapa abstracto del lenguaje donde los términos ‘perro’ y ‘gato’ sí están más próximos que cualquiera de ellos con ‘manzana’.

Para las máquinas, el español tiene su propio tejido de palabras, distinto del inglés, el francés o el chino mandarín. Los sistemas de traducción automática actuales, como el popular Google Translate, funcionan construyendo esa representación numérica de las expresiones en el idioma de origen, para luego encontrar su equivalencia más probable en el de destino —no siempre con acierto, eso sí—.

El Traductor de Google acertará más con sus resultados gracias a la red neuronal

El Confidencial
Las traducciones tendrán menos errores gramaticales y se parecerán más a la lengua hablada de los usuarios que ejecuten la búsqueda

El problema es que estas herramientas presentan una importante limitación: solo saben descifrar aquellas lenguas para las que han sido entrenadas previamente con millones de traducciones de textos hechas por personas. Es lo que se conoce en 'machine learning' como aprendizaje supervisado, es decir, sus redes neuronales "necesitan asistencia humana para aprender", indica a Teknautas el investigador Mikel Artetxe, cuyo reciente trabajo abre la puerta a la creación de traductores universales.

Junto a otros expertos de la Universidad del País Vasco (UPV), este joven ingeniero ha desarrollado un traductor basado en aprendizaje no supervisado. "Simplemente, coge un texto en un idioma, otro texto diferente en otro idioma y el sistema aprende a traducir de uno a otro por sí mismo", explica Aretxe. Es como si una persona pudiera aprender las equivalencias entre palabras de dos lenguas distintas a partir de dos libros, sin ayuda de un diccionario. "Esto, que parece imposible para los humanos, es totalmente factible para un ordenador", asegura.

Los sistemas de traducción automática transforman las palabras y frases en vectores (Fuente: Blickpixel I Pixabay)

De momento, el sistema solo funciona para texto, pero las aplicaciones de traducción por voz constan de varios módulos independientes que transforman las palabras habladas a texto para traducirlas. "En este

sentido, nuestro sistema podría también combinarse con estos módulos adicionales para trabajar con voz, imágenes u otros", dice el ingeniero de la UPV.

Traducción en varios pasos

Los traductores automáticos tradicionales están compuestos, en general, por dos redes neuronales. La primera actúa como codificador: toma una frase, por ejemplo en inglés, y construye su correspondiente representación vectorial. La segunda recoge ese vector que contiene información sobre la oración textual y produce todas las traducciones posibles, por ejemplo al español, para escoger la que tiene una mayor probabilidad de coincidir con la correcta.

Además, muchos sistemas actuales incorporan un mecanismo de atención que añade complejidad al proceso y actúa en el segundo paso, combinando los vectores que representan las palabras fijándose en las partes más importantes.

Si se entrenan con millones de ejemplos de traducciones, estas redes neuronales aprenden poco a poco cuáles son las verdaderas correspondencias entre las palabras y frases en uno y otro idioma y acaban traduciendo bien expresiones que no han visto en los ejemplos. Lo malo es que este método lleva mucho tiempo y requiere de gran cantidad de datos que no siempre están disponibles.

Para conseguir resultados aceptables hace falta partir de millones de frases. "A veces es muy difícil encontrar tantas traducciones"

Para conseguir resultados aceptables hace falta partir, como mínimo, de millones de frases. "A veces es muy difícil encontrar tantas traducciones, sobre todo para lenguas que no están muy relacionadas como, por ejemplo, el letón y el euskera", indica Núria Bel, del Instituto de Lingüística Aplicada de la Universidad Pompeu Fabra.

La investigadora asegura que hasta hace muy poco no había un corpus suficientemente extenso y disponible de traducciones chino–español. De los cerca de 7.000 idiomas que existen en todo el mundo, Google Translate solo entiende 103. Por eso, Bel ve muy positiva la dirección tomada por el equipo de la UPV: "En cualquier sistema de procesamiento de lenguaje natural, no depender de la supervisión y mejorar las técnicas de aprendizaje y generalización intentando reducir la necesidad de datos siempre es una buena línea de investigación".

Google utiliza un sistema colectivo para aumentar el corpus de traducciones (Fuente: jonrussell I Visualhunt)

Al sistema desarrollado por Aretxe y sus compañeros le basta con que se disponga de un texto escrito suficiente en los dos idiomas. "Es muchísimo más fácil de obtener que los centenares de miles de traducciones que requiere como mínimo un traductor tradicional".

El equipo se ha basado en el esquema de los modelos tradicionales, pero han añadido ciertas modificaciones. Por un lado, en lugar de establecer una sola dirección de traducción, esta se produce en ambos sentidos a la vez. Lo consiguen porque la red neuronal que construye los vectores a partir del texto toma frases tanto del inglés como del español y en la segunda parte del proceso interviene no una, sino dos redes neuronales encargadas de traducir sendos idiomas.

De esta forma, el sistema produce representaciones abstractas bilingües que contienen información de las palabras en ambas lenguas. Es decir, teje un diccionario a partir de las equivalencias que existen entre los mapas vectoriales de los dos idiomas, porque las palabras están agrupadas de igual manera —los términos 'perro' y 'gato' estarán próximos tanto en castellano como en inglés—. Después, evalúa y corrige el resultado traduciendo repetidamente en un sentido y otro.

Una línea de investigación en ciernes

El método desarrollado en la UPV se parece mucho al diseñado por otro grupo de investigadores de Facebook y la Universidad de la Sorbona de París. Su traductor también construye una representación abstracta de las frases en ambos idiomas, pero, a diferencia del de padres españoles, añade un paso en el proceso para verificar que ese lenguaje intermedio es verdaderamente abstracto. Tanto el sistema patrio como el galo se basan en el trabajo del ingeniero de Microsoft Di He sobre aprendizaje dual.

Los dos estudios, recogidos en el repositorio virtual arXiv con un día de diferencia, están pendientes de ser revisados por pares y se presentarán en la International Conference on Learning Representations de este año. Ambos parten del mismo corpus de textos en francés e inglés y obtienen una tasa de precisión en las traducciones que ronda el 15%. Si bien no es un porcentaje tan alto como el de Google Translate (alrededor del 40%) o el de una persona (puede superar el 50%), los resultados son muy positivos tratándose de una línea de investigación tan incipiente.

Aretxe advierte que "los traductores tradicionales llevan décadas entre nosotros y esta nueva aproximación no ha hecho más que nacer". Además, es comprensible que los sistemas basados en millones de ejemplos reales sean mejores, al menos de momento, que uno que aprende por sí solo. "La ventaja del nuestro reside, precisamente, en que requiere muchos menos recursos, haciendo la tecnología mucho más accesible", dice el investigador de la UPV. Entrenándolo con unos pocos ejemplos, podrían aumentar su tasa de acierto.

Estamos aún muy lejos del nivel del droide C3PO de Star Wars

Las herramientas desarrolladas por ambos equipos suponen un "avance importante" hacia el desarrollo de un verdadero traductor universal, aunque aún es pronto para beneficiarnos de un dispositivo con el dominio de la palabra de C-3PO, el famoso droide de protocolo de 'Star Wars'. Más allá de los idiomas, estos nuevos traductores podrían ser útiles para 'traducir' los lenguajes especializados, como el científico y el médico, aunque se trata aún de un campo poco explorado en el que habría que realizar nuevas pruebas.

Bel no se atreve a predecir si el aprendizaje no supervisado es el futuro, después de vivir el enorme y abrupto salto de la traducción automática estadística a la neuronal. Lo que sí parece factible es que, a medida que progrese y madure, esta tecnología acabe integrada en aplicaciones o programas que podamos usar en móviles y ordenadores. Aunque, como Google Translate, tampoco sean perfectos.

→
Siguiente

Nueva economía, vieja desigualdad: solo 1 de cada 4 jefes en 'startups' españolas es mujer

←
Anterior