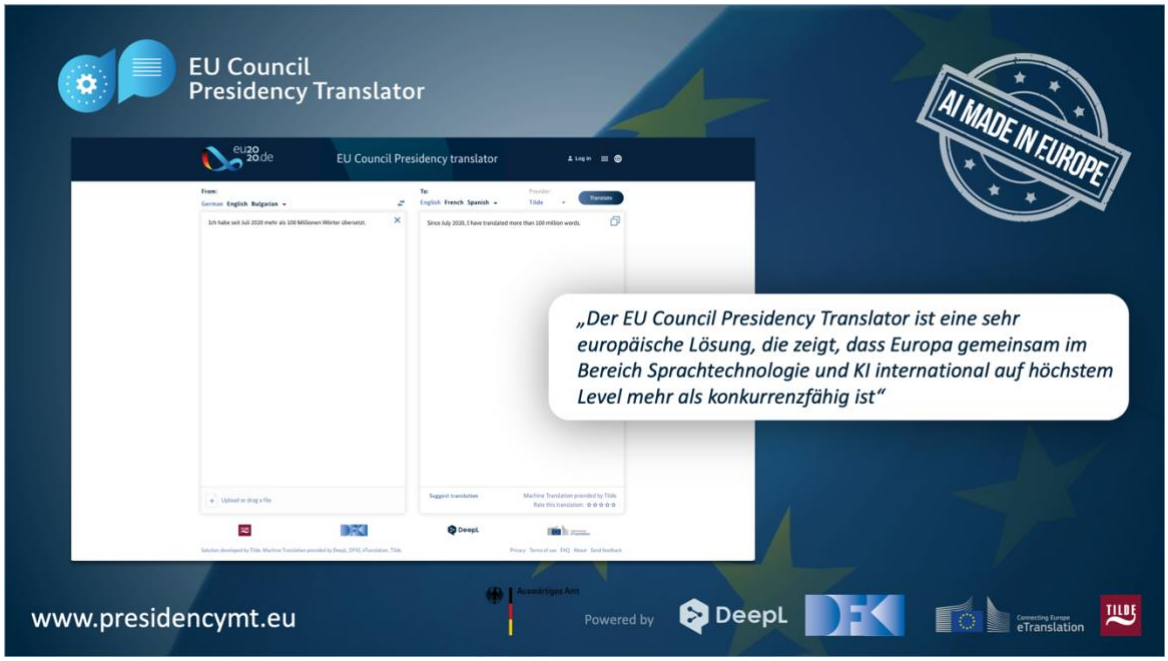
100 millones de palabras traducidas con el traductor de la Presidencia del Consejo de la UE

19.11.2020 | Idioma y comprensión de textos |
Tecnología de lenguaje y multilingüismo | Saarbrücken

Entrevista con el Prof. Dr. Josef van Genabith, director del Departamento de Investigación de Tecnologías de Lenguaje Multilingüe en el DFKI de Saarbrücken, sobre traducción mecánica y el Traductor de la Presidencia del Consejo de la UE, que se utiliza en el contexto de la presidencia alemana del Consejo de la UE desde julio de 2020.



Nota: Versión en español traducida por un sistema DFKI y ligeramente posteditada de la entrevista original alemana generada por EUCPT.

- Prof. van Genabith, usted es director científico del DFKI y desde 2014 dirige el área de investigación de tecnologías lingüísticas multilingües (MLT) en Saarbrücken. ¿Cómo fue su carrera científica antes de cambiar a Saarbrücken?

Los logros del Traductor de la Presidencia del Consejo de la UE son una buena ocasión para nuestro equipo MLT y nuestros socios en DeepL, Tilde y eTranslation! Estoy muy orgulloso de los equipos y el trabajo que han realizado en estrecha coordinación con el Ministerio Federal de Relaciones Exteriores. Yo mismo he estado muy interesado en el lenguaje y la tecnología durante mucho tiempo, estudiando ingeniería eléctrica y anglicología en la RWTH de Aachen y después tuve mucha suerte: A través de una beca del British Council y más tarde del Ministerio de Relaciones Exteriores y de la Mancomunidad de Naciones pude hacer primero un MA en la Universidad de Essex y luego doctorarme en Louisa Sadler. A principios de la década de los 90, estuve de postdoctorado en Hans Kamp, en el Instituto de procesamiento del lenguaje natural (IMS) de Stuttgart. ¡Un gran momento! Después de eso, estuve 17 años en Irlanda en la Escuela de Computación, la Universidad de Dublín City, donde pasé por toda la gama de Lecturer, Senior Lecturer y Profesor Asociado. En Dublín tenía mucha libertad y excelentes colegas de la DCU, otra de las universidades de Dublín, y de muchas empresas de alta tecnología con sede en Irlanda (IBM, Microsoft, Symantec), y pudimos aprovechar estas libertades: Reconstruí el Centro Nacional de Tecnología de la Industria (NCLT) y fui el director fundador del CNGL (Centro para la Localización de Próxima Generación, ahora ADAPT y dirigido por

***Light post-editing** implies minimal intervention by the post-editor to make the text understandable. Grammar, punctuation and spelling are correct, the translation is complete and accurate in content, but not necessarily idiomatic and fluent.

Vinny Wade). A través de estos trabajos y, en particular, del CNGL, a principios de la segunda mitad de los años 2000-2010 nos vimos cada vez más involucrados en proyectos internacionales, por ejemplo, de la UE, en los que el anterior director de nuestro laboratorio en Saarbrücken, Hans Uszkoreit, fue muy activo. A través de Hans Uszkoreit, quien mientras tanto había construido el laboratorio hermano en Berlín (hoy SLT, dirigido por Sebastian Möller), en 2014, después de 17 años en Irlanda, llegué a Saarbrücken y al DFKI.

- Además de su actividad en DFKI, también tiene una cátedra en la Universidad del Sarre. ¿Cómo se complementan los trabajos académicos y orientados a la aplicación?

Lo más importante en nuestro trabajo son los colaboradores y colaboradoras: ¡a través de ellos nuestro trabajo se convierte en un éxito! Mis colaboradores en la universidad y en el DFKI trabajan conjuntamente en equipos mezclados. En nuestras reuniones conjuntas semanales no importa si alguien está en DFKI o en la universidad. Somos parte del SFB1102 (Information Density and Linguistic Encoding) en la Universidad, tenemos un proyecto del DFG en la Universidad sobre Post-Edición multimodal, donde trabajamos con gran éxito con el Prof. Antonio Krüger también del DFKI; lidero el programa de Masters europeo en Tecnologías del Lenguaje y Comunicación (LCT, Erasmus +), liderado por una de mis colaboradoras en la dirección en el MLT-Lab (DFKI). Todos mis jefes de equipo en los cuatro grupos de MLT -Machine Translation, Question Answering and Information Extraction, Talking Robots y Data and Resources dan clases, imparten seminarios y s supervisan a estudiantes de doctorado, master y carrera. Del mismo modo, muchos empleados del MLT son activos en la universidad. Por supuesto, formalmente y financieramente todo está en proyectos separados. Pero la conexión con la universidad es muy fuerte. El Departamento de Ciencia y Tecnología del Lenguaje de la Universidad del Sarre es uno de los mejores de Europa. En el MLTLab del DFKI somos especialmente fuertes en investigación: por ejemplo, hemos publicado en 2020 más de 10 artículos en las principales conferencias internacionales (ACL, ICML, EMNLP, Coling, IJCAI) en el campo de la tecnología del lenguaje, la IA y el aprendizaje automático. Es un gran éxito y muestra la calidad del equipo. Por otro lado, la investigación orientada a aplicaciones del DFKI es una atracción para los estudiantes e investigadores de la universidad: ¿dónde, de lo contrario, se utiliza de forma abierta y segura para todos el trabajo propio, como por ejemplo pasa en con el Traductor de la Presidencia del Consejo de la UE , que permite que 100 millones de palabras se hayan traducido en un plazo de 4,5 meses (hasta la fecha)? ¡Esto es genial!

- El Traductor de la Presidencia del Consejo de la UE ha promovido aún más la visibilidad de los servicios de traducción automática en Alemania. Es una actuación conjunta de varios actores, pero usted ha dirigido este proyecto. ¿Cuándo inició usted el trabajo? ¿Cómo ha compilado el consorcio? ¿Y cuántos científicos participaron?

El traductor de la Presidencia del Consejo de la UE es una solución muy europea que demuestra que Europa es más que competitiva a nivel internacional en el ámbito de la tecnología de la lengua y la IA: Se basa en una combinación de destacados expertos en tecnología y en IA de Alemania (DeepL, DFKI), Letonia (Tilde) y la CE (eTranslation). Una asociación entre la industria (DeepL, Tilde), el sector público (CE, eTranslation) y un instituto de investigación (DFKI). El DFKI dirige el proyecto, el financiamiento proviene del Ministerio Federal de Relaciones Exteriores, que ostenta la presidencia alemana del Consejo de la UE. En este sentido, las competencias de los miembros del consorcio se complementan idealmente: Tilde ha desarrollado y gestionado durante muchos años, con apoyo europeo, el andamio básico de los Traductores de la Presidencia , en el que se integran los motores de traducción de muchos proveedores. DeepL ofrece motores de traducción de excelente calidad

para 8 idiomas. eTranslation (CE) proporciona una base de traductores automáticos para los 24 idiomas oficiales de la UE. En estrecha colaboración con los departamentos de traducción de los ministerios, DFKI ha desarrollado sistemas de traducción automática especialmente adaptados a los datos y necesidades de los ministerios para el alemán, francés y español. Tilde hace lo mismo para inglés, italiano y polaco. En DFKI, Stephan Busemann atiende administrativamente al Traductor de la Presidencia. Yo estoy a cargo de los aspectos científicos y técnicos. Cristina España i Bonet, la directora del equipo de traducción en el MLT-Lab y su colaboradora Jingyi Zhang desarrollan los sistemas con el apoyo de dos estudiantes, Damyana Gateva y Anastasija Amman, del master "Ciencia y Tecnología del Lenguaje " de la Universidad. El DFKI también dirige el trabajo de relaciones públicas y medios de comunicación del Traductor de la Presidencia. Esto es supervisado por Eileen Schnur y su colega Marlies Thönnissen en el MLT y apoyado activamente por el Departamento de Comunicación del DFKI.

- Ustedes utilizan redes neuronales artificiales para la traducción. ¿Puede describir cómo funciona su motor de traducción?

En los últimos años, los modelos neuronales han permitido saltos cuánticos en la calidad de muchas tecnologías de lenguaje y otras aplicaciones en la IA. Nuestros sistemas utilizan redes neuronales profundas basadas en “Transformers”. Estos modelos utilizan diferentes tipos de atención y son altamente paralelizables en muchas partes.

- Las redes neuronales artificiales se entrenan/testean con cantidades muy grandes de datos del idioma. ¿De dónde vienen estos datos de entrenamiento y pruebas y, solo como estimación, cuántas palabras en curso son?

Para muchos pares de idiomas, nuestros datos de entrenamiento son decenas de millones de pares de frases, cada par de frases incluye un conjunto inicial en un idioma y su traducción al otro idioma. A partir de ellos, aprenden las máquinas a traducir por sí mismas. Estos datos provienen de traducciones ya hechas por personas. Por lo tanto, la máquina aprende de los seres humanos. Los datos provienen de recopilaciones de datos de la UE, de ELRC (la Coordinación de Recursos del Lenguaje de la Unión Europea, que también gestionamos en el MLT en DFKI) y de otras fuentes. Además, trabajamos muy estrechamente con los equipos de transferencia de los ministerios para creartraductores especiales, especialmente orientados a las necesidades de los ministerios, con datos de los ministerios. Estos son evaluados constantemente por los traductores y traductoras de los ministerios, de modo que los motores puedan mejorarse continuamente a lo largo del proyecto.

- El Traductor de la Presidencia ha sido ampliamente utilizado por los usuarios y usuarias durante los últimos 150 días. Se tradujeron más de 100 millones de palabras. ¿Cuáles fueron los pares de idiomas más demandados? ¿Y quizás también hubo frases que eran particularmente frecuentes?

A diferencia de otras ofertas, el traductor de la presidencia está protegido y es seguro, todos los servidores están en la UE, las transmisiones están cifradas, y después de una traducción creada, todos los datos se borran de inmediato. Sólo tenemos información de alto nivel. Las cifras muestran que la traducción con un solo clic de la página web de la Presidencia del Consejo en lengua alemana es muy bien aceptada: ca. 47% de los 100 millones de palabras traducidas hasta la fecha vienen de aquí. Los idiomas preferidos para la traducción automática en el sitio web de la Presidencia del Consejo son el español, el italiano y el portugués (las versiones en francés e inglés fueron elaboradas

manualmente). La otra mitad resulta de traducciones de texto (22%), de documentos (30%) y de páginas web (2%) en la página web del traductor, y aquí la traducción entre alemán e inglés es la más exigida.

- ¿Qué dicen los traductores sobre la nueva calidad de la traducción automática? ¿Los traductores ven las máquinas como competidores o como herramientas que apoyan su trabajo? ¿Y cómo cambia la imagen profesional del traductor?

En el proyecto "EU Council Presidency Translator" cooperamos muy estrechamente con los colegas de los equipos de traducción de los ministerios: dirigen la recopilación y puesta a disposición de datos dentro de los ministerios para adaptar motores especiales a las necesidades de los ministerios. Además, prueban y evalúan los motores y, gracias a sus resultados, contribuyen de forma centralizada a mejorar los sistemas. En el flujo de trabajo de la traducción, los motores son entonces una herramienta: con una buena calidad de traducción, el traductor automático puede ayudar a aumentar la productividad de un traductor humano. El concepto profesional del traductor está cambiando hacia el control de calidad, el aseguramiento de la calidad mediante el reeditado (corrección) de traducciones creadas automáticamente y la certificación de traducciones y su calidad. La formación moderna de traductores tiene en cuenta estos cambios: El curso de postgrado "Translation Science and Technology" en la Universidad del Sarre tiene un alto porcentaje de tecnología en la que se familiariza a los futuros traductores y traductoras con las tecnologías del lenguaje desarrolladas por sus compañeros en los cursos de lingüística computacional (Ciencia y Tecnología del Lenguaje) e informática.

- La presidencia alemana del Consejo de la UE finaliza el 31 de diciembre de 2020. ¿Cómo se utiliza el traductor Presidency a continuación? E independientemente, ¿cuáles son sus planes posteriores?

El Traductor de la Presidencia ha sido ampliamente aceptado y ha superado todos los récords del anterior traductor. Estoy muy orgulloso de lo que el equipo MLT ha hecho en el DFKI junto con los colegas de DeepL, Tilde y eTranslation. Existe un gran interés en utilizar el traductor en otras presidencias. Se están llevando a cabo conversaciones al respecto. También existe un gran interés por parte de la industria en la tecnología del lenguaje alemana y europea: Tecnología del lenguaje e IA "made in Europe". La traducción automática es sólo una de las competencias de nuestro laboratorio MLT: Otras son las del grupo de "Question-Answering and Information Extraction" (especialmente en el ámbito biomédico), las del grupo "Talking Robots" (centrado en los sistemas de diálogo y robótica salvavidas) y las del grupo "Data and Resources" (que dirige grandes proyectos de la UE como ELRC desde hace muchos años). A ello se suma nuestro laboratorio hermano SLT (Speech and Language Technology) en Berlín. Los dos laboratorios (MLT en Saarbrücken y SLT en Berlín) cooperan estrechamente y se complementan en su experiencia.