MINTZAI: Sistemas de Aprendizaje Profundo E2E para Traducción Automática del Habla

MINTZAI: End-to-end Deep Learning for Speech Translation

Thierry Etchegoyhen[1], Haritz Arzelus[1], Harritxu Gete[1], Aitor Alvarez[1], Inma Hernaez[2], Eva Navas[2], Ander González-Docasal[1], Jaime Osácar[1], Edson Benites[1], Igor Ellakuria[3], Eusebi Calonge[4], Maite Martin[4]
[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
{tetchegoyhen,harzelus,hgete,aalvarez,agonzalezd,josacar,eebenites}@vicomtech.org
[2]HiTZ Center - Aholab, University of the Basque Country (UPV/EHU)
{eva.navas,inma.hernaez}@ehu.eus
[3]ISEA - iellakuria@isea.eus
[4]Ametzagaiña - ecalonge@ametza.com, maite@adur.com

Resumen: La traducción automática del habla consiste en traducir el habla de un idioma origen en texto o habla de un idioma destino. Sistemas de este tipo tienen múltiples aplicaciones y son de especial interés en comunidades multilingües como la Unión Europea. El enfoque estándar en el ámbito se basa en componentes principales distintos que encadenan el reconocimiento del habla, la traducción automática, y la síntesis del habla. Con los avances obtenidos mediante redes neuronales artificiales y aprendizaje profundo, la posibilidad de desarrollar sistemas de traducción del habla extremo a extremo (end-to-end), sin descomposición en etapas intermedias, está dando lugar a una fuerte actividad en investigación y desarrollo. En este artículo, se hace un repaso del estado del arte en este área y se presenta el proyecto MINTZAI, que se está realizando en el ámbito.
Palabras clave: Traducción del Habla, Traducción Automática, Reconocimiento del Habla, Síntesis del Habla, Aprendizaje Profundo

Abstract: Speech Translation consists in translating speech in one language into text or speech in a different language. These systems have numerous applications, particularly in multilingual communities such as the European Union. The standard approach in the field involves the chaining of separate components for speech recognition, machine translation and speech synthesis. With the advances made possible by artificial neural networks and Deep Learning, training end-to-end speech translation systems has given rise to intense research and development activities in recent times. In this paper, we review the state of the art and describe project MINTZAI, which is being carried out in this field.
Keywords: Speech Translation, Machine Translation, Speech Recognition, Text to Speech, Deep Learning

1 Participantes y entidades financiadoras

MINTZAI es un proyecto de investigación subvencionado por el Gobierno Vasco a través de la convocatoria de ayudas ELKARTEK 2019 de la Agencia Vasca de desarrollo empresarial Spri.[1] Su principal objetivo es la investigación y el desarrollo de sistemas de traducción automática neuronal del habla, con particular énfasis en la traducción entre euskera y castellano.

El proyecto tiene una duración total de 21 meses, con comienzo el 1 de abril de 2019 y finalización el 31 de diciembre de 2020.

MINTZAI se está llevando a cabo por el siguiente consorcio: Vicomtech[2], grupo Aholab de la UPV/EHU[3], ISEA[4] y Ametzagaina[5], siendo empresas adheridas Argia, EiTB y MondragonLingua. El proyecto tiene asignado el código KK-2019/00065 y el sitio web asociado es: http://mintzai.eus/

[1]http://www.spri.eus

[2]https://www.vicomtech.org
[3]https://aholab.ehu.eus/
[4]https://www.isea.eus/
[5]https://www.ametza.com

2 Contexto y motivación

Los métodos de aprendizaje profundo (Deep Learning) se han impuesto como el nuevo paradigma en el campo de las tecnologías de la lengua, con mejoras significativas logradas por ejemplo en traducción automática, conversión de texto en habla, y reconocimiento automático del habla. Actualmente, tanto la investigación científica como la explotación comercial en estos ámbitos se basan mayoritariamente en variantes de redes neuronales artificiales profundas.

Una de las aportaciones importantes del enfoque basado en redes neuronales es la posibilidad de diseñar y entrenar sistemas extremo a extremo (E2E), i.e., sistemas que convierten información de entrada en información de salida mediante un sistema de aprendizaje neuronal único. Los sistemas de traducción automática neuronal, por ejemplo, modelan así de forma conjunta los procesos de alineamiento entre palabras y de traducción (Bahdanau y otros, 2015); los sistemas E2E de reconocimiento del habla, a su vez, aprenden a asociar directamente señales de sonido con transcripciones para modelar el proceso completo de reconocimiento (Graves y otros, 2013), y los sistemas E2E de conversión de texto a habla generan la señal partiendo de una representación fonética u ortográfica de la entrada (Wang y otros, 2017).

Se puede contrastar este tipo de arquitectura con su alternativa estándar, donde distintos componentes se encadenan para convertir información de entrada en información de salida. En un sistema de traducción habla-habla estándar, por ejemplo, destacarían un módulo de reconocimiento del habla, cuya salida en forma de texto sería tratada por un módulo de traducción automática que produzca un texto en el idioma de destino, a partir del cual se pueda generar un contenido leído mediante una voz sintética.

Las diferencias de arquitectura se trasladan en diferencias importantes en cuanto a ventajas y deficiencias respectivas; a continuación se describen las principales diferencias, centradas en las ventajas obtenibles con sistemas E2E:

- *Reducción de errores*: Los sistemas en cadena tienen como desventaja la acumulación de errores generados por los distintos módulos de la cadena. Estos errores se propagan en la cadena de procesamiento global debido a la independencia de los módulos de tratamiento de los distintos aspectos (reconocimiento, traducción o síntesis). Un error de transcripción de la señal de voz de entrada, por ejemplo, generaría así errores de traducción automática por la simple presencia del error en la transcripción inicial. Los sistemas E2E, en comparación, no sufren de esta limitación y eliminan por lo tanto esta clase de errores.
- *Optimización de modelado*: Los sistemas E2E modelan la transformación de la información de entrada en información de salida mediante una red neuronal única. Esta característica permite modelar de forma conjunta los diferentes aspectos necesarios para obtener una solución óptima al problema considerado por la red. En ámbitos ajenos, como la traducción automática, se obtienen resultados significativamente mejores con modelos de este tipo; en contraste, los sistemas en cadena delegan la determinación de las representaciones óptimas a cada módulo de forma independiente, lo cual puede resultar en un modelado subóptimo del problema global a solucionar.
- *Desarrollo y despliegue*: Al ofrecer una arquitectura reducida a una única red neuronal, los sistemas E2E permiten una simplificación significativa de la preparación de sistemas para distintos idiomas y dominios. En comparación, los sistemas en cadena requieren entrenamientos separados de módulos complejos en una primera fase, y un encadenamiento de los distintos módulos, mediante entradas y salidas adecuadas, que requiere un esfuerzo específico de adaptación y desarrollo. Los sistemas E2E permiten así el despliegue ágil que requieren los numerosos escenarios distintos que ocurren en la práctica.
- *Eficiencia*: Por la simplificación de arquitectura, los sistemas E2E pueden ofrecer una mayor eficiencia computacional, en términos de espacio y tiempos de procesamiento. Aunque sea posible en teoría desarrollar redes para sistemas E2E similares en complejidad a la suma de los componentes de sistemas encadenados, no suele ser el caso en la práctica y la eliminación de los componentes

intermedios suele proveer una mejora a nivel de eficiencia.

Estas ventajas potenciales de los sistemas E2E han impulsado su presencia cada vez mayor en aplicaciones de tecnologías del lenguaje. Los principales sistemas comerciales en reconocimiento del habla, así como alguno de los sistemas de conversión de texto en habla, son actualmente de tipo E2E, como pueden serlo también los sistemas de traducción automática adoptados en los ámbitos académicos y comerciales.

En el campo de la traducción del habla se han desarrollado muestras del potencial de la tecnología, como se describe en más detalle en la siguiente sección. En ciertas condiciones, los sistemas E2E iniciales pueden lograr resultados superiores a los obtenidos con sistemas estándar en cadena, y se considera una tecnología clave para el desarrollo de sistemas de traducción automática habla-habla y habla-texto, cuyas aplicaciones son múltiples y en alta demanda. En comunidades multilingües como la Comunidad Autónoma Vasca o la Unión Europea, por ejemplo, las necesidades de comunicación multilingüe se extienden a toda la red socioeconómica, con una presencia impactante de barreras lingüísticas que frenan la presencia cultural, la igualdad de idiomas y los desarrollos económicos.

Pese a resultados preliminares de cierto éxito con sistemas E2E de traducción del habla, los sistemas en cadena suelen obtener actualmente mejores resultados que los sistemas E2E en cuanto a calidad de traducción habla-texto en la mayoría de los casos. Por lo cual, existe un reto importante en investigación y desarrollo de arquitecturas E2E que permitan alcanzar o superar de forma consistente los sistemas en cadena clásicos. Por otro lado, mientras los sistemas de traducción habla-texto dominan el campo de la traducción del habla, la traducción habla-habla neuronal directa constituye un campo de investigación poco explorado, con retos propios importantes. Por último, para ciertos pares de idiomas, la escasez de recursos lingüísticos y componentes tecnológicos son obstáculos significativos para el desarrollo de tecnología de traducción del habla de calidad.

El proyecto MINTZAI se propone responder a estos retos, con la investigación y el desarrollo de métodos avanzados en traducción neuronal del habla, y su validación en casos de uso centrados en el par de idiomas euskera-castellano.

3 Estado del arte

Los sistemas estándares de traducción automática del habla se basan en tres componentes principales encadenados: un sistema de reconocimiento del habla, un sistema de traducción automática, y un sistema de síntesis del habla. Estos tres componentes principales se entrenan por separado, y el procesamiento opera en cascada: la salida del reconocedor de habla alimenta el sistema de traducción automática, el cual produce una traducción en forma de texto que sirve de entrada al sistema de síntesis del habla.

Para optimizar el funcionamiento de los sistemas encadenados, la comunicación entre componentes se suele adaptar a la tarea, en particular explotando hipótesis múltiples (Ney, 1999; Matusov y otros, 2005). Otros enfoques clásicos se han centrado en métodos estadísticos y en autómatas de estados finitos, integrando los modelos acústicos y de traducción en transductores estocásticos (Vidal, 1997; Casacuberta y otros, 2004).

Como fue mencionado previamente, los sistemas encadenados sufren de la acumulación de errores y los avances a este nivel han consistido en mejorar los componentes individuales, como el reconocedor de habla, o en mejorar la conectividad entre componentes mediante rasgos específicos a la frontera entre componentes (Kumar y otros, 2015).

Para resolver el problema de la propagación de errores y mejorar la calidad general de los sistemas, en trabajos recientes se ha explorado el enfoque extremo a extremo neuronal para la traducción habla-texto. Los primeros resultados obtenidos han sido prometedores con sistemas codificador-decodificador y mecanismos de atención, en particular en cuanto a la reducción de errores acumulados (Duong y otros, 2016; Bérard y otros, 2016; Weiss y otros, 2017). En cuanto a la traducción habla-habla, los trabajos son escasos, con primeros trabajos exploratorios (Jia y otros, 2019).

Pese a estos primeros resultados, donde sistemas E2E pueden superar a los sistemas encadenados en condiciones similares, en la mayoría de los casos los sistemas encadenados siguen obteniendo los mejores resultados, como muestran por ejemplo los resultados de las tareas compartidas internacionales en tra-

ducción del habla (Niehues y otros, 2019). Una de las principales razones es la escasez de datos de entrenamiento paralelos para la tarea de traducción del habla, en comparación con los datos paralelos existentes para el entrenamiento de los componentes individuales como la traducción automática. Otro factor relevante es la necesidad de mejorar las arquitecturas E2E para la traducción del habla en general.

4 MINTZAI

El proyecto MINTZAI se propone contribuir a los avances en la investigación de sistemas E2E tanto para la traducción automática habla-texto como para la traducción habla-habla. El proyecto pretende además avanzar en el estado del arte para la traducción del habla en el par de idiomas euskera-castellano, en las dos direcciones de traducción.

Tras su puesta en marcha en 2019, el proyecto ha logrado los primeros resultados resumidos a continuación:

- Creación de corpus paralelos habla-texto y habla-habla euskera-castellano y castellano-euskera a partir de los contenidos de audio, transcripciones y traducciones de las sesiones del Parlamento Vasco. Los corpus se compartirán con la comunidad científica bajo licencia Creative Commons CC-BY-NC.
- Investigación y desarrollo de sistemas de traducción del habla encadenados en euskera y castellano, con componentes neuronales E2E para reconocimiento del habla, traducción automática y síntesis del habla.
- Investigación y desarrollo de sistemas E2E para traducción habla-texto y habla-habla en euskera y castellano.

Los primeros resultados del proyecto son satisfactorios, en particular con la creación de un corpus relativamente amplio para la traducción del habla en un par de idiomas con bajo soporte a nivel de recursos, y de primeros sistemas encadenados y E2E de traducción del habla para este par de idiomas.

Durante el 2020, el esfuerzo se centrará en extender y mejorar los primeros sistemas desarrollados, y en validar los resultados obtenidos. La convocatoria en la que se enmarca el proyecto apoya a proyectos de investigación con alto potencial industrial y se validará el potencial de los sistemas desarrollados en entornos profesionales.

Bibliografía

Bahdanau, D. et al. 2015. Neural machine translation by jointly learning to align and translate. En *Proc. of ICLR*.

Bérard, A. et al. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. En *Proc. of NIPS*.

Casacuberta, F. et al. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Comput. Speech Lang.*, 18(1):25–47.

Duong, L. et al. 2016. An attentional model for speech translation without transcription. En *Proc. of NAACL*, páginas 949–959.

Graves, A. et al. 2013. Speech recognition with deep recurrent neural networks. En *Proc. of ICASSP*, páginas 6645–6649.

Jia, Y. et al. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv:1904.06037*.

Kumar, G. et al. 2015. A coarse-grained model for optimal coupling of ASR and SMT systems for speech translation. En *Proc. of EMNLP*, páginas 1902–1907.

Matusov, E. et al. 2005. On the integration of speech recognition and statistical machine translation. En *Proc. of Eurospeech 2005*.

Ney, H. 1999. Speech translation: Coupling of recognition and translation. En *Proc. of ICASSP 1999*, páginas 517–520.

Niehues, J. et al. 2019. The IWSLT 2019 Evaluation Campaign. En *Proc. of IWSLT*.

Vidal, E. 1997. Finite-state speech-to-speech translation. En *Proc. of ICASSP*, páginas 111–114.

Wang, Y. et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv:1703.10135*.

Weiss, R. J. et al. 2017. Sequence-to-sequence models can directly translate foreign speech. *arXiv:1703.08581*.