Abstract / Resum / Resumen / Résumé

In the last decade, automatic speech recognition (ASR) and machine translation (MT) have improved enormously through the use of constantly evolving deep neural network (DNN) models. If at the beginning of the 2010s the then pre-DNN ASR and MT systems were ready to tackle with success some real-life applications such as offline video lecture transcription and translation, now in the 2020s much more challenging applications are within grasp, such as live broadcast media subtitling.

At the same time in this period, media accessibility for everyone, including deaf and hard-of-hearing people, is being given more and more importance. ASR and MT, in their current state, are powerful tools to increase the coverage of accessibility measures such as subtitles, transcriptions and translations, also as a way of providing multilingual access to all types of content.

In this PhD thesis, we present research results on automatic speech recognition and machine translation based on deep neural networks in three very active domains: open educational resources, parliamentary contents and broadcast media.

Regarding open educational resources (OER), we first present work on the evaluation and post-editing of ASR and MT with intelligent interaction approaches, as carried out in the framework of EU project transLectures: Transcription and Translation of Video Lectures. The results obtained confirm that the intelligent interaction approach can make post-editing automatic transcriptions and translations even more cost-effective. Then, in the context of subsequent EU project X5gon, we present research on developing DNN-based neural MT systems, and making the most of larger MT corpora through automatic data filtering. This work resulted in a first-rank classification in an international evaluation campaign on MT, and we show how these new NMT systems improved the quality of multilingual subtitles in real OER scenarios.

In the also growing domain of language technologies for parliamentary contents, we describe research on speech data curation techniques for streaming ASR in the context of European Parliament debates. This research resulted in the release of Europarl-ASR, a new, large speech corpus for streaming ASR system training and evaluation, as well as for the benchmarking of speech data curation techniques.

Finally, we present work in a domain on the edge of the state of the art for ASR and MT: the live subtitling of broadcast media, in the context of the 2020–2023 R&D collaboration agreement between the Valencian public broadcaster À Punt and the Universitat Politècnica de València for real-time computer assisted subtitling of media contents. This research has resulted in the deployment of high-quality, low-latency, real-time streaming ASR systems for a less-spoken language (Catalan) and a widely spoken language (Spanish) in a real broadcast use case.

Resum

En l'última dècada, el reconeixement automàtic de la parla (RAP) i la traducció automàtica (TA) han millorat enormement mitjançant l'ús de models de xarxes neuronals profundes (XNP) en constant evolució. Si a principis dels 2010 els sistemes de RAP i TA previs a les XNP van arribar a afrontar amb èxit algunes aplicacions reals com la transcripció i traducció de vídeos docents pregravats, ara en els 2020 són abordables aplicacions que suposen un repte molt major, com la subtitulació de retransmissions audiovisuals en directe.

En aquest mateix període, s'estan invertint cada vegada majors esforços en l'accessibilitat als mitjans audiovisuals per a tots, incloses les persones sordes. El RAP i la TA, en el seu estat actual, són grans eines per a incrementar la disponibilitat de mesures d'accessibilitat com subtítols, transcripcions i traduccions, també com una manera de proporcionar accés multilingüe a tota classe de continguts.

En aquesta tesi doctoral presentem resultats d'investigació sobre RAP i TA basades en XNP en tres camps molt actius: els recursos educatius oberts, els continguts parlamentaris i els mitjans audiovisuals.

En l'àrea dels recursos educatius oberts (REO), presentem primerament treballs sobre l'avaluació i postedició de RAP i TA amb mètodes d'interacció intel·ligent, en el marc del projecte d'investigació europeu "transLectures: Transcripció i traducció de vídeos docents". Els resultats obtinguts confirmen que la interacció intel·ligent pot reduir encara més l'esforç de postedició de transcripcions i traduccions automàtiques. Seguidament, en el context del posterior projecte europeu X5gon, presentem una investigació sobre el desenvolupament de sistemes de TA neuronal basats en XNP, i sobre traure el màxim partit de corpus de TA massius mitjançant filtratge automàtic de dades. Aquest treball va donar com a resultat sistemes de TA neuronal classificats entre els millors en una competició internacional de TA, i mostrem com aquests nous sistemes milloren la qualitat dels subtítols multilingües en casos reals de REO.

En l'àmbit també en creixement de les tecnologies del llenguatge per a continguts parlamentaris, descrivim una investigació sobre tècniques de filtratge de dades de parla per al RAP en temps real en el context de debats del Parlament Europeu. Aquesta investigació va permetre la publicació d'Europarl-ASR, un corpus de parla nou i extens per a l'entrenament i l'avaluació de sistemes de RAP en continu, així com per a l'avaluació comparativa de tècniques de filtratge de dades de parla.

Finalment, presentem un treball en un àmbit en l'avantguarda tecnològica del RAP i de la TA: la subtitulació de retransmissions audiovisuals en directe, en el context del Conveni de col·laboració R+D+i 2020–2023 entre la radiotelevisió pública valenciana À Punt i la Universitat Politècnica de València per a la subtitulació assistida per ordinador de continguts audiovisuals en temps real. Aquesta investigació ha donat com a resultat la implantació de sistemes de RAP en temps real, amb alta precisió i baixa latència, per a una llengua no majoritària en el món (el català) i una de les llengües més parlades del món (el castellà) en un mitjà audiovisual real.

Resumen

En la última década, el reconocimiento automático del habla (RAH) y la traducción automática (TA) han mejorado enormemente mediante el uso de modelos de redes neuronales profundas (RNP) en constante evolución. Si a principios de los 2010 los sistemas de RAH y TA previos a las RNP llegaron a afrontar con éxito algunas aplicaciones reales como la transcripción y traducción de vídeos docentes pregrabados, ahora en los 2020 son abordables aplicaciones que suponen un reto mucho mayor, como la subtitulación de retransmisiones audiovisuales en directo.

En este mismo período, se están invirtiendo cada vez mayores esfuerzos en la accesibilidad a los medios audiovisuales para todos, incluidas las personas sordas. El RAH y la TA, en su estado actual, son grandes herramientas para aumentar la disponibilidad de medidas de accesibilidad como subtítulos, transcripciones y traducciones, y también para proporcionar acceso multilingüe a todo tipo de contenidos.

En esta tesis doctoral presentamos resultados de investigación sobre RAH y TA basadas en RNP en tres campos muy activos: los recursos educativos abiertos, los contenidos parlamentarios y los medios audiovisuales.

En el área de los recursos educativos abiertos (REA), presentamos primeramente trabajos sobre la evaluación y postedición de RAH y TA con métodos de interacción inteligente, en el marco del proyecto de investigación europeo "transLectures: Transcripción y Traducción de Vídeos Docentes". Los resultados obtenidos confirman que la interacción inteligente puede reducir aún más el esfuerzo de postedición de transcripciones y traducciones automáticas. Seguidamente, en el contexto del posterior proyecto europeo X5gon, presentamos una investigación sobre el desarrollo de sistemas de TA neuronal basados en RNP, y sobre sacar el máximo partido de corpus de TA masivos mediante filtrado automático de datos. Este trabajo dio como resultado sistemas de TA neuronal clasificados entre los mejores en una competición internacional de TA, y mostramos cómo estos nuevos sistemas mejoraron la calidad de los subtítulos multilingües en casos reales de REA.

En el ámbito también en crecimiento de las tecnologías del lenguaje para contenidos parlamentarios, describimos una investigación sobre técnicas de filtrado de datos de habla para el RAH en tiempo real en el contexto de debates del Parlamento Europeo. Esta investigación permitió la publicación de Europarl-ASR, un nuevo y extenso corpus de habla para entrenamiento y evaluación de sistemas de RAH en continuo, así como para la evaluación comparativa de técnicas de filtrado de datos de habla.

Finalmente, presentamos un trabajo en un ámbito en la vanguardia tecnológica del RAH y de la TA: la subtitulación de retransmisiones audiovisuales en directo, en el marco del Convenio de colaboración I+D+i 2020–2023 entre la radiotelevisión pública valenciana À Punt y la Universitat Politècnica de València para la subtitulación asistida por ordenador de contenidos audiovisuales en tiempo real. Esta investigación ha resultado en la implantación de sistemas de RAH en tiempo real, de alta precisión y baja latencia, para una lengua no mayoritaria en el mundo (el catalán) y una de las lenguas más habladas del mundo (el castellano) en un medio audiovisual real.

Résumé

Au cours de la dernière décennie, la reconnaissance automatique de la parole (RAP) et la traduction automatique (TA) se sont énormément améliorées grâce à l'utilisation de modèles de réseaux de neurones profonds (RNP) en constante évolution. Si au début des années 2010 les systèmes pré-RNP de RAP et de TA étaient prêts à s'attaquer avec succès à certaines applications réelles, telles que la transcription et la traduction de cours en vidéo préenregistrés, dans les années 2020 des applications beaucoup plus complexes sont à portée de main, telles que le sous-titrage de retransmissions audiovisuelles en direct.

Dans cette même période, l'accessibilité des médias pour tous, y compris pour les personnes sourdes et malentendantes, est devenue de plus en plus importante. La RAP et la TA, dans leur état actuel, sont des outils puissants pour accroître la disponibilité des mesures d'accessibilité telles que les sous-titres, les transcriptions et les traductions, ainsi que pour fournir un accès multilingue à des nombreux contenus.

Dans cette thèse de doctorat, nous présentons des résultats de recherche sur la reconnaissance automatique de la parole et la traduction automatique basées sur des réseaux de neurones profonds dans trois domaines très actifs : les ressources éducatives libres, les contenus parlementaires et les médias audiovisuels.

En ce qui concerne les ressources éducatives libres (REL), nous présentons tout d'abord des travaux sur l'évaluation et la post-édition de la RAP et de la TA avec des approches d'interaction intelligente, dans le cadre du projet européen "transLectures : Transcription et traduction de cours en vidéo". Les résultats obtenus confirment que l'interaction intelligente peut rendre la post-édition des transcriptions et des traductions automatiques encore plus efficiente. Ensuite, dans le contexte du projet européen X5gon, nous présentons des recherches sur le développement de systèmes de TA neuronale basés sur des RNP, et sur l'exploitation de corpus massifs pour la TA grâce au filtrage automatique des données. Ces travaux ont abouti à une classification de premier rang dans le cadre d'une campagne internationale d'évaluation de la TA, et nous montrons comment ces nouveaux systèmes de TA neuronale ont amélioré la qualité des sous-titres multilingues dans des scénarios réels de REL.

Dans le domaine en pleine expansion des technologies linguistiques pour les contenus parlementaires, nous décrivons des recherches sur les techniques de filtrage des données pour la RAP en temps réel dans le contexte des débats du Parlement européen. Cette recherche a abouti à la publication d'Europarl-ASR, un nouveau et vaste corpus de parole pour l'entraînement et l'évaluation des systèmes de RAP en continu, ainsi que pour l'évaluation des techniques de filtrage des données vocales.

Enfin, nous présentons des travaux à la pointe de la technologie de la RAP et de la TA : le sous-titrage de retransmissions audiovisuelles en direct, dans le contexte de la convention de partenariat R&D 2020–2023 entre le radiodiffuseur public valencien À Punt et l'Universitat Politècnica de València pour le sous-titrage assisté par ordinateur des contenus audiovisuels en temps réel. Cette recherche a abouti au déploiement de systèmes de RAP en temps réel, de haute qualité et à faible latence pour une langue moins parlée (le catalan) et une des langues les plus parlées au monde (l'espagnol) dans un média audiovisuel réel.