

Caso grupal aplicaciones prácticas de PLN en investigación y en el mercado final

Procesamiento del Lenguaje Natural, UNIR 2023



realizado por:

* Nicolás felipe trujillo montero
* Jesús carlos avecilla de la herrán
* alejandro magdaleno gascón

ÍNDICE

1.- Introducción

2.- Artículos acerca de la traducción automática

2.1.- Artículo i

2.2.- Artículo ii

2.3.- Artículo iii

3.- solución comercial en la actualidad

4.- Conclusión

5.- Bibliografía

1.- INTRODUCCIÓN

En el ámbito de la traducción automática neuronal, se han desarrollado diversas tecnologías y metodologías para mejorar la calidad y eficiencia de los sistemas de traducción. En este trabajo, se exploran tres artículos científicos que abordan diferentes enfoques y aplicaciones de la traducción automática neuronal.

El primer artículo se centra en Tensor2Tensor (T2T), una biblioteca de modelos y conjuntos de datos de aprendizaje profundo diseñada para agilizar la investigación en este campo. T2T utiliza TensorFlow como base y se enfoca en el rendimiento y la facilidad de uso. El artículo describe la tecnología utilizada en el experimento, la metodología experimental, que incluye la selección de un problema específico, la elección de un modelo predefinido, la configuración de hiperparámetros y la selección de conjuntos de datos, así como los resultados obtenidos y las métricas utilizadas para evaluar el experimento.

El segundo artículo se centra en la traducción neuronal automática del euskera, un idioma con características notables que pueden representar un desafío para los enfoques convencionales debido a su particularidad tipológica desde lo lingüístico. El artículo describe la tecnología utilizada en el experimento, que incluye modelos de segmentación basados en Byte Pair Encoding (BPE) y FLATCATV2, así como la recopilación de corpus paralelos en euskera y español. Además, se presentan los resultados obtenidos y las métricas utilizadas para validar el experimento en el contexto específico de la traducción del euskera.

El tercer artículo aborda la traducción automática no supervisada para idiomas con recursos limitados. Se describen las tecnologías utilizadas, incluyendo sistemas de traducción automática neuronal y enfoques no supervisados. La metodología experimental implica el uso de datos monolingües y paralelos de diferentes fuentes, como el repositorio WMT. Los resultados obtenidos se evalúan utilizando métricas de puntuación BLEU y se discute su relevancia para el ámbito de aplicación de la traducción automática en lenguas no tan documentadas.

En conjunto, estos artículos proporcionan una visión amplia de las tecnologías, metodologías y resultados actuales en el campo de la traducción automática neuronal, abarcando desde la mejora de la eficiencia y accesibilidad de los modelos hasta la aplicación de estos enfoques a idiomas en concreto que no tienen tanto impacto internacional

2.- ARTÍCULOS ACERCA DE LA TRADUCCIÓN AUTOMÁTICA

2.1.- ARTÍCULO I

**Tensor2Tensor for Neural Machine Translation**

Describir la tecnología que se ha utilizado en el experimento descrito.

Tensor2Tensor (T2T) es una librería de modelos y conjuntos de datos de aprendizaje profundo diseñada para hacer que la investigación en aprendizaje profundo sea más rápida y accesible. T2T utiliza TensorFlow (Abadi et al., 2016) en su totalidad y se centra en el rendimiento y la facilidad de uso. A través de su uso de TensorFlow y varias abstracciones específicas de T2T, los investigadores pueden entrenar modelos en CPU, GPU (individuales o múltiples) y TPU, tanto localmente como en la nube, por lo general sin necesidad de escribir código específico del dispositivo o realizar configuraciones complicadas.

Describir la metodología experimental. ¿En qué consistió el experimento?, ¿Qué corpus se utilizó?

El desarrollo comenzó enfocado en la traducción neuronal automática, por lo que Tensor2Tensor incluye muchos de los modelos de traducción neuronal automática más exitosos y conjuntos de datos estándar.

Existen cinco componentes clave que especifican una ejecución de entrenamiento en Tensor2Tensor:

Problem: El problema es la tarea específica que se desea abordar, como la traducción automática, el reconocimiento de voz, la generación de texto, entre otros. En Tensor2Tensor, se selecciona un problema específico para entrenar un modelo.

Model: El modelo se refiere a la arquitectura de aprendizaje profundo utilizada para abordar el problema. Tensor2Tensor proporciona una amplia gama de modelos predefinidos para diferentes tareas, como Transformer, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), entre otros.

HParams: Los HParams (parámetros hiper) son los hiperparámetros del modelo, que determinan aspectos como el tamaño de los lotes (batch size), la tasa de aprendizaje, el número de capas, las dimensiones de las representaciones ocultas, entre otros. Estos parámetros se configuran según las necesidades del problema y el modelo seleccionado.

Data: Los datos se refieren al conjunto de entrenamiento y validación utilizados para entrenar el modelo. Tensor2Tensor proporciona conjuntos de datos estándar para tareas comunes, como WMT, COCO, CIFAR, Librispeech, entre otros. También es posible utilizar conjuntos de datos personalizados.

Tabla

Descripción generada automáticamenteTraining Schedule: El horario de entrenamiento define los detalles del proceso de entrenamiento, como el número de iteraciones, la programación del aprendizaje (learning rate schedule), la política de regularización, entre otros. Estos parámetros determinan cómo se ajusta el modelo a lo largo del tiempo durante el entrenamiento

Describir los resultados obtenidos: ¿qué métricas se utilizaron para validar el experimento?, ¿qué resultados se obtuvieron? Valora si dichos resultados son adecuados para el ámbito de aplicación.

Entre las métricas comunes se incluyen la puntuación BLEU (Bilingual Evaluation Understudy) y la exactitud de las traducciones generadas por los modelos entrenados. Estas métricas se aplicaron para evaluar la calidad de las traducciones automáticas obtenidas utilizando diferentes problemas y modelos de T2T.

BLEU compara una traducción automática con una o varias traducciones de referencia humanas y calcula la similitud entre ellas. Se basa en la coincidencia de palabras o secuencias de palabras entre la traducción automática y las referencias. La puntuación BLEU se expresa como un valor entre 0 y 1, donde 1 indica una coincidencia perfecta con las referencias humanas. Cuanto mayor sea la puntuación BLEU, se considera que la traducción automática es de mayor calidad. Es importante tener en cuenta que BLEU es una métrica basada en n-gramas, lo que significa que compara secuencias de palabras de diferentes longitudes. Sin embargo, BLEU no captura aspectos semánticos o de coherencia más amplios de las traducciones, por lo que se deben considerar otras métricas y evaluaciones humanas para obtener una imagen más completa de la calidad de las traducciones automáticas.

Los resultados obtenidos demostraron mejoras significativas en la calidad de las traducciones en comparación con otros enfoques existentes. T2T logró rendimientos competitivos en varios conjuntos de datos estándar para tareas de traducción automática, como el conjunto de datos de traducción automática del WMT logrando puntuaciones de BLEU altas con una mayor precisión en las traducciones generadas.

2.2.- ARTÍCULO II

**Neural Machine Translation of Basque**

Describir la tecnología que se ha utilizado en el experimento descrito.

En este artículo se explora la aplicabilidad de la traducción neuronal automática al euskera, un idioma con características notables que pueden representar un desafío para los enfoques de codificador-decodificador con mecanismos de atención.

Describir la metodología experimental. ¿En qué consistió el experimento?, ¿Qué corpus se utilizó?

En este artículo, tomamos conjuntos de BPE tanto en datos en euskera como en español para mejorar la consistencia entre la segmentación de origen y destino. El Byte Pair Encoding (BPE) es un algoritmo de compresión que fue adaptado para la segmentación de palabras en la traducción neuronal automática por (Sennrich et al., 2016). Reemplaza iterativamente el par de caracteres más frecuente en una secuencia con un símbolo no utilizado, sin tener en cuenta los pares de caracteres que atraviesan los límites de las palabras. BPE permite representar un vocabulario abierto mediante un vocabulario de tamaño fijo de secuencias de caracteres de longitud variable, lo que tiene la ventaja de producir secuencias de símbolos aún interpretables como unidades de palabras.

FLATCATV2 es un sistema basado en MORFESSOR que se desarrolló para implementar una reducción de vocabulario motivada lingüísticamente para la traducción neuronal automática y que se propuso originalmente para el turco (Ataman et al., 2017). El proceso de segmentación consta de dos pasos. Primero, se utiliza MORFESSOR para inferir la morfología del idioma considerado de manera no supervisada, basándose en un corpus monolingüe sin etiquetar. Luego, las segmentaciones morfológicas aprendidas se ajustan a un vocabulario de tamaño fijo, que en nuestro caso fue de 45,000 tokens.

A diferencia del método de aprendizaje conjunto que seleccionamos para la segmentación BPE, la segmentación de FLATCATV2 se aprendió por separado en los datos monolingües, ya que esta técnica está diseñada para extraer una segmentación lingüísticamente sólida del texto.

Para construir modelos de traducción representativos para el par de idiomas euskera-español, se recopilaron y prepararon corpora paralelos de tres fuentes diferentes: traducciones profesionales en diferentes dominios, páginas web bilingües y datos comparables en el ámbito de noticias.

Describir los resultados obtenidos: ¿qué métricas se utilizaron para validar el experimento?, ¿qué resultados se obtuvieron? Valora si dichos resultados son adecuados para el ámbito de aplicación.

Tabla

Descripción generada automáticamente

2.3.- ARTÍCULO III

**Harnessing Multilinguality in Unsupervised Machine Translation for Rare Languages**

Describir la tecnología que se ha utilizado en el experimento descrito.

Los sistemas de traducción automática neuronal (Kalchbrenner y Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015; Wu et al., 2016) han demostrado resultados de vanguardia para un conjunto diverso de pares de idiomas cuando se les proporciona grandes cantidades de datos paralelos relevantes. Sin embargo, dado lo prohibitivo de este requisito para pares de idiomas con pocos recursos, ha surgido un creciente interés en la traducción automática no supervisada (Ravi y Knight, 2011) y su contraparte neuronal, la traducción automática neuronal no supervisada (UNMT) (Lample et al., 2018a; Artetxe et al., 2018), que aprovechan únicamente corpora de origen y destino monolingües para el aprendizaje. Los sistemas bilingües no supervisados (Lample y Conneau, 2019; Artetxe et al., 2019; Ren et al., 2019; Li et al., 2020a) han logrado resultados sorprendentemente sólidos en pares de idiomas de alta disponibilidad de recursos como inglés-francés e inglés-alemán.

Describir la metodología experimental. ¿En qué consistió el experimento?, ¿Qué corpus se utilizó?

Los trabajos existentes sobre traducción no supervisada multilingüe (Liu et al., 2020; García et al., 2020; Li et al., 2020b; Bai et al., 2020), que también utilizan datos paralelos auxiliares, emplean un esquema de entrenamiento de dos etapas que consiste en pre-entrenamiento con objetivos de reconstrucción ruidosos y ajuste fino con términos de retrotraducción y traducción cruzada en tiempo real (iterativa) (§4). Mostramos que esto conduce a un rendimiento subóptimo para pares de baja disponibilidad de recursos y proponemos una etapa de entrenamiento intermedia adicional en nuestro enfoque.

Obtenemos la mayor parte de nuestros datos de WMT. Los datos monolingües provienen de News Crawl6 cuando están disponibles. Para todas las parejas no supervisadas, excepto el turco, complementamos los conjuntos de datos de News Crawl con datos monolingües de Common Crawl y Wikipedia7. Los datos paralelos que utilizamos provienen de diversas fuentes, todas disponibles a través de WMT. Obtuvimos nuestros datos paralelos en inglés-hindi de IITB (Kunchukuttan et al., 2017); datos paralelos en inglés-ruso, inglés-árabe e inglés-chino del Corpus de la ONU (Ziemski et al., 2016); y datos paralelos en inglés-tamil e inglés-telugu de Wikimatrix (Schwenk et al., 2019). Utilizamos los scripts de Moses (Koehn, 2009) para normalizar la puntuación, eliminar caracteres no imprimibles y reemplazar los caracteres Unicode con sus equivalentes no Unicode. Además, utilizamos el script de normalización de Indic NLP (Kunchukuttan, 2020) para gujarati, nepalí, telugu y sinhala.

Describir los resultados obtenidos: ¿qué métricas se utilizaron para validar el experimento?, ¿qué resultados se obtuvieron? Valora si dichos resultados son adecuados para el ámbito de aplicación.

Evaluamos el rendimiento de nuestros modelos utilizando puntuaciones BLEU (Papineni et al., 2002). Se sabe que las puntuaciones BLEU dependen del preprocesamiento de los datos (Post, 2018), por lo que se requiere cuidado adecuado para asegurar que las puntuaciones entre nuestros modelos y las líneas de base sean comparables. Por lo tanto, solo consideramos líneas de base que informen puntuaciones BLEU no-tokenizadas con sacreBLEU (Post, 2018) o informen pasos de preprocesamiento explícitos.

Tabla

Descripción generada automáticamente

3.- Solución Comercial en la Actualidad

A lo largo de la evolución del campo de la traducción automática se han ido desarrollando nuevas técnicas que han mejorado las MT (Machine Translation). Dichos algoritmos han pasado de realizar traducciones literales palabra a palabra de un corpus a realizar dichas traducciones incorporando un contexto asociado al corpus (introducido por el usuario) para mejorar la forma de relacionar los hechos y de redactar, ya que no es lo mismo escribir un mensaje de texto a un familiar que escribir un artículo para la prensa.

Una de las soluciones comerciales (en este caso, un servicio web) más importantes en la actualidad es DeepL (DeepL SE, 2023). DeepL es un servicio de traducción automática neural propuesto por la empresa del mismo nombre DeepL SE en agosto de 2017 (dicha empresa nació de un proyecto de la empresa Linguee GmbH en el 2016 (DeepL SE, 2023))

Su principal producto es el servicio online de traducción gratuita localizada en su página web (DeepL SE, 2023), pero posee numerosos tecnologías relacionadas con la traducción automática en las licencias DeepL para empresas y DeepL Pro, enfocadas tanto a particulares como a empresas que sean de ámbito internacional que usualmente necesiten traducir documentos a muchos idiomas (DeepL posee 31 idiomas para traducir y la capacidad de automatizar traducciones de ficheros), por lo que es muy buena opción para dicho público objetivo.

Interfaz de usuario gráfica

Descripción generada automáticamente

Modelo de negocio desglosado

La tecnología que hay detrás de DeepL es el uso de Redes Neuronales entrenadas con algoritmos supervisados gracias al uso de varios millones de textos traducidos y de modificaciones directas de la arquitectura de los Transformers.

4.- CONCLUSIONES

En conclusión, los tres artículos abordan avances significativos en el campo de la traducción automática neuronal:

Tensor2Tensor (T2T) proporciona una plataforma eficiente y accesible para la investigación en aprendizaje profundo, lo que facilita el desarrollo de modelos de traducción automática más rápidos y precisos. Esto implica un avance importante en la mejora de la eficiencia y la accesibilidad de los sistemas de traducción automática. Por otro lado, la aplicación de la traducción neuronal automática a idiomas con características lingüísticas específicas, como el euskera, demuestra resultados prometedores. Estos enfoques permiten abordar los desafíos particulares de estos idiomas y mejorar la calidad de las traducciones automáticas.

Además, el enfoque de la traducción automática no supervisada para idiomas con recursos limitados presenta una alternativa prometedora al uso de datos paralelos. Esto abre nuevas posibilidades para la traducción automática en contextos donde la disponibilidad de datos paralelos es limitada, lo que tiene un impacto significativo en la accesibilidad y la efectividad de los sistemas de traducción automática.

En general, estos avances en la traducción automática neuronal ofrecen nuevas perspectivas y oportunidades para mejorar la comunicación interlingüística en diversos ámbitos, desde la traducción profesional hasta la accesibilidad de contenidos en diferentes idiomas. Sin embargo, aún existen desafíos por superar, como la calidad y la adaptabilidad de los modelos, así como la necesidad de explorar enfoques más eficientes y efectivos en diferentes contextos lingüísticos.

5.- Bibliografía

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., Sepassi, R., Shazeer, N., & Uszkoreit, J. (2018). Tensor2Tensor for Neural Machine Translation.

Garcia, X., Siddhant, A., Firat, O., & Parikh, A. P. (2020). Harnessing multilinguality in unsupervised machine translation for rare languages.

Antonio Pérez-Ortiz, J., Sánchez-Martínez, F., Esplà-Gomis, M., Popović, M., Rico, C., Martins, A., Van Den Bogaert, J., & Forcada, M. L. (s/f). Proceedings of the 21st annual conference of the European association for machine translation

DeepL SE. (17 de Junio de 2023). Obtenido de El mejor traductor del mundo - DeepL Translate: https://www.deepl.com/es/translator

DeepL SE. (17 de Junio de 2023). Obtenido de Perfil corporativo de DeepL: https://static.deepl.com/files/press/companyProfile\_ES.pdf