UNIVERSITAT POLITÈCNICA DE VALÈNCIA
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

MASTER'S THESIS

Machine Translation of Open Educational Resources: Evaluating Translation Quality and the Transition to Neural Machine Translation

Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

Academic Course 2019/2020

Gonçal Garcés Díaz-Munío

Advisers:
Dr. Alfons Juan Ciscar
Dr. Jorge Civera Saiz

Experimental adviser:
Adrià A. Martínez Villaronga

Abstract / Resum / Resumen / Résumé

Open Education has become a revolutionary approach towards the future of education, enabling worldwide free access to a huge volume of Open Educational Resources (OER). The rapid growth of OER and MOOCs has not gone unnoticed by governments and international organizations, as is demonstrated by the UNESCO-sponsored 2012 Paris OER Declaration and the 2017 Ljubljana OER Action Plan, the latter addressing five strategic actions to support the mainstreaming of OER around UN Sustainable Development Goal 4 (SDG4) on "Quality Education". In the EU, the European Commission's 2013 "Opening up Education" agenda recognized that the EU lacks a critical mass of good quality educational content in multiple languages. Although there is a clear need for multilingual services in Open Education, current OER platforms and MOOCs do not offer multilingual communication and seldom offer multilingual content.

Based on this evidence, this master's thesis aims to foster Open Education with contributions on machine translation for the provision of multilingual access to OER and MOOC platforms. With the trans-disciplinary tools of automatic speech recognition (ASR), machine translation (MT), text-to-speech synthesis (TTS), and dialogue, multilingual access to OER will be possible for everyone regardless of their mother tongue or learning abilities.

Firstly, we present work on the evaluation of MT, including intelligent interaction approaches to post-editing, as carried out in the framework of EU project transLectures. Evaluating MT is still an open question, and so exploring different ways to address this effectively and applying them in real scenarios has a clear interest. The results obtained confirm that the intelligent interaction approach can make post-editing automatic transcriptions and translations even more cost-effective.

Secondly, we present work on developing state-of-the-art neural machine translation (NMT) systems to transition from the previous phrase-based MT models. The new NMT paradigms provide significant improvements in MT quality, and so at this point it is key to move towards them, research on how to improve them, and, again, apply them in real scenarios to verify their usefulness. Our work resulted in a first-rank classification in an international evaluation campaign on MT, and we show the impact that these new NMT systems have in real OER scenarios. A structured comparison shows that our results are on par with the high quality of recent Google Translate results, and also shows that it is possible to go beyond Google Translate's quality through domain adaptation of MT systems.

Resum

L'Educació Oberta ha esdevingut un model revolucionari de cara al futur de l'educació, permetent l'accés lliure a nivell global a un enorme volum de Recursos Educatius Oberts (REO). El ràpid creixement dels REO i dels cursos MOOC no ha passat desapercebut per als governs i les organitzacions internacionals, com es comprova per la Declaració de París sobre els REO de 2012 i el Pla d'Acció de Ljubljana sobre els REO de 2017. Aquest últim proposa cinc accions estratègiques per a la implantació dels REO entorn del 4t Objectiu de Desenvolupament Sostenible de l'ONU (ODS4) sobre «Educació de Qualitat». A l'UE, l'agenda «Obertura de l'educació» de la Comissió Europea (2013) reconeix que a l'UE manca una massa crítica de continguts educatius d'alta qualitat en múltiples llengües. Malgrat la clara necessitat de serveis multilingües per a l'Educació Oberta, les plataformes actuals de REO i MOOC no ofereixen comunicació multilingüe i rarament ofereixen continguts multilingües.

Sobre la base d'aquesta evidència, aquest Treball de Fi de Màster es proposa fomentar l'Educació Oberta amb contribucions sobre la traducció automàtica (TA) per a la provisió d'accés multilingüe a plataformes de REO i MOOC. Les eines interdisciplinàries del reconeixement automàtic de la parla, la traducció automàtica, la síntesi de text a veu i les tecnologies del diàleg faran possible l'accés multilingüe als REO per a tots, independentment de la llengua materna o de les destreses d'aprenentatge.

En primer lloc, presentem un treball sobre l'avaluació de la TA, incloent-hi mètodes d'interacció intel·ligent per a la postedició, en el marc del projecte europeu transLectures. L'avaluació de la TA és encara una qüestió oberta, de manera que explorar diferents vies per a abordar-la de forma efectiva i aplicar-les en escenaris reals té un interés clar. Els resultats obtinguts confirmen que la interacció intel·ligent redueix l'esforç necessari per a la postedició de transcripcions i traduccions automàtiques.

En segon lloc, presentem el desenvolupament de sistemes de traducció automàtica neuronal (TAN) punters per a superar els anteriors models de TA basada en frases. Els nous paradigmes de la TAN aporten millores significatives en la qualitat de la TA, i és fonamental abraçar-los, investigar sobre com millorar-los, i, novament, aplicar-los en escenaris reals per a verificar-ne la utilitat. El sistema de TAN desenvolupat s'ha classificat entre els millors en una competició internacional de TA, i ací mostrem la repercussió que tenen aquests nous sistemes de TAN en escenaris reals de REO. Ens hem comparat i hem comprovat que els nostres resultats competeixen amb l'alta qualitat dels resultats més recents de Google Translate, i que és possible anar més enllà d'aquesta qualitat amb l'adaptació al domini dels sistemes de TA.

Resumen

La Educación Abierta es cada vez más un modelo revolucionario de cara al futuro de la educación, permitiendo el acceso libre a nivel global a un enorme volumen de Recursos Educativos Abiertos (REA). El rápido crecimiento de los REA y de los cursos MOOC no ha pasado desapercibido para los gobiernos y las organizaciones internacionales, como se comprueba por la Declaración de París sobre los REA de 2012 y el Plan de Acción de Liubliana sobre los REA de 2017. Este último propone cinco acciones estratégicas para la implantación de los REA en torno al 4.º Objetivo de Desarrollo Sostenible de la ONU (ODS4) sobre «Educación de Calidad». En la UE, la agenda «Apertura de la educación» de la Comisión Europea (2013) reconoce que la UE carece de una masa crítica de contenidos educativos de alta calidad en múltiples lenguas. A pesar de la clara necesidad de servicios multilingües para la Educación Abierta, las plataformas actuales de REA y MOOC no ofrecen comunicación multilingüe y raramente ofrecen contenidos multilingües.

En base a esta evidencia, este Trabajo de Fin de Máster propone fomentar la Educación Abierta con contribuciones sobre la traducción automática (TA) para la provisión de acceso multilingüe a plataformas de REA y MOOC. Las herramientas interdisciplinares del reconocimiento automático del habla, la traducción automática, la síntesis de texto a voz y las tecnologías del diálogo harán posible el acceso multilingüe a los REA para todos, independientemente de la lengua materna o de las capacidades de aprendizaje de cada uno.

En primer lugar, presentamos un trabajo sobre la evaluación de la TA, incluyendo métodos de interacción inteligente para la postedición, en el marco del proyecto europeo transLectures. La evaluación de la TA es aún una cuestión abierta, de forma que explorar diferentes vías para abordarla de forma efectiva y aplicarlas en escenarios reales tiene un claro interés. Mostraremos resultados que confirman que la interacción inteligente reduce el esfuerzo necesario para la postedición de transcripciones y traducciones automáticas.

En segundo lugar, presentamos un esfuerzo de desarrollo de sistemas de traducción automática neuronal (TAN) punteros para superar los anteriores modelos de TA basada en frases. Los nuevos paradigmas de la TAN aportan mejoras significativas en la calidad de la TA, y es fundamental adoptarlos, investigar sobre cómo mejorarlos, y, de nuevo, aplicarlos en escenarios reales para verificar su utilidad. El sistema de TAN desarrollado se ha clasificado entre los mejores en una competición internacional de TA, y aquí mostramos la repercusión que tienen estos nuevos sistemas de TAN en escenarios reales de REA. Nos hemos comparado y hemos comprobado que nuestros resultados compiten con la alta calidad de los resultados más recientes de Google Translate, y que es posible ir más allá de esta calidad con la adaptación al dominio de los sistemas de TA.

Résumé

L'éducation ouverte est devenue un modèle révolutionnaire pour le futur de l'éducation, permettant l'accès libre au niveau mondial à un immense volume de ressources éducatives libres (REL). La croissance rapide des REL et des cours MOOC a été notée par les gouvernements et les organisations internationales, comme nous pouvons le voir dans la Déclaration de Paris sur les REL 2012 et le Plan d'action de Ljubljana sur les REL 2017. Ce dernier propose cinq actions stratégiques pour l'implantation des REL autour du 4ème objectif de développement durable de l'ONU (ODD4) sur l'« Éducation de qualité ». Au niveau de l'UE, le programme « Ouvrir l'éducation » de la Commission Européenne (2013) reconnaît la manque d'une masse critique de contenus éducatifs d'haute qualité en plusieures langues. Malgré une claire nécessité de services multilingues pour l'éducation ouverte, les plateformes actuelles de REL et MOOC n'offrent pas de communication multilingue et rarement de contenus multilingues.

Sur cette évidence, ce projet de master prétend promouvoir l'éducation ouverte avec des contributions sur la traduction automatique (TA) pour fournir d'accès multilingue les plateformes de REL et MOOC. Les outils interdisciplinaires de la reconnaissance automatique de la parole, la traduction automatique, la synthèse vocale et les technologies du dialogue rendront possible l'accès multilingue aux REL pour tous, indépendamment de la langue maternelle ou des capacités d'apprentissage.

En premier lieu, nous présentons notre travail sur l'évaluation et la post-édition de la TA, avec des méthodes d'interaction intelligente pour la post-édition, dans le cadre du projet européen transLectures. L'évaluation de la TA est encore une question ouverte, ce qui signifie que l'exploration des diverses voies pour l'aborder d'une manière effective a un intérêt clair. Les résultats obtenus confirment que l'interaction intelligente réduit l'effort nécessaire pour la post-édition de transcriptions et traductions automatiques.

En deuxième lieu, nous présentons notre développement de systèmes de traduction automatique neuronale (TAN) de pointe pour dépasser les antérieurs modèles de TA basée sur les séquences de mots. Les nouveaux paradigmes de la TAN apportent des améliorations significatives de la qualité de la TA, et il est donc fondamentale de les adopter, de rechercher sur la façon de les améliorer et aussi de les appliquer dans des environnements réels pour en vérifier l'utilité. Le système de TAN développé a été classifié entre les meilleurs dans une compétition internationale de TA, et ici nous montrons l'impact que ces nouveaux systèmes de TAN ont dans des environnements réels de REL. Avec une comparaison systématique, nous avons vérifié que nos résultats rivalisent avec l'haute qualité des résultats les plus récents de Google Translate, et qu'il est possible de dépasser cette qualité avec l'adaptation au domaine des systèmes TA.

Contents

Chapter 1

Motivation and objectives

1.1 Motivation

The growing importance and new challenges of Open Education

Open Education has become a revolutionary approach towards the future of education, enabling worldwide free access to a huge volume of Open Educational Resources (OER).

A prominent example of OER are the OpenCourseWare (OCW) courses produced at universities and published for free via the Internet since the early 2000s. Although OCW projects have had a large impact on Open Education, the so-called Massive Open Online Courses (MOOCs) are increasing this impact even further in recent years. In contrast to online repositories of (OCW) lectures, MOOCs offer more structured formal courses, discussion forums and the possibility of earning academic certificates.

In Spain, many educational organizations have increased their production of OER and joined, or even launched, MOOC initiatives such as MiríadaX, UNED COMA, UniMOOC and UPV[X].

The rapid growth of OER and MOOCs has not gone unnoticed by governments and international organizations concerned with Education, as is demonstrated by the 2012 Paris OER Declaration [5] adopted at the World OER Congress held by UNESCO. This Declaration showed the importance of OER in widening access to Education and enumerated 10 recommendations to states regarding international collaborations and accessibility. The worldwide push for OER was confirmed more recently in 2017 with the adoption of the 2017 Ljubljana OER Action Plan [7] at the 2nd World OER Congress, held again by UNESCO. This plan addresses five strategic actions to support the mainstreaming of OER around the United Nations Sustainable Development Goal 4 (SDG4) on "Quality Education".

In the EU, following the Paris Declaration, the European Commission launched the "Opening up Education" agenda [6] in September 2013 for stimulating high-quality, innovative ways of learning and teaching through new technologies and digital content. In it, it is recognized that the EU lacks a critical mass of good quality educational

content in multiple languages. Although there is a clear need of multilingual services in Open Education, current OER-based platforms, and MOOCs in particular, do not offer multilingual communication and only occasionally offer multilingual content.

Machine Translation of Open Educational Resources

Based on this evidence, this master's thesis aims to foster Open Education with contributions on machine translation for the provision of multilingual access to OER and MOOC platforms. These purposes can be achieved by using a combination of trans-disciplinary tools: automatic speech recognition (ASR), machine translation (MT), text-to-speech synthesis (TTS), and dialogue. By using ASR, MT and TTS, multilingual access to OER will be possible for everyone regardless of their mother tongue or learning abilities. This approach has been successfully applied by the MLLP research group of Universitat Politècnica de València in the context of different EU and Spanish research projects (transLectures, EMMA, X5gon, MORE).

The work presented in this master's thesis focuses on MT and was carried out in the framework of these projects.

Firstly, we present work on the evaluation of machine translation, including intelligent interaction approaches to post-editing, as carried out in the framework of EU project transLectures (2011–2014). Evaluating MT is still an open question, and so exploring different ways to address this effectively and applying them in real scenarios has a clear interest. We will see how different methods of evaluation have different applications, and how they complement each other to provide a clearer and more complete assessment[1].

Secondly, we present work on developing state-of-the-art neural machine translation (NMT) systems to transition from the previous phrase-based models. The new NMT paradigms provide significant improvements in MT quality, and so at this point it is key to move towards them, research on how to improve them, and, again, apply them in real scenarios to verify their usefulness[2].

UN Sustainable Development Goals: Quality Education for all

The work of this master's thesis fits with the United Nations' Sustainable Development Goals (SDG). In particular, with SDG 4, on "Quality Education", which aims to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" [2].

Within SDG 4, the three targets more directly related to this master's thesis are:

- 4.3 By 2030, ensure equal access for all women and men to affordable and quality technical, vocational and tertiary education, including university.

[1] The work covered in this chapter resulted in several scientific publications, including an article in the indexed international journal *Open Learning*. See Section 3.5.1 for the details of these publications.

[2] Part of the work covered in this chapter resulted in a first-rank classification in an international evaluation campaign on machine translation, accompanied by its publication as an article in the international Third Conference on Machine Translation (WMT18). See Section 4.6.1 for the details of this publication.

- 4.4 By 2030, substantially increase the number of youth and adults who have relevant skills, including technical and vocational skills, for employment, decent jobs and entrepreneurship.
- 4.5 By 2030, eliminate gender disparities in education and ensure equal access to all levels of education and vocational training for the vulnerable, including persons with disabilities, indigenous peoples and children in vulnerable situations.

The work of this master's thesis contributes to increasing accessibility to Open Educational Resources for everyone (target 4.3), including persons with hearing disabilities (target 4.5), speakers of minority or minoritized languages, and persons with fewer resources or without access to formal education systems (target 4.5), for whom access to Open Educational Resources (at the level of primary, secondary and higher education) in a language they understand constitutes an essential source for learning and professional renewal (target 4.4).

1.2 Objectives

1. **Contributions to evaluation and post-editing of machine translation in the context of the automatic transcription and translation of Open Educational Resources, with hybrid intelligent interaction approaches**

   Evaluating machine translation is still and open question, and so exploring different ways to address this effectively and applying them in real scenarios has a clear interest. In this master's thesis, we will present a review of work on evaluation of machine translation in the context of automatic transcription and translation of Open Educational Resources in a real scenario, the EU research project transLectures and its pilot cases. We set out to analyse how different methods of evaluation have different applications, how they complement each other to provide a clearer and more complete assessment, and to see how they are used to evaluate the cost-effectiveness of automatic transcriptions and translations.

   Among these evaluation methods, an important one is measuring post-editing times, in order to confirm whether revising the automatic transcriptions and translations saves time and costs with respect to transcribing and translating from scratch. In this context, we will evaluate an *intelligent interaction* approach proposed in transLectures in which the system first identifies which sections of an automatic transcription or translation are likely to contain the most errors (based on automatic confidence measures), and then presents only these sections to the user for correction; these corrections are then fed back into the system and used to re-train the underlying models with a view to avoiding the same errors in the future.

   The challenge is to implement in real use cases this intelligent interaction approach to post-editing, as well as hybrid approaches, and to determine whether it can make post-editing automatic transcriptions and translations even more cost-effective.

2. **Developing state-of-the-art neural machine translation systems for their application in the context of Open Educational Resources**

   The new neural machine translation paradigms provide significant improvements in machine translation quality, and so at this point it is key to move towards them, research on how to improve them, and apply them in real scenarios to verify their usefulness. In this master's thesis, we will present work on developing state-of-the-art neural machine translation systems to transition from the previous phrase-based models.

   The challenge is to develop neural machine translation systems based on state-of-the-art neural machine translation architectures, making the most of available multilingual and monolingual corpora, and applying appropriately techniques which are essential to improve the quality of NMT systems: corpus filtering, data augmentation through backtranslations, fine-tuning for domain adaptation.

   Additionally, we will show the impact that these new neural machine translation systems have in real Open Educational Resource scenarios within the framework of EU research project X5gon.

1.3 Structure of this master's thesis

According to the objectives set out above, two topics constitute the core of this master's thesis, all of it in the context of machine translation of Open Educational Resources, based on work and publications by the author: on the one hand, evaluation and post-editing of machine translation; on the other hand, the transition from phrase-based machine translation to neural machine translation.

Before Chapters 3 and 4, where these two main topics are covered, the current Chapter1 contains the motivation and objectives for this master's thesis, as well as the current section on its structure. Then, in Chapter 2, some preliminary matters are presented, including a brief review of the state of the art in machine translation, an introduction to the different methods for machine translation evaluation, and a compilation of current tools available for machine translation.

Chapter 3, "Evaluating automatic transcriptions and translations in a real scenario" (based on work carried out by the author as a researcher in EU project transLectures, 2011–2014, including a 2014 journal publication[3]), begins with an introduction (to the topic, and to the specific context), then goes into the topic with three sections covering different evaluation methods (automatic evaluations, human evaluations by experts, and human evaluations by users, this latter including intelligent interaction approaches to post-editing), before finishing with some conclusions (including a summary of publications and contributions by the author).

Chapter 4, "The transition from phrase-based to neural machine translation" (based on work carried out for the Third Conference of Machine Translation, WMT18,

[3]The work covered in this chapter resulted in several scientific publications, including an article in the indexed international journal *Open Learning*. See Section 3.5.1 for the details of these publications.

and on its corresponding 2018 publication[4]), begins with an introduction to the context of the chapter and a description of the datasets used. Afterwards, the development of a new NMT system for WMT18 is summarized in a section based on our WMT18 publication, followed by a comparison of the results obtained with PBMT and NMT technologies, an analysis of the impact of the new NMT systems on real OER scenarios, and finally by the chapter's conclusions (including a summary of publications and contributions by the author).

Finally, Chapter 5 contains a summary of achievements and the overall conclusions for this master's thesis.

[4]Part of the work covered in this chapter resulted in a first-rank classification in an international evaluation campaign on machine translation, accompanied by its publication as an article in the international Third Conference on Machine Translation (WMT18). See Section 4.6.1 for the details of this publication.

Chapter 2

Introduction

In this chapter, some preliminary matters are presented, including a brief review of the state of the art in machine translation, an introduction to the different methods for machine translation evaluation, and a compilation of current tools available for machine translation.

2.1 State of the art in machine translation technologies

2.1.1 Introduction

Machine translation (MT) is translation carried out by computers. One of its main goals is to be able to translate texts with a quality as close as possible to that achievable by an expert human translator (although there are other very important goals such as producing translations quickly/instantly, and there are many tasks where machine translation is useful even in cases where it is not close to human translation quality).

Improvements in machine translation since its beginnings in the 1950s have come from several sides: from using and improving different *methods* for machine translation, from increasing the amount and quality of the linguistic *data* used (monolingual, bilingual and multilingual text corpora), and from improvements in *computer hardware*, which allow for the use of better and better methods with more and more linguistic data (to extents which at previous points in time could only be theorized).

Methods for machine translation have moved from *rule-based machine translation* (based on codifying linguistic rules for translation from expert knowledge) to *statistical machine translation* (based on machine learning, computing translation probabilities from multilingual text corpora). Within statistical machine translation (which is currently the most successful approach), in the 2000s the best results were obtained using *phrase-based machine translation* (where the units of translation are sequences of words or *phrases*, and the MT model is composed of a combination of a translation model, a reordering model and a language model); nowadays, since the mid-2010s, the

best results have been obtained with MT methods based on (deep) artificial neural networks, known as *neural machine translation*[1].

This latter trend, the huge advancements brought by new uses of deep artificial neural networks (DNNs), has been common in the 2010s to all fields of machine learning and natural language processing, including (statistical) machine translation. The increasing capacities of recent graphics processing units (GPUs) for matrix computations and parallel computing have multiplied the possibilities for experimenting with and applying DNNs, which is also facilitating the theoretical development of this mathematical tool for machine learning.

The recent use of DNNs showed its effects first in the related field of automatic speech recognition (ASR), with the 2012 proposal by Li Deng's team at Microsoft Research [25]. This work opened the door to a whole series of improvements in the accuracy of ASR systems, which has gone beyond 20% relative improvement in many cases [91].

In the case of MT, after the pioneering work by Yoshua Bengio at Université de Montréal in early 2015 [11], many works have been published in which DNNs have clearly surpassed the previously standard phrase-based machine translation (PBMT) models. In 2016, neural machine translation (NMT) systems consistently prevailed over PBMT systems [63], even achieving levels of quality close to those of human translators [88]. In 2017, 2018 and 2019, new DNN architectures for NMT have again significantly surpassed previous results, in some cases with relatively low computational costs (which is key in order to facilitate both research tasks and practical applications) [84, 21, 86]. These advances in NMT have not been confined to research laboratories: commercial MT systems are making the same transition concurrently [88].

These new applications of DNNs in MT [46] (and in other fields of machine learning and natural language processing) are resulting in improved results year after year, and they show no signs of stagnating any time soon. On the other hand, a this point we are not able to interpret an NMT model in as detailed a way as it is possible with PBMT models (where all the information and computations are explicit), and we still do not have a full comprehension of the reasons behind the superiority of DNN models over PBMT models [52].

Taking these circumstances into account, this is a field with many possibilities for research and development, both as relates to its theoretical aspects (deepening our knowledge of DNNs and NMT), and to its practical applications (developing new and better NMT systems for an increasing number of uses).

In this section, after this introduction, we will briefly review the state of the art in phrase-based machine translation and the currently more successful neural machine translation, and we will briefly summarize how the technology presented in this master's thesis fits within the state of the art in machine translation.

[1] Note that, in part of the bibliography, the term *neural machine translation* is treated as separate from *statistical machine translation*, the latter being reserved for word-based and phrase-based machine translation. In this master's thesis, we consider neural machine translation a part of statistical machine translation, since it is equally based on statistics and probability, just using a different method to compute the probabilities.

Preliminary concepts: Machine learning and deep learning (and neural networks)

For readers unfamiliar with these terms, we will have a brief look at their meaning before going on.

As already mentioned in this section, the state of the art in machine translation are the methods known as statistical machine translation, which are based on machine learning (as opposed to the methods known as rule-based machine translation, based on codifying linguistic rules for translation from expert knowledge).

Machine learning (ML), a part of the techniques for artificial intelligence, is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. ML algorithms build a mathematical model based on sample data in order to make predictions or decisions without being explicitly programmed to perform the task[2].

What about *deep learning*, then? Deep learning is an approach to ML based on the use of large, many-layered artificial neural networks.

Let's take here a step back to briefly explain the concept of *artificial neural networks*, since this is key in modern machine translation and machine learning. Artificial neural networks (ANNs)[3] are a mathematical tool used in ML to compute and apply statistical models, vaguely inspired by the biological neural networks of animal brains. Simply stated, an ANN will take as input an observation, will process the data of this observation through a network of non-linear mathematical functions (the "neurons"), and will output a decision based on the input observation. The (mathematical) behaviour of the ANN will be "trained" by processing a set of examples (each example being made up of an observation and the expected decision to be made from it), thus "learning" a statistical model that minimizes the decision errors made by the ANN on the set of examples. The goal is for the trained ANN to be able to make appropriate decisions when confronted with new observations that are not the same ones that it saw before in the set of examples[4].

The mathematical structure of an ANN is typically organized in

[2]Some additional information about terminology: ML is closely related to the field of *pattern recognition*. Pattern recognition has been defined as "the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories". In fact, ML is one of the possible approaches to pattern recognition, but it has been so successful that frequently the two concepts are treated as synonymous (in the same way that statistical machine translation has been so successful as to frequently be treated as synonymous with machine translation in general).

[3]Also known as connectionist systems.

[4]In the case in point of machine translation, an ANN is trained on a set of examples of parallel source sentences and target sentences from two different languages, and the goal is for the trained ANN to be able to produce correct translations when confronted with new source sentences that it has not seen before.

"layers" of neurons through which the input data is processed. Larger, more complex, "deeper" ANNs with more layers are capable of better modelling more complex realities. Back in the 20th century, hardware limitations restricted us to use only very simple, "shallow" ANNs (with few layers). In the 2010s, however, as already mentioned in this section, new hardware setups allowed for the use of more and more complex, deeper and deeper ANNs.

And now we can come back to the concept of deep learning (DL): the use of large, many-layered, deep artificial neural networks (DNNs) to create better statistical models for complex realities as a powerful technique for ML.

2.1.2 Phrase-based machine translation

As mentioned above, state-of-the-art machine translation systems work on the principle of statistical models. The statistical machine translation (SMT) approach was originated in the 1980s by a research group at IBM which applied principles initially pioneered for automatic speech recognition to the task of human language translation. The main principles of SMT are: decomposition of the translation process into a sequence of rule application; extracting the translation rules automatically from a large parallel corpus of translated text (consisting of millions, if not billions of words); defining a probabilistic model over the translation rules and optimizing it to best fit the data.

In the 2000s, the best results were obtained using phrase-based machine translation (PBMT). In PBMT systems, given a text sentence to be translated from a source language into a target language, the system decides its most likely translation by combining (as a log-linear model) a translation model, a reordering model and a language model. Translation models are usually implemented in terms of phrase tables, learnt by maximum likelihood estimation (although discriminative training is also being increasingly used); language models are usually based on n-gram statistics (although DNN-based language models are replacing n-gram language models); while the reordering model can be a predefined distance-based model (although it can otherwise be learnt from the data) [45].

SMT and PBMT were born at IBM Research in the late 1980s, inspired by the success of statistical methods in automatic speech recognition. While the foundations of PBMT were developed in the 1980s and 1990s at IBM and some other research groups, it was not until the 2000s that it saw widespread and commercial implementation. In 1998, participants at a Johns Hopkins University workshop re-implemented most of the IBM methods and released the resulting SMT engine, called Moses [48], as free software. Access to Moses, a powerful and efficient SMT engine, made research, development and deployment of full-fledged PBMT systems possible for researchers and industry.

Since the mid-2010s, neural machine translation (NMT) methods have surpassed PBMT as the state of the art in MT [18]. However, PBMT still finds use for some applications. A notable recent example is the successful use of PBMT as part of the

techniques used in the state of the art for unsupervised machine translation [9] (where a machine translation system is trained solely from monolingual corpora in the two languages of the desired translation pair, without the usual parallel corpora).

2.1.3 Neural machine translation

In the last two decades, there has been significant progress in SMT: the emergence of phrase-based machine translation models around 2000 and grammar-based approaches which matured at the end of the 2010s, and the increased use of advanced machine learning methods in the last 5 years, such as deep neural networks (DNNs). Since 2017, this latter approach, known as neural machine translation (NMT), has been the state of the art in SMT [18].

Following the NMT approach, the application of convolutional models [43] and sequence-to-sequence models [74, 24] provided the first accurate results, but only for short sentences. The incorporation of the attention mechanism allowed for competitive results [11, 37]. In 2017, the idea of the attention mechanism was generalized to self-attention [84], a technique that has been providing the best results since then. At the same time, other efficient DNN architectures, such as convolutional networks, were being explored [31].

In 2018, an enhanced RNN with attention mechanism was developed that could compete with self-attention, although at a much greater computational cost [21]. Interesting results were also obtained by using this latter topology with a character-based approach, at an even greater cost [23]. The current state of the art relies again on self-attention, refining the technique so that deeper self-attentive networks can be trained, instead of focusing on widening networks as was the trend previously [86].

There has also been more and more research in domain adaptation and data filtering [20, 41, 47], multilingual machine translation [39, 53, 8], online machine translation [92] and integration of neural machine translation systems into real industry applications [88, 51]. As MT at the segment level seems to be approaching human parity in many cases, interest is growing in turning the focus towards MT at the document level [49, 42].

2.1.4 This master's thesis in the context of the state of the art

The two core chapters of this master's thesis are based on the state of the art of machine translation at the time of the work they respectively cover. Chapter 3 covers work carried out during 2011–2014, using state-of-the-art phrase-based machine translation techniques of the time. While machine translation technology has evolved since then, this chapter focuses on machine translation evaluation, based on the methods that are still current nowadays. Chapter 4 covers work carried out in 2018, using the then recently published state-of-the-art Transformer neural machine translation architecture, and so here we will see the basis of current machine translation systems.

In this master's thesis we can see the evolution from one machine translation paradigm to the next one, and the machine translation evaluations we show confirm

the appropriateness of continuing to work on developing advanced neural machine translation systems.

2.2 Evaluation of machine translation

Evaluating the quality and usefulness of MT systems is not straightforward. This is in great part caused by a characteristic of translation itself: the fact that any source text can be translated correctly in many different ways in the target language. Thus, unlike in other related areas of natural language processing or machine learning (such as automatic speech recognition or image recognition), there is no single correct answer, no single correct translated text against which to compare the output of the MT system under evaluation. This fact makes both automatic and human evaluations of MT more complex.

In this section, we will explain some of the most common ways of evaluating MT systems. Firstly, we will look at automatic metrics (which are relatively quick and cheap to obtain, although their reliability is disputed because of the problem explained above). Secondly, we will consider human evaluations (which are much more costly in time and effort, but in the case of MT they continue to be especially important given the limitations of MT automatic metrics). Thirdly, we will talk about the relation and balance between automatic and human evaluations. Finally, we will discuss measuring the time and memory efficiency of MT systems.

These evaluation methods will come up in Chapters 3 and 4 of this master's thesis.

> ***Preliminary concepts*: Sentence level versus document level**
>
> MT system training is usually carried out based on corpora composed of millions of parallel sentences, but at the *sentence level*, that is, learning from one sentence pair at a time. The trained MT system will then translate input texts sentence by sentence, which means that only the current sentence is used as context for translation decisions.
>
> At the same time, the focus of MT evaluations has also been frequently at the sentence level, that is, the reference and automatic translations are compared sentence by sentence. This is a convenient way of evaluating: it is ideal for automatic evaluation metrics based on edit distances, and it has also been used as a way to organize human evaluations (i.e., human evaluators may be asked to rate translation quality by giving a score to each translated sentence).
>
> However, as MT quality metrics at the sentence level become higher and higher (to the point that it is becoming more and more difficult to improve them), more importance is being given to MT training and evaluation at the *document level*, that is, taking into account contexts larger than a single sentence, and thus considering questions of linguistic coherence between different sentences and all along the full text to be translated (e.g., word agreement between different sentences, coherence in references between points far apart in the text...).

> In this master's thesis, MT training and automatic evaluations have been carried out at the sentence level, and as we will see, some of the human evaluations too, although human evaluations have also provided information on quality at the document level.

2.2.1 Automatic metrics of machine translation accuracy

Automatic metrics are important because they can be calculated quickly and cheaply, without the effort required for human evaluations. Thus, during research and development of MT systems, automatic metrics can quickly provide an indication of whether the new systems actually provide better translation quality than the previous ones.

The basic idea behind the most commonly used automatic metrics is to compare the output of the MT system against a *reference* (correct) *translation*. The closer the MT output to the reference translation, the better its quality. Thus, to automatically evaluate an MT system, generally we need a bilingual parallel corpus that has not been used for the training of the MT system; the MT system is used to translate the source text of the corpus, and then the MT system is evaluated by comparing its output against the target text of the corpus.

The question of automatic metrics for MT is still open. As explained in the introduction to this section, the fact that there is no single correct translation with which to compare the output of an MT system makes it complex to define a useful automatic metric.

As a way to overcome this problem, some of the metrics can evaluate MT output against several correct translations into the same language (instead of against a single one). However, it is not usual to have corpora with different translations into the same language (nor is it easy to create them), and we would still have to determine how much does comparing against two or more correct translations contribute to covering the possible space for acceptable variations in translation.

In any case, the most commonly used automatic metrics nowadays have been shown to have a high degree of *correlation to human judgement* [59, 69, 12], which makes them useful even if not definitive.

Additionally, to make up for the limitations of any single automatic metric, it is usual to report evaluations using more than one of them (we will see this in some sections of this master's thesis). If a significant difference is found between the evaluation results with different metrics, this can be an indication that a more detailed analysis is necessary.

In this section we cover the most commonly used automatic metrics for MT: BLEU and TER. Before explaining them, we will introduce WER, a simpler automatic metric that is commonly used in the related field of automatic speech recognition (ASR), and which can be useful as a stepping stone towards understanding TER and BLEU. We will also introduce the related concept of confidence measures (which, despite not being actually an automatic evaluation metric, are a related concept that will be applied in Chapter 3 for intelligent interaction).

***Preliminary concepts*: n-grams**

The concept of *n-gram* is an important one in machine translation and in natural language processing in general. An n-gram is a contiguous sequence of n items (phonemes, syllables, letters, words...) from a given text. In the case of machine translation, we usually consider *word* n-grams. Thus, a 1-gram (unigram or one-gram) is made up of 1 word; a 2-gram (bigram or two-gram) is a sequence of 2 words; a 3-gram (trigram or three-gram) is a sequence of 3 words; and so on.

The concept of n-gram is used frequently in automatic metrics of machine translation accuracy. Since a concept or term can be expressed with a different number of words in different languages (or even in different correct translations in the same language), and these multi-word concepts or terms can be placed differently within equivalent sentences (in different languages or in the same language), it is common for automatic metrics to compare word sequences of different sizes; that is, n-grams of different size n. We will see this below, in the definitions of TER and BLEU.

In a different context in the field of machine translation, n-grams are the basis of the most successful language models used in phrase-based machine translation in the 2000s.

2.2.1.1 WER: The standard metric in automatic speech recognition

We explain here an automatic metric that is commonly used in the related field of automatic speech recognition (ASR), but which is not as useful for MT (and thus, is not usually used by itself for MT).

The Word Error Rate (WER) is the ratio, expressed as a percentage, between the number of basic word editing operations required to convert an automatic transcription (or translation) into the reference transcription (or translation), and the total number of words in the reference transcription (or translation)[5].

The WER is derived from the Levenshtein (edit) distance working at the word level, adding length normalization.

It is computed as:

$$\text{WER} = \frac{I + D + S}{N} = \frac{I + D + S}{S + D + C} \tag{2.1}$$

where I is the number of insertion errors, D is the number of deletion errors, S is the number of substitution errors, C is the number of correct words, and N is the number of words in the reference ($N = S + D + C$).

This results in a score (usually) between 0 and 1 (frequently reported between 0 and 100, for better readability). Being an error metric, the lower the WER, the higher the quality of our system's output.

[5]In this context, an automatic transcription is the output of automatic speech recognition, while an automatic translation is the output of machine translation (see the introduction to Section 3.1 for a more detailed explanation of the use of these terms in this master's thesis).

In the case of ASR, since there is only one reference (correct) transcription with which to compare the ASR system's output, WER is a very useful metric (it simply tells us how far from the single correct transcription is our ASR system's output), and thus it is very commonly used[6].

In the case of MT, however, since there can be many different correct translations for any source text, our MT system's output might be a good translation even if it has a low WER with respect to the reference translation we are comparing it to (i.e., even if it has a very different wording).

This is why more suitable metrics have been designed and are commonly used for MT, such as BLEU and TER, which we explain below.

2.2.1.2 TER

The Translation Edit Rate (TER) [69] is the ratio, expressed as a percentage, between the number of word editing operations required to convert the automatic translation into the reference (correct) translation, and the number of words in the reference translation.

TER is similar to WER, but also considers any swaps of consecutive word groups (n-grams) required to make the MT system output match the reference translation.

The TER is computed as:

$$\text{TER} = \frac{\text{Number of edits required}}{\text{Number of words in the reference}} \tag{2.2}$$

where the edits can be: insertion of a word; deletion of a word; substitution of a word; and movement of a block of contiguous words to a different part of the sentence.

This results in a score (usually) between 0 and 1 (frequently reported between 0 and 100, for better readability). Being an error metric, the lower the TER, the higher the translation quality.

2.2.1.3 BLEU

BLEU (BiLingual Evaluation Understudy) is a general method for automatic MT evaluation, and also a specific "baseline metric" based on this method, which were proposed in 2002 by an IBM research team [59].

Based on the central idea that "the closer a machine translation is to a professional human translation, the better it is", the authors described BLEU as "a method of automatic machine translation evaluation that is quick, inexpensive, and language-independent, that correlates highly with human evaluation, and that has little marginal cost per run".

BLEU computes an n-gram precision[7], based on the question: "what percentage of n-grams from the automatic translation can be found in the reference translation(s)?".

[6] The WER can actually be found to have limitations even for ASR, but we will not go into them here as they are out of the scope of this master's thesis.

[7] *Precision*, in the field of pattern recognition, being itself a basic evaluation metric. Precision measures how much of a system's output is correct with regard to the reference. When used by itself as a metric, it is usually reported together with *recall*, which measures how much of the reference is reflected in the system's output.

In other words, BLEU measures the degree of overlap between the automatic translation and a (corpus of) good quality human reference translation(s), comparing isolated words and sequences of up to n words (that is, 1-grams, 2-grams... up to n-grams, for a given value of n).

The BLEU metric is defined as:

$$\text{BLEU-}n = \text{brevitypenalty} \exp \sum_{i=1}^{n} \lambda_i \log \text{precision}_i \tag{2.3}$$

where

$$\text{precision}_n = \frac{\text{Number of correct } n\text{-grams in the output}}{\text{Number of } n\text{-grams in the output}}$$

$$\lambda_n : \text{Weight attributed to precision}_n$$

$$\text{brevitypenalty} = \min\left(1, \frac{\text{Number of words in the output}}{\text{Number of words in the reference}}\right)$$

The brevity penalty in the formula is used to address a problem with metrics based on precision: we could obtain high precision scores by outputting only the words that we are sure we know how to translate correctly (as there is no penalty in the precision metric for leaving out from our output parts of the source text). Thus, the brevity penalty reduces the BLEU score if the output is too short with respect to the reference.

In practice, the maximum order n of n-grams to be matched is usually 4 (BLEU-4) [45].

BLEU provides a score from 0 to 1 (frequently reported from 0 to 100, for better readability). Being a quality metric, the higher the BLEU score, the better the MT output evaluated.

2.2.1.4 Confidence measures

We introduce here the concept of confidence measures, which are related to automatic evaluation metrics, but different to them in that the former are computed by an ASR or MT system as an automatic estimation of the correctness of each part of its output, without actually evaluating it against a reference transcription or translation.

Confidence estimation aims at providing confidence measures of the automatic speech recognition or machine translation output at a certain level of granularity such as sub-word, word, utterance, sentence or sub-sentence [26, 16]. Confidence measures are represented by scores usually between 0 and 1 which are intended to reflect the reliability of a given ASR or MT output.

Confidence measures have been shown to provide useful estimations in the case of ASR [26]. In the case of MT, confidence estimation is a more difficult problem. So far, no method for confidence estimation in MT has been clearly shown to provide useful results [16].

Confidence measures can be used to predict which points of an automatic transcription or translation are likely to contain errors, which is why they are the basis for intelligent interaction techniques[8].

2.2.2 Human evaluations

When well organized, human evaluations can provide us the most reliable information about the quality of the output of an MT system. They are, however, more costly in time and effort than automatic metrics.

In this section, we will first describe how direct assessment works, followed by explaining human evaluation based on measurements of post-editing effort.

2.2.2.1 Direct assessment and relative ranking

The most straightforward form of human evaluation consists in asking evaluators to judge the correctness of the output of the MT system being evaluated. Ideally, evaluators would be bilingual, in order to evaluate the automatic translation with respect to the source text in the original language. However, since it is not easy to gather a sufficient number of bilingual evaluators, the evaluation can be carried out by evaluators that are monolingual on the target language: in this case, they do not judge the automatic translation against the source text, but against a reference translation in the target language.

Usually, the evaluation is done sentence by sentence. Evaluators may be asked to assign a score to each automatically translated sentence, in what is called *direct assessment*; or they may be shown automatically translated sentences generated by different MT systems, and be asked to rank them by correctness, in what is known as *relative ranking*. The results of both methods correlate very strongly [13].

Traditionally, direct assessment scores are given for two criteria: adequacy (how much of the meaning of the source text has been carried over to the automatic translation) and fluency (how correct is the language in the automatic translation). Two scores can be given to judge each sentence, one for each criterion; or a single score can be given taking into account both criteria.

2.2.2.2 Post-editing effort: RTF and other relative metrics

One of the ways to evaluate the quality of automatic translations is to measure how much time is required to correct the errors of the automatic output (the idea being that, the better the quality of the automatic translation, the less time will be required to correct it).

In the case of automatic translations of video subtitles, such as automatic translations of subtitles of video lectures (which is the case covered in Chapter 3 of this master's thesis), the time required to post-edit an automatic translation can be measured relatively to the length of the video.

[8] This will be important in Chapter 3. See Section 3.1.1.2 for an introduction to the concept of intelligent interaction, and Section 3.4.1 for its application to post-editing ASR output.

In this master's thesis, we refer to this measure as the *post-editing Real Time Factor (RTF)*[9], the ratio between the time devoted to correcting the automatic translation of a video and the duration of that video:

$$\text{Post-editing RTF} = \frac{\text{Post-editing time}}{\text{Video length}} \tag{2.4}$$

The lower the RTF, the less effort required for correction, and thus, the higher the quality of the automatic translation (e.g., an MT system the output of which takes RTF 20 to correct [20 times the length of the video] is probably better than a different MT system with a post-editing RTF of 30 [30 times the length of the video]).

This metric is useful because it is centred around a value that is immediately available and easy to interpret, the length of the subtitled video. This makes it easy to compute and understand the measure and to make estimations from our measurements for new videos. However, it has a downside: different videos with the same length can actually show large differences in the amount of subtitles, since there can be different amounts of segments without speech or speakers with very different speech paces (thus, with different amounts of speech for an equal length of time). This means that from an RTF measurement we would estimate the same post-editing time for two new videos of equal video length, but the actual post-editing time could be very different if one of the videos had more speech time than the other.

To take this into account, a different metric can be used, based on the amount of text in the video (if this data is available to us, that is, if a transcription is already available), the *words per post-editing hour*, where we measure the ratio between the length of the source subtitles in words and the time devoted to correcting the automatic translation:

$$\text{Words per post-editing hour} = \frac{\text{Length of the source subtitles (in words)}}{\text{Post-editing time (in hours)}} \tag{2.5}$$

In this case, the more words per post-editing hour, the less effort required for correction, and thus, the higher the quality of the automatic translation.

Based on the definition of these two metrics of post-editing effort, we can derive the formula to convert between them:

$$\text{Words per post-editing hour} = \frac{\text{Length of the source subtitles (in words)}}{\text{RTF} \times \text{Video length (in hours)}} \tag{2.6}$$

2.2.3 Relation and balance between automatic and human evaluations in MT

As already mentioned earlier on in this Section 2.2, while there remain important caveats as to the correspondence between automatic metrics and human judgement of

[9] This is an application to post-editing of the concept of Real Time Factor from automatic speech recognition: the ratio between the time devoted to applying ASR to an audio and the duration of that audio; or, generalizing, the ratio between the time devoted to processing an audio and the duration of that audio.

MT output, the most commonly used automatic metrics such as BLEU and TER have been shown to have a strong correlation with human judgement; i.e., it is generally the case that the better the automatic measure of an MT system, the better its human judgement.

At the current stage in the evolution of MT technology, we are at a point in which MT systems are rapidly improving thanks to advances in technology (both in ML techniques and in hardware) and to the use of larger and better linguistic corpora as training data. In accordance with this, automatic metrics such as BLEU and TER are registering constant increases for reference MT tasks, measuring improvements in MT output that are clearly noticeable for a human evaluator.

Thus, while automatic metrics do not provide us with a fine-grained evaluation of the linguistic reasons behind the improvement of MT outputs, at this stage improvements are so constant and clear that automatic metrics are being a very useful indicator of which direction to follow during experiments on the improvement of MT systems.

At the same time, even though more fine-grained human evaluations are less immediate and more costly, they are also providing additional information to better understand the current NMT models and to guide decisions on the refinement of NMT architectures and techniques. This more fine-grained information will probably become even more important in case the stream of improvements in MT results reaches a new plateau; in such a situation, automatic metrics become less useful in distinguishing between the quality of MT outputs, and they do not provide information about what aspects of the MT outputs compared might make one better than the other (and so, about where attempts at improvement could be focused).

In Chapter 3, we will study an example of a typical balance between automatic metrics and human evaluations of MT in a large project: day to day work on MT is guided by automatic evaluations, supplemented at specific points in time by several types of human evaluations that provide valuable additional information on the evolution of the MT systems being developed.

2.2.3.1 Correlation between automatic and human evaluations in MT

The expected correlation between automatic metrics and human evaluations of MT output is also something that can be used to confirm whether evaluations are being performed correctly, and to detect points that require further analysis.

When both automatic metrics and human evaluations are used to measure the evolution of an MT system, we can measure if there is a statistically significant correlation between the automatic and human measures that we produce, which would be the expected outcome. This can be computed using an appropriate method such as the Pearson correlation coefficient [45]. If no correlation is found, then this can be an indication that a more detailed analysis is necessary (to check whether measurements and evaluations are actually being correctly made, or whether additional insight can be gained on what is it that is causing an unexpected difference between the results of automatic and human evaluations).

2.2.4 Time and space efficiency

When comparing different MT systems, their output is not the only thing that we can evaluate. We can also measure their efficiency in terms of time and space complexity.

There are two points at which we can measure an MT system's efficiency. One is model training, and the other is inference or decoding (the application of the trained model to generate a translation)[10].

In both cases, we can empirically measure time efficiency by measuring how much time the process takes, and space efficiency by measuring how much RAM/video RAM is used during the process. We can also perform a theoretical analysis of our MT system to determine how efficient it is in terms of model parameters[11], MT techniques used, software implementation, and amount of training data (size of the parallel and monolingual corpora)[12].

As a relevant example, the Transformer (multi-headed self-attentional DNN) [84] has become very popular as a state-of-the-art NMT architecture in large part because it provides high-quality MT output with high time and space efficiency (it can reach comparable levels of quality using less parameters than other MT models).

Time and space efficiency for inference

When considering the application of an MT system, it is the system's efficiency for inference that we would first think about. That is, how fast the system is in generating translations, and how much computer memory will it require. In actual application scenarios, it can be important for an MT system to be quick in producing translations, so that a quicker system could be considered better than a slower one, even at the expense of some output quality. A system that is less time-efficient for inference means that the user will have to wait longer for a result, and/or that more hardware resources (CPU, GPU) will be needed to achieve an acceptable response time. Similarly, space efficiency is important in that it will determine how much RAM or video RAM is needed to generate translations (which can mean needing more hardware resources, or making the system unusable in a given hardware setup).

Time and space efficiency for MT model training

As to efficiency for MT model training, this is a process that might be done less frequently in an actual practical scenario. Training a large, high-quality MT system is generally expected to be quite costly in time and in hardware resources (both processing and memory), e.g., it could take days or weeks on relatively powerful computer hardware. Indeed, when training an MT system for actual use, we will usually give it as much time and as many hardware resources as we can, in order to maximize output quality (here, we will not usually sacrifice expected output quality

[10] *Inference* is the more general term in ML for the task of applying a trained statistical model to make predictions, while *decoding* is the usual term for this task in the subfield of MT.

[11] A model with more parameters can better represent complex realities, but at the cost of a higher time and space complexity.

[12] Using larger linguistic corpora usually implies that the MT models will have larger space requirements.

lightly just to accelerate the process). Nevertheless, this does not mean that model training efficiency is not important when evaluating an MT system. A higher time and space efficiency means developing, maintaining and improving MT systems becomes accessible for researchers and companies with less resources (CPU/GPU, RAM/video RAM). Being able to quickly train new MT models is also important in that it speeds up the process of trying new ideas in order to improve the MT system. Additionally, efficient model training can be important if we expect to perform frequent MT system adaptations or to apply intelligent interaction protocols (implying model retraining).

2.3 Tools for machine translation

Current research in the fields of machine learning and natural language processing with deep neural networks (DNNs) requires appropriate hardware and software resources, as well as large amounts of data (text corpora) for MT model training.

Regarding *hardware*, a powerful computing infrastructure with high-performance GPUs is required, to allow for the execution of multiple experiments in parallel, each of them composed of massively parallel computations. Nevertheless, it is more and more the case that high-end standard consumer computer hardware can provide the performance required (lowering the entry barrier to participate in state of the art research).

As to *software*, fortunately there are many free software tools and environments released and maintained by top research groups (both in industry and in academia) which are used assiduously for research and for commercial applications. This software is frequently developed in the C/C++ and Python programming languages (the focus usually being on computational efficiency and modifiability).

Using these frameworks, libraries and software toolkits for one's own research is very convenient, as the researcher thus avoids having to reimplement the basic techniques and can focus on experimenting with new ideas to go beyond the state of the art. However, one should be careful not to become overly dependent on any single framework, as this is a field in constant evolution and it is not infrequent for popular software to be abandoned, which then forces the researcher to move to a different framework.

In this section, we will identify some of the most relevant hardware and software tools in the fields of machine translation and natural language processing. Firstly, we will explain the typical hardware needs. Secondly, we will talk about software tools for machine translation, including: software for phrase-based machine translation; software for neural machine translation (deep learning frameworks, specific software tools for NMT, and deep learning GPGPU frameworks and libraries); software for language modelling (which are also part of the toolset necessary for phrase-based machine translation); and software for machine translation evaluation. Finally, we will briefly talk about the data available for training machine translation models.

In each subsection below, we have listed items in alphabetic order (except where a different order is explicitly indicated). As to the description for each item, we have not limited ourselves to their official description, but we have compiled our own descrip-

tions from their websites, scientific articles and technical papers, as well as reliable tertiary sources, trying to highlight for each tool some of their key distinguishing features. In some cases, we have also added information based on our own experience in working with them.

2.3.1 Hardware for machine translation

Computing (i.e., "training" or "learning") the statistical models on which statistical machine translation (including phrase-based machine translation and neural machine translation) is based requires powerful computer hardware, although in recent times it is more and more the case that high-end standard consumer computer hardware can provide the performance required (which is positive in that it makes it easier for smaller research groups to participate in state of the art research).

In the 2000s and early 2010s, training PBMT systems required *high-performance CPUs* (i.e., multi-core processors with clock rates over 2 GHz) and *large amounts of RAM* (i.e., 32 GB and over, to load in RAM full statistical models based on hundreds of millions of words of text). Parallel computing began to be exploited up to some point as multicore CPUs became more and more common.

In the late 2010s, the hardware setup required for state-of-the-art machine translation changed when neural machine translation became the paradigm with the best results. Nowadays, research and innovation in the fields of natural language processing is being driven by advances in deep neural network technologies that have been enabled by the use of modern GPUs. General-purpose computing on graphics processing units (GPGPU) enables running the complex matrix computations required for deep neural network model training in a massively parallel way on thousands of computing cores, speeding up computations by several orders of magnitude compared to the few cores offered by standard CPUs, all while still working with regular high-end consumer computer hardware.

Thus, current hardware for state-of-the-art machine translation includes, in addition to high-performance CPUs and large amounts of RAM, *high-performance GPUs with large amounts of GPU memory* (i.e., GPUs with over 3000 cores at a clock rate over 1.5 GHz with over 10 GB of memory).

A typical hardware setup for research groups in machine translation nowadays may be a computer cluster with tens of high-performance CPUs and GPUs, with large amounts of primary memory (both RAM and GPU memory). This allows for running multiple experiments in parallel, each of them composed of massively parallel computations.

2.3.2 Software for machine translation

2.3.2.1 Phrase-based machine translation software

These are the most relevant software tools for training and running phrase-based machine translations systems.

Development of software for PBMT was at its peak during the 2000s and early 2010s (when this MT paradigm was itself at its best moment). It has slowed down in

the late 2010s, as the focus of research and development has turned towards neural machine translation.

We can say that the reference software for PBMT is Moses, listed as the first item in this section, which can handle most of the tasks required for training and running PBMT models. However, Moses is not the only relevant PBMT software tool available. There are some software tools that are designed to handle a specific task of the PBMT pipeline, of which we include as a relevant example the alignment model software GIZA++ (as it has been used together with Moses for a long time). And there are other comprehensive PBMT toolkits that have tried to work as an alternative to Moses with their own distinctive features, of which we include as examples Jane and Thot.

Note also that one of the components of PBMT systems, the language model, is trained with separate specific software which is covered in Section 2.3.2.5 (although some of these separate language model toolkits have been integrated into PBMT toolkits).

Moses
Website: `http://www.statmt.org/moses/`
Licence: Free software (GNU LGPL 2.1-or-later)
Latest release: 4.0 (Oct 2017)

Moses is a free software statistical machine translation engine, including many of the software tools required for training SMT models and using them to produce automatic translations (*decoding*), and also for evaluating the output (among other secondary tasks).

Developed by an international community of top researchers in machine translation, since its release in the late 2000s Moses has become the reference software for phrase-based machine translation, having been used extensively for research and industrial applications.

In addition to phrase-based machine translation, Moses supports other PBMT-related MT techniques in which the models incorporate additional linguistic information: syntax-based machine translation, hierarchical phrase-based machine translation and factored machine translation.

Regarding parallelism, for the decoding step, the Moses2 decoder (released in 2016) supports multithreading.

While Moses provides most of the software tools required to train and run an SMT system, it used to rely on external tools mainly for two steps of the process: computing the word alignments from the parallel corpus (which can be done, e.g., with GIZA++, see below) and training language models (which can be done with the software we explain in Section 2.3.2.5). The language model training software KenLM was included into Moses in 2010 to cover this second missing step (although users still have the option to use other external language model software instead).

GIZA++
Website: `https://github.com/moses-smt/giza-pp`
Licence: Free software (GNU GPL 2.0-or-later)
Latest release: v2 (2003)

GIZA++ is a free software statistical machine translation toolkit that is used to train IBM alignment models (1–5) and an HMM word alignment model. It was developed by F.J. Och at RWTH Aachen University.

Moses relies on Giza++ (or other equivalent software) for computing the word alignments from the parallel corpus.

Regarding parallelism, mgiza[13] is an extended version of GIZA++ supporting multi-threading, resume training and incremental training.

Jane
Website: `http://www-i6.informatik.rwth-aachen.de/jane/`
Licence: Open source (RWTH Jane Licence)
Latest release: 2.3 (Feb 2014)

Jane is RWTH Aachen University's open source statistical machine translation toolkit. It supports state-of-the-art techniques for phrase-based and hierarchical phrase-based machine translation, as well as for system combination. Many advanced features are implemented in the toolkit, such as forced alignment phrase training for the phrase-based model and several syntactic extensions for the hierarchical model.

Jane, developed by the RWTH's i6 Human Language Technology and Pattern Recognition research group, has a long history and it has been used with success in numerous MT evaluation campaigns.

Thot
Website: `http://www.statmt.org/moses/`
Licence: Free software (GNU LGPL 3.0-or-later)
Latest release: 3.2.0 Beta (Jul 2017)

Thot is a free software toolkit for statistical machine translation. Thot incorporates tools to train PBMT models, a state-of-the-art PBMT decoder, as well as tools to estimate all of the models involved in the translation process. In addition to this, Thot is able to incrementally update its models in real time after presenting an individual sentence pair using online learning (also known as adaptive machine translation). Thot was created by Daniel Ortiz Martínez at Universitat Politècnica de València.

According to its author, Thot has a strong focus on online and incremental learning, and so it includes its own programs to carry out language and translation model estimation. Specifically, Thot includes tools to work with n-gram language models based on incrementally updateable sufficient statistics. Also, Thot includes a set of tools and a software library to estimate IBM 1, IBM 2 and HMM-based word alignment models, using batch and incremental EM algorithms (avoiding depending on

[13] *Website:* `https://github.com/moses-smt/mgiza`; *Licence:* Free software (GNU GPL 2.0-or-later); *Latest release:* 2010.

GIZA++ for this functionality). Other relevant features of this toolkit are that it incorporates interactive machine translation functionality, and a stable and robust translation server.

Other PBMT software

Apache Joshua SMT decoder
`https://cwiki.apache.org/confluence/display/JOSHUA/`

Stanford Phrasal: A Phrase-Based Translation System
`https://nlp.stanford.edu/phrasal/`

UCAM-SMT: The Cambridge Statistical Machine Translation System
`https://ucam-smt.github.io/`

2.3.2.2 Deep learning frameworks

Deep learning (DL) has revolutionized many of the fields of machine learning in the 2010s, including machine translation in the late 2010s. DL frameworks offer high-level programming interfaces devised specifically to make it easier to design, train and apply deep neural networks (DNNs) for machine learning. DL frameworks typically use underneath GPGPU libraries to train and run DNNs in a massively parallel way, making this as transparent as possible for their users (as we will see below in Section 2.3.2.3).

Here we list some of the most relevant DL frameworks, on top of which specific software for neural machine translation can be built (as we will see right below in Section 2.3.2.4). We focus especially on the frameworks that have been used or tried as part of the work of this master's thesis. We also include here, as a last item, ONNX, an initiative for a common machine learning model format to allow for models to be exchanged between different ML/DL frameworks.

Apache MXNet
Website: `https://mxnet.apache.org/`
Licence: Free software (Apache Licence 2.0)
Latest release: 1.7.0 (Jul 2020)

Apache MXNet is a free software deep learning framework used to train and deploy deep neural networks. It allows mixing symbolic and imperative programming. MXNet contains a dynamic dependency scheduler that automatically parallelizes both symbolic and imperative operations on the fly, with a graph optimization layer on top to make symbolic execution faster and more memory efficient. MXNet can scale to multiple GPUs and multiple machines.

MXNet is integrated into Python, and also supports Scala, Julia, Clojure, Java, C++, R and Perl.

MXNet is the deep learning framework of choice for Amazon.

PyTorch
Website: `https://pytorch.org/`
Licence: Free software (Modified BSD Licence)
Latest release: 1.5.1 (Jun 2020)

PyTorch is a free software machine learning framework, used for applications such as natural language processing and computer vision. PyTorch provides two high-level features: tensor computation with GPU acceleration; and dynamic deep neural networks built on a tape-based autodiff system.

PyTorch is integrated into Python, and also offers a C++ interface.

PyTorch is primarily developed by Facebook's AI Research lab (FAIR).

TensorFlow
Website: `https://www.tensorflow.org/`
Licence: Free software (Apache Licence 2.0)
Latest release: 2.2.0 (May 2020)

TensorFlow is a free software end-to-end machine learning framework.

TensorFlow (1.x) computations are expressed as stateful dataflow graphs (the name TensorFlow derives from the operations that such neural networks perform on multidimensional data arrays, which are referred to as "tensors"). TensorFlow 2.0, released on September 2019, introduced eager execution as the new default programming approach, as an alternative to the previous static computational graph–approach. In these recent releases, an effort has also been made to incorporate a high-level API (Keras) to make programming neural networks simpler. TensorFlow can run on multiple CPUs and GPUs (with optional CUDA and SYCL extensions for general-purpose computing on graphics processing units).

TensorFlow is one of the most successful ML/DL frameworks. It was released as free software by Google in 2015, making it a pioneer in its class. The ML community was quick to adopt TensorFlow for DL developments, and Google has always very actively developed the framework, also releasing frequently their ML research results implemented in TensorFlow for ML researchers to tinker with.

TensorFlow primarily supports Python and C++, and also offers APIs for JavaScript, Swift, Java and Go. Community-developed APIs are also available for other languages, including C#, Haskell, Julia, Matlab, Ruby, Rust and Scala.

TensorFlow was created and is primarily developed by Google.

Theano
Website: `http://www.deeplearning.net/software/theano/`
Licence: Free software (Modified BSD Licence)
Latest release: 1.0.4 (Jan 2019)

Theano is a Python library that allows one to define, optimize, and evaluate mathematical expressions involving multi-dimensional arrays efficiently. It can use GPUs and perform efficient symbolic differentiation.

Since its release in 2008 and until the release of TensorFlow in 2015, Theano was frequently used as a basis for programming in the fields of machine learning and

deep learning. However, in 2017 development of Theano was cancelled, and so ML researchers moved to other more recently released DL frameworks.

Theano was created by the Montréal Institute for Learning Algorithms (MILA) of Université de Montréal.

ONNX: Open Neural Network Exchange Format
Website: `https://onnx.ai/`
Licence: Free software (MIT/Expat Licence)
Latest release: 1.7.0 (May 2020)

ONNX, the Open Neural Network Exchange, is a free software artificial intelligence ecosystem, providing an open source format for AI models, both deep learning and traditional ML, to facilitate the exchange of models between different ML/DL frameworks. It defines an extensible computation graph model, as well as definitions of built-in operators and standard data types. Currently, it is focused on the capabilities needed for inferencing (scoring).

ONNX was introduced by Facebook and Microsoft, and currently has some degree of support also from Amazon, Baidu, IBM, Huawei and Nvidia (among other organizations).

ONNX models are currently supported in CNTK, MXNet, PaddlePaddle and PyTorch (among other DL frameworks). Converters are also in development for Tensorflow, among other DL frameworks.

Other DL frameworks

CNTK: The Microsoft Cognitive Toolkit
`https://docs.microsoft.com/en-us/cognitive-toolkit/`

DeepSpeed (DL optimization library for distributed training), by Microsoft
`https://github.com/microsoft/DeepSpeed`

DyNet: The Dynamic Neural Network Toolkit, by Carnegie Mellon Univ's LTI
`http://dynet.io/`

Keras (high-level DL API over TensorFlow, CNTK or MXNet)
`https://keras.io/`

NNEF: Neural Network Exchange Format, by the Khronos Group consortium
`https://www.khronos.org/nnef/`

PaddlePaddle: PArallel Distributed Deep LEarning, by Baidu
`https://github.com/PaddlePaddle/Paddle`

RETURNN: RWTH extensible training framework for universal RNNs (based on TensorFlow)
`https://github.com/rwth-i6/returnn`

2.3.2.3 Deep learning GPGPU frameworks and libraries

DL frameworks (presented above in this same section) typically use underneath GPGPU libraries to train and run DNNs in a massively parallel way, making this as transparent as possible for their users.

Currently, the main GPGPU framework used by most DL frameworks (and thus, by most NMT toolkits and libraries) is Nvidia's CUDA. This has two important negative effects: firstly, it limits this software to work exclusively on Nvidia GPUs (which enables Nvidia to hold a hardware monopoly in this field); and secondly, as CUDA has a proprietary licence, it is exclusively developed by Nvidia, and they could arbitrarily decide to halt development or to block anyone from using CUDA at any point (which would prevent all DL and NMT software from using GPGPU, until they spent the time necessary to develop an alternative).

The most promising open, vendor-agnostic alternative to CUDA seems to be OpenCL. However, at this point support for OpenCL is very limited among the most important DL frameworks.

CUDA
Website: `https://developer.nvidia.com/cuda-zone`
Licence: Proprietary (Nvidia Software License Agreement and CUDA Supplement)
Latest release: 11.0.194 (May 2020)

Nvidia CUDA (Compute Unified Device Architecture) is a parallel computing platform, programming model and toolkit for general-purpose computing on graphics processing units (GPGPU). It allows developers to use CUDA-enabled GPUs (i.e., only Nvidia GPUs since 2006) for GPGPU.

When using CUDA, developers can program in common languages such as C, C++, Fortran, Python and Matlab and express parallelism through extensions in the form of additional keywords (thus not requiring specialist skills in graphics programming). The CUDA platform is a software layer that gives direct access to the GPU's virtual instruction set and parallel computational elements, for the execution of compute kernels. CUDA provides both a low level API (CUDA Driver API) and a higher level API (CUDA Runtime API).

The CUDA Toolkit is designed for the development of GPU-accelerated applications. The CUDA Toolkit includes GPU-accelerated libraries, a compiler, development tools and the CUDA runtime.

All of the DL frameworks listed above in this same section support (and require) the use of CUDA for GPGPU.

CUDA-X Deep Learning Libraries: cuDNN and TensorRT
Website: `https://developer.nvidia.com/gpu-accelerated-libraries`
Licence: Proprietary (cuDNN Software License Agreement + TensorRT Software License Agreement + Nvidia Registered Developer Agreement)
Latest release: cuDNN 7.6.5, TensorRT 7.1.3 (Jun 2020)

Nvidia CUDA-X, built on top of CUDA, is a collection of libraries, tools and technologies that deliver higher performance compared to CPU-only alternatives across

multiple application domains.

cuDNN, the Nvidia CUDA Deep Neural Network library, is a GPU-accelerated library of primitives for the training of deep neural networks. cuDNN provides implementations for standard routines such as forward and backward convolution, pooling, normalization, and activation layers.

Nvidia TensorRT is a software development kit for high-performance deep learning inference. It includes a deep learning inference optimizer and runtime with the goal of delivering low latency and high-throughput for deep learning inference applications. TensorRT provides INT8 and FP16 optimizations for deep learning inference applications such as video streaming, speech. Reduced-precision inference significantly reduces application latency, which is a requirement for many real-time services, auto and embedded applications.

While CUDA can be downloaded openly from Nvidia's website, cuDNN has to be downloaded separately and requires previous registration in the "Nvidia Developer Program", which carries with it the acceptance of additional terms and conditions supposes an additional barrier to access (registration is currently free of cost, but Nvidia has the ability to change this at any time).

All of the DL frameworks listed above in this same section support the use of cuDNN and TensorRT for GPGPU.

OpenCL: Open Computing Language
Website: `https://www.khronos.org/opencl/`
Licence: OpenCL Specification licence
Latest release: 3.0 (Apr 2020)

OpenCL (Open Computing Language) is a multi-vendor, open, royalty-free standard for general purpose parallel programming of heterogeneous systems including CPUs, GPUs, and other processors. It is maintained by the Khronos Group consortium.

OpenCL supports several different applications through a low-level, high-performance, portable abstraction. It is intended to be the foundation layer (in the same vein as the CUDA Driver API) of a parallel computing ecosystem of platform-independent tools, middleware and applications.

OpenCL consists of an API for coordinating parallel computation across heterogeneous processors; and a cross-platform intermediate language with a well-specified computation environment.

OpenCL also includes an OpenCL C++ compiler reference implementation and an OpenCL C++ standard library reference implementation, both of them released as free software.

OpenCL support has been added to various degrees to the following DL frameworks (and the NMT software based on them): Keras, PlaidML, PyTorch, TensorFlow, Theano.

SYCL
Website: `https://www.khronos.org/sycl/`
Licence: SYCL Specification licence
Latest release: 1.2.1 revision 7 (Apr 2020)

SYCL (C++ Single-source Heterogeneous Programming for OpenCL) is a royalty-free, cross-platform abstraction C++ programming model for OpenCL. It is maintained by the Khronos Group consortium.

SYCL is designed to be as close to standard C++ as possible (SYCL 1.2.1 builds on the features of C++11, with additional support for C++14 and C++17). It includes templates and generic lambda functions to enable higher-level application software to be coded with optimized acceleration of kernel code across OpenCL 1.2 implementations. Developers program at a higher level than OpenCL C or C++ (in the same vein as the CUDA Runtime API), but have access to lower-level code through integration with OpenCL C/C++ libraries and frameworks.

SYCL is the way by which support for OpenCL has been added to the DL framework TensorFlow (as an alternative to the Nvidia-specific CUDA).

Other DL GPU/CPU frameworks and libraries

HIP: C++ Heterogeneous-Compute Interface for Portability, by AMD
`https://gpuopen.com/compute-product/hip-convert-cuda-to-portable-c-code/`

hipSYCL: an implementation of SYCL over Nvidia CUDA/AMD HIP, by AMD
`https://gpuopen.com/compute-product/hipsycl/`

Metal, by Apple
`https://developer.apple.com/metal/`

OpenACC: Open Accelerators API, by the OpenACC.org consortium
`https://www.openacc.org/`

OpenMP: Open Multi-Processing API specification for parallel programming, by the OpenMP Architecture Review Boards consortium
`http://openmp.org/`

PlaidML: A framework for making deep learning work everywhere, by Intel
`https://www.intel.ai/plaidml/`

ROCm: Radeon Open Compute Platform for HPC and UltraScale GPU Computing, by AMD
`https://rocm.github.io/`

2.3.2.4 Neural machine translation software

On top of the general DL frameworks presented above, several libraries and toolkits have been released implementing specific neural network architectures and techniques for neural machine translation.

Here we list some of the most relevant ones, focusing especially on the ones that have been used or tried as part of the work of this master's thesis. For each item in this list, next to the name of the library we indicate in parentheses the underlying DL framework.

Fairseq (PyTorch)
Website: `https://ai.facebook.com/tools/fairseq/`
Licence: Free software (MIT/Expat Licence)
Latest release: 0.9.0 (Dec 2019)

Fairseq is the Facebook AI Research (FAIR) Sequence-to-Sequence Toolkit, based on PyTorch. It is a sequence modelling toolkit for training custom models for machine translation, summarization, language modelling and other text generation tasks.

Fairseq provides reference implementations of various sequence-to-sequence models, including Transformer self-attentive networks, Long Short-Term Memory (LSTM) networks and a novel convolutional neural network (CNN) that can generate translations many times faster than comparable recurrent neural network (RNN) models. Fairseq also features multi-GPU training on one or across multiple machines, fast beam search generation on both CPU and GPU, large mini-batch training even on a single GPU via delayed updates, and mixed-precision training (faster training with less GPU memory).

Fairseq is currently one of the most useful NMT toolkits. It is very actively developed, implementing up-to-date NMT architectures, and including features to make for more efficient training and running of neural networks. Fairseq's code is relatively easy to understand and modify, and it is based on PyTorch, a DL framework that is also efficient and very actively developed.

OpenNMT (PyTorch, TensorFlow)
Website: `http://opennmt.net/`
Licence: Free software (MIT/Expat Licence)
Latest release: OpenNMT-tf 2.11.1 / OpenNMT-py 1.1.1 (Jun 2020)

OpenNMT is a free software framework for neural machine translation and neural sequence learning, created by the Harvard NLP group and Systran. Originally based on Torch (now abandoned), currently OpenNMT has implementations on PyTorch and TensorFlow.

While OpenNMT initially focused on standard sequence to sequence models applied to machine translation, it has been extended to support many additional models and features, for ML tasks including image to text, language modelling, sequence classification, sequence tagging, sequence to sequence, speech to text and summarization. For machine translation, it implements some of the most important NN architectures, including Transformer, RNN with attention and ConvS2S.

Sockeye (MXNet)
Website: `https://github.com/awslabs/sockeye`
Licence: Free software (Apache Licence 2.0)
Latest release: 2.1.13 (Jul 2020)

Sockeye is a sequence-to-sequence statistical modelling framework with a focus on neural machine translation, based on Apache MXNet. It was created by Amazon Research.

Sockeye implements key MT encoder-decoder architectures, including Transformer models with self-attention [84], deep recurrent neural networks with attention [11] and fully convolutional sequence-to-sequence models [31]. Sockeye also supports a wide range of optimizers, normalization and regularization techniques, and inference improvements from recent NMT literature.

At the time of WMT18 (Chapter 4), Sockeye implemented the state of the art models for NMT. Sockeye's code, written in Python and based on MXNet, is relatively easy to understand and modify, which made it a good choice at that point in order to research with Transformer models. Since then, it seems to have lagged a little bit behind other NMT toolkits, not having incorporated the most recent advances published in 2018 and 2019, although recently it seems to have picked up the pace somewhat with the addition in version 2.0 of important features such as low-precision training.

Tensor2Tensor (TensorFlow)
Website: `https://github.com/tensorflow/tensor2tensor`
Licence: Free software (Apache Licence 2.0)
Latest release: 1.15.7 (Jun 2020)

Tensor2Tensor is a free software library of deep learning models and datasets, including models and tools of special interest for neural machine translation, based on TensorFlow. It was created by the Google Brain team.

Some of its features are: many state of the art and baseline models are built-in and new models can be added; many datasets across modalities (text, audio, image) are available for generation and use, and new ones can be added; support for multi-GPU machines and synchronous and asynchronous distributed training.

Tensor2Tensor was the software on which the original implementation of the Transformer self-attentional neural network architecture for NMT was released by Google. This made it very interesting for MT researchers. Nevertheless, several other NMT libraries were quick to implement this architecture and further refinements.

Tensor2Tensor is designed to be a powerful deep learning library, with the aim of being applicable to any field of machine learning. From our experience with it, it could be said that this focus on being general rather than specific makes its structure more abstract, making its functionalities not so simple to understand and modify. Scarce documentation is also a factor that complicates using and modifying Tensor2Tensor.

In October 2019, a successor library was released by the Google Brain team to supersede Tensor2Tensor: Trax[14]. Information and documentation for this new library

[14] *Website:* `https://github.com/google/trax`; *Licence:* Free software (Apache Licence 2.0);

is scarce at the moment.

Other NMT software

MarianNMT (C++), by Microsoft Translator and EdinburghNLP
`https://marian-nmt.github.io/`

Nematus (Tensorflow), by EdinburghNLP
`https://github.com/EdinburghNLP/nematus`

Nvidia OpenSeq2Seq: distributed and mixed precision training of sequence-to-sequence models (TensorFlow)
`https://github.com/NVIDIA/OpenSeq2Seq`

xnmt: eXtensible Neural Machine Translation (DyNet), by Carnegie Mellon University's LTI
`https://github.com/neulab/xnmt`

2.3.2.5 Language modelling software

A statistical language model (LM) is a probability distribution over sequences of words. Given a sequence of words, it assigns a probability to the whole sequence (meaning "this is the probability of this sequence of words being used in this language (according to the data on which the language model was trained)").

LMs are have important applications in many fields of natural language processing. In the case of machine translation, LMs are one of the main components in PBMT systems. Thus, LM software has always been a part of the toolset necessary for PBMT (see Section 2.3.2.1 above for PBMT software).

Here we list some of the most relevant software for statistical language modelling, beginning with more traditional n-gram-based[15] LM software, and then adding more recent DNN-based LM software (which is the current state of the art).

KenLM
Website: `https://kheafield.com/code/kenlm/`
Licence: Free software (GNU LGPL 2.1-or-later)
Latest release: May 2020

KenLM is a free software statistical language model toolkit that estimates, filters, and queries n-gram language models. It was created by Kenneth Heafield at the University of Edinburgh.

KenLM estimates from text unpruned language models with modified Kneser-Ney smoothing; estimation is done on disk, in a fast and scalable way, using streaming algorithms co-developed by the toolkit's author. Regarding LM querying, it is faster and lower-memory than with other contemporary LM toolkits. As to LM filtering,

Latest release: 1.3.1 (Jul 2020).
[15] See Section 2.2.1 for a brief introduction to the concept of n-gram.

KenLM filters LMs to test sets, that is, n-grams are removed if they cannot be generated during decoding (in a non-lossy process).

KenLM was released in 2010, and at the same time it was incorporated into the leading PBMT toolkit Moses (which had depended until that point on other LM toolkits listed in this section, which could not be included into Moses for licence reasons). Other PBMT toolkits (including Apache Joshua, Jane, and Stanford Phrasal) followed in Moses' footsteps and also included KenLM as their default language model training software.

SRILM
Website: `http://www.speech.sri.com/projects/srilm/`
Licence: Open source (SRILM Research Community Licence 1.1)
Latest release: 1.7.3 (Sep 2019)

SRILM is a toolkit for building and applying statistical n-gram language models (LMs). It has been under development in the SRI Speech Technology and Research Laboratory since 1995.

SRILM consists of the following components: a set of C++ class libraries implementing language models; a set of executable programs built on top of these libraries to perform standard tasks such as training LMs and testing them on data; a collection of miscellaneous scripts facilitating minor related tasks.

CUED-RNNLM
Website: `http://mi.eng.cam.ac.uk/projects/cued-rnnlm/`
Licence: Free software (Modified BSD Licence)
Latest release: 1.1 (Nov 2017)

CUED-RNNLM is a free software statistical DNN-based language model toolkit offering an implementation of RNNLM training (on GPU) and efficient evaluation (on CPU). It is developed by Xie Chen at the University of Cambridge.

Some of its main features are: improved RNNLM training criteria (Cross Entropy, Variance Regularization, Noise Contrastive Estimation); efficient RNNLM training on GPU using spliced sentence bunch; support for various deep structures, including su-RNNLMs; RNNLM evaluation via perplexity or N-best rescoring; random sampling of sentences up to a specified number of words from a trained RNNLM (for n-gram LM interpolation).

Other LM software

IRSTLM: The FBK IRST n-gram Language Modelling Toolkit
`https://hlt-mt.fbk.eu/technologies/irstlm`

2.3.2.6 Machine translation evaluation software

For the most successful evaluation methods, we can find free software implementations that we can use so that we don't need to implement ourselves the software from the method's original publication.

However, performing automatic or manual MT evaluations is never as straightforward as one might imagine. The way in which we preprocess and postprocess the input and output data and the reference translations, the way in which we train our systems, the specific implementation of the evaluation method that we are using... We must be very clear about each one of these aspects to make sure that our evaluation will be reproducible and our results will be comparable to the current reference and for future researchers.

Here is just a brief list of examples of available MT evaluation software. Among them, sacreBLEU is especially interesting, since it is a recent effort intending to facilitate performing BLEU evaluations keeping every aspect of the evaluation under control, so that we and everyone can be sure of how the evaluation is performed, which reference data is used, how the data is processed... Apart from that, mteval is one of the classic references for computing BLEU scores (used in many MT evaluation campaigns), but in this case that is all it does, it is up to the user to follow strictly the appropriate evaluation protocols so that the results are comparable. TER COMpute is a classic piece of software along the lines of mteval but for TER. Finally, we have included TLP: The transLectures-UPV Platform as an example of a post-editing platform that can be used for human post-editing and evaluations, keeping track of post-editing times as one of the measures of MT quality.

sacreBLEU
Website: `https://github.com/mjpost/sacreBLEU`
Licence: Free software (Apache Licence 2.0)
Latest release: 1.4.12 (Jul 2020)

mteval (NIST BLEU)
Website: `https://www.nist.gov/itl/iad/mig/tools`
Licence: Public domain
Latest release: 14c (Nov 2016)

TER COMpute
Website: `http://www.cs.umd.edu/~snover/tercom/`
Licence: Open source
Latest release: 0.7.25 (Aug 2008)

TLP: The transLectures-UPV Platform
Website: `https://aplicat.upv.es/exploraupv/ficha-tecnologia/patente_software/15324`
Licence: Free software (Apache Licence 2.0)
Latest release: 3.7.8 (Apr 2020)

Other MT evaluation software

Moses multi-bleu
`https://github.com/moses-smt/mosesdecoder/tree/master/scripts/generic`

Moses multi-bleu-detok
`https://github.com/moses-smt/mosesdecoder/tree/master/scripts/generic`

WER++
`https://github.com/nsmartinez/WERpp`

2.3.3 Data for machine translation

Finally, a very brief section on text corpora or datasets for machine translation. While monolingual datasets are easy to gather and easy to obtain, parallel bilingual corpora have been traditionally much more scarce. Recently there have been some large-scale efforts such as ParaCrawl to automatize and refine parallel data crawling from the web, which is producing parallel data in many languages (big and small) and in amounts much more vast than we were used to with more traditional, more manually crafted parallel corpora.

We list here just a few other examples of important sources of MT data. The most important yearly MT evaluation campaigns, such as WMT, are a great source of MT data, since they usually try and gather as many quality data as possible for their participants, and they usually produce new high-quality test sets every year. International organizations are also a good source: in this list we can see the UN multilingual corpus, while included in the WMT corpus we will find the classic Europarl corpus (with parallel data in every official language of the EU). Finally, we include Europarl-ST as an example of an ambitious corpus combining acoustic data and parallel text data gathered from publicly available European Parliament session recordings; the aim of this corpus is to be a reference corpus for Spoken Language Translation, that is, language processing from an acoustic signal in one language to an automatic text translation in another language.

WMT20 News Translation corpora
Website: `http://www.statmt.org/wmt19/translation-task.html`
Latest release: 2020

WMT20 Biomedical Translation corpora
Website: -
Latest release: 2020

ParaCrawl
Website: `https://paracrawl.eu/`
Licence: Creative Commons CC0
Latest release: 6.0 (Sep 2020)

United Nations Parallel Corpus
Website: `https://conferences.unite.un.org/UNCorpus`
Latest release: 1.0 (May 2016)

Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates
Website: `https://www.mllp.upv.es/europarl-st/`
Latest release: 1.0 (Nov 2019)

Chapter 3

Evaluating automatic transcriptions and translations in a real scenario

In this chapter, we will study how to evaluate the quality and usefulness of machine translation (MT) systems in several ways. To do this, we will review the work done in this respect within the EU research project transLectures (which the author of this master's thesis was a part of). This work was carried out in a real scenario, the poliMèdia video lecture repository of Universitat Politècnica de València[1].

While the focus of this master's thesis is on MT, we will also cover in this chapter part of the work done on evaluation of automatic speech recognition (ASR), given that ASR is the first step in the workflow of producing multilingual subtitles for video lectures using MT. In this setting, ensuring high ASR quality is essential towards obtaining usable, cost-effective automatic translations using MT.

In transLectures, three types of evaluation were carried out periodically to monitor the progress in automatic transcription (ASR) and translation (MT) quality along the

[1] Part of the contents of this chapter (especially Sections 3.1 and 3.4.1) are adapted from 2 publications by the author of this master's thesis:

- J. D. Valor Miró, R. N. Spencer, A. Pérez González de Martos, G. Garcés Díaz-Munío, C. Turró, J. Civera, and A. Juan. Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures. *Open Learning*, 29(1):72–85, 2014.
  URL: `http://dx.doi.org/10.1080/02680513.2014.909722`
- J. D. Valor Miró, R. N. Spencer, A. Pérez González de Martos, G. Garcés Díaz-Munío, C. Turró, J. Civera, and A. Juan. Evaluación del proceso de revisión de transcripciones automáticas para vídeos Polimedia. In *Proc. of the I Jornadas de Innovación Educativa y Docencia en Red (IN-RED 2014)*, pages 272–278, València (Spain), 2014.
  URL: `http://hdl.handle.net/10251/54397`

span of the project: automatic evaluations, human evaluations by language experts, and human evaluations by users at the case study sites, both with internal and external users.

Among the evaluation methods employed, an important one is measuring post-editing times by experts or users, in order to confirm whether revising the automatic transcriptions and translations saves time and costs with respect to transcribing and translating from scratch. In the context of human evaluations by users, we will evaluate an intelligent interaction approach to post-editing in which the system first identifies which sections of an automatic transcription or translation are likely to contain the most errors (based on automatic confidence measures), and then presents only these sections to the user for correction; these corrections are then fed back into the system and used to re-train the underlying models with a view to avoiding the same errors in the future.

The structure of this chapter is as follows: In Section 3.1, we present the context and the evaluation framework for the different evaluation methods employed that we will review in this chapter. Then, the next sections will be devoted to reviewing how each of the evaluation methods was applied: Section 3.2 for automatic evaluations; Section 3.3 for human evaluations by language experts; and Section 3.4 for human evaluations by users, including the implementation and study of intelligent interaction approaches to post-editing, as well as hybrid approaches. Finally, in Section 3.5 we sum up this chapter's conclusions.

3.1 Introduction. Evaluation framework.

Online video lecture repositories are fast becoming a staple feature of the Internet and an everyday educational resource in higher education. Used to supplement traditional lectures, they are being incorporated into existing university curricula around the world, to enthusiastic responses from students [71]. In 2007 the Universitat Politècnica de València (UPV) implemented its poliMèdia lecture recording system for the cost-effective creation and publication of quality educational video content [78]. In the early 2010s, it had a collection of over 10 000 video objects created by more than 1300 lecturers, in part incentivized by the Docència en Xarxa (Networked Teaching) action plan to boost the use of digital resources at the university. It had also been successfully exported to other universities within Spain and South America.

In 2011, the UPV embarked on the EU-subsidized project transLectures [67] to develop innovative, cost-effective techniques for the transcription and translation of video lectures. This followed the recommendations of research underlining the importance of transcriptions being available for these video lectures [29], not only for the purposes of providing subtitles for non-native speakers or people with hearing impairments [85], but also to allow lecture content searches and other advanced repository management functions [60]. Users might also choose to watch the video lectures with subtitles in order to aid comprehension of a foreign language, even viewing the transcript and translation simultaneously to help with the language learning process itself. Subtitles can also mean that videos can be followed in both noisy and noise-restricted

environments.

As part of transLectures, automatic subtitles in Spanish, English and Catalan were made available for all videos in the UPV's poliMèdia repository and were continually updated as technologies were improved during the course of the project. These subtitles were provided using a state-of-the-art automatic speech recognition toolkit developed as part of the transLectures project, known as TLK: The transLectures-UPV toolkit [75], and the well-known phrase-based machine translation toolkit Moses [48]. These toolkits are able to create statistical models and use them to produce high quality automatic transcriptions and translations. Both toolkits were integrated into a complex and distributed system that automatically transcribed and translated the complete poliMèdia repository [68].

Evaluating these automatic multilingual subtitles as the technology improved was an important part of the transLectures project, and was key in order to measure how well current ASR and MT can provide a solution to the need for multilingual subtitles. This section presents the context and the evaluation framework for the different evaluation methods employed that we will review in this chapter. First, we present the transLectures project and the poliMèdia pilot case in which UPV worked directly to apply and evaluate automatic speech recognition and machine translation. Then, we describe the poliMèdia datasets that were the reference for evaluations, and the ASR and MT technologies that were used and evaluated in transLectures. Finally, we give an overview of the types of evaluation that were used in the project (which clearly reflect the different evaluation methods introduced in Section 2.2, including both automatic and human evaluations).

Preliminary concepts: **Automatic transcriptions, automatic translations and automatic multilingual subtitling**

In the context of this master's thesis, an *automatic transcription* is the output of automatic speech recognition, while an *automatic translation* is the output of machine translation.

Automatic multilingual subtitling is the process of obtaining the synchronized automatic transcription of a recording (in its original language) and, from it, synchronized automatic translations in other languages.

Regarding the difference between transcriptions and subtitles, in this master's thesis these terms are used with their simpler meanings. A *transcription* is a written representation of speech (in the original language); *subtitles* are a synchronized transcription of a recording that is segmented in some way (to be shown in synchronization with the recording), or the translation of the original language subtitles into other languages.

Through automatic speech recognition, we can obtain automatic transcriptions that are time-stamped at the word level. From this information, we can apply automatic segmentation to obtain automatic subtitles in the original language. And we can apply machine transla-

tion to obtain automatic translated subtitles in other languages (using the same segmentation as in the original language subtitles, or using different automatic segmentations for the translated versions).

Since the automatic transcriptions we are considering are already time-stamped as an output of automatic speech recognition, and we apply to them automatic segmentation as they are produced, at some points in this master's thesis we might see that the terms "transcription" and "subtitles" can be used somewhat interchangeably.

3.1.1 The project: transLectures

transLectures: Transcription and translation of video lectures (2011–2014) [67, 76] was an EU FP7 research project with the aim of developing innovative, cost-effective technologies for the automatic transcription and translation of large video lecture repositories[2]. Advanced techniques for Automatic Speech Recognition and Machine Translation were developed for use in this specific context.

The transLectures consortium included video lecture providers (users), experts in automatic speech recognition (ASR) and machine translation (MT) and professional transcription and translation providers, from four academic and three industrial partners (Universitat Politècnica de València, RWTH Aachen University, Institut Jožef Stefan, Knowledge 4 All Foundation, Deluxe Media Europe, European Media Laboratory GmbH, XEROX Research Centre Europe). The MLLP research group of Universitat Politècnica de València was the project's coordinating partner.

With online collections of video material fast becoming a staple feature of the Internet and a key educational resource, in transLectures, technologies and tools were developed to enable organizations to add *multilingual subtitles* to these videos, making their contents available to a much wider audience in a cost-effective and sustainable way. By adding multilingual subtitles, the content of these videos is made accessible for the deaf and hard-of-hearing and for non-native speakers. In addition, the text data generated can be used in combination with other pattern recognition technologies to enable advanced repository management functions, such as lecture search, classification, recommendation and plagiarism detection among others.

The main objectives of the project were the improvement of transcription and translation quality using *massive adaptation* (of automatic speech recognition and machine translation models) and *intelligent interaction* (to reduce the effort required for post-editing), and the *integration* of the technology developed in the Opencast platform to enable real-life evaluation.

These ideas were integrated and tested on two real-life case studies: VideoLectures.NET, an EU-based online repository with over 17 000 videos (at the time of the project) of talks given by top researchers in various academic settings, and the UPV's poliMèdia, described below.

[2] Automatic transcriptions and translations are considered *cost-effective* if they require less human intervention time to be produced and revised compared to producing transcriptions and translations from scratch

In this master's thesis we will focus on the work done on the poliMèdia pilot, which was where the UPV team worked directly.

3.1.1.1 Project languages and language pairs

3 languages and 6 language pairs were covered in the transLectures project:

ASR English, Spanish, Slovenian.

MT English↔Spanish, English↔Slovenian, English→French, English→German.

As we will see in the following Sections 3.1.2 and 3.1.3, only some of these languages and language pairs were part of the poliMèdia pilot, and thus will be part of the focus of this master's thesis: Spanish for ASR and Spanish→English for MT.

3.1.1.2 Key techniques for the improvement of ASR and MT in transLectures

Massive adaptation This is the process whereby general purpose ASR and MT models are adapted to the specific context using lecture-specific variables, such as speaker [50] and topic [55], in order to produce more accurate output. Text data extracted from the metadata and time-aligned presentation slides provided by the lecturers are also used to inform these new "in-domain" models.

Intelligent interaction User interaction is not new in this field. In transLectures, however, an intelligent approach was adopted in which the system first identifies which sections of a lecture's automatic transcription or translation contain the most errors [66] (that is, which sections, based on automatic confidence measures, are most likely to contain errors), and then presents only these sections to the user for correction. These corrections are then fed back into the system and used to re-train the underlying models with a view to avoiding the same errors in the future.

In Section 3.4.1 we will study the application of the intelligent interaction technique to post-editing.

3.1.2 The pilot: poliMèdia

poliMèdia [4, 77] is a service for the creation and publication of quality educational video content at the Universitat Politècnica de València. Launched in 2007, it is primarily designed to allow UPV lecturers to record pre-scripted short lectures (around 10 minutes long) for use by students to supplement the traditional live lecture. The videos are accompanied by time-aligned presentation slides. poliMèdia video lectures are published through the UPV Mèdia repository [3].

The video recordings are filmed at specialized studios under controlled conditions to ensure maximum recording quality and homogeneity. Lecturers are filmed against

a constant-colour background in order to be able to post-produce videos in which the lecturer is shown (properly scaled) next to the corresponding presentation slides.

In addition to the presentation slides, lecturers are requested to provide any metadata and additional textual resources related to the subject of the video lecture. These are used for domain adaptation during the automatic transcription process developed in transLectures, as part of which the entire poliMèdia repository is regularly retranscribed.

By the end of the transLectures project (2014), poliMèdia comprised a collection of over 10 000 video objects created by over 1 300 lecturers, in part incentivized by the Docència en Xarxa (Networked Teaching) action plan to boost the use of digital resources at Universitat Politècnica de València. This translated into a total running time of over 1 600 hours of educational material across a range of subjects taught at the UPV, all published online via the UPV Mèdia catalogue. In terms of uptake by the student body, in 2012 some 30 000 students accessed these videos.

Regarding languages, most of the poliMèdia video lectures are in Spanish, English or Catalan.

3.1.3 The poliMèdia datasets for ASR and MT

The initial poliMèdia video dataset used in project transLectures was created in November 2011 (at the beginning of the project). It was a copy of the live poliMèdia video repository at that time, comprising almost 6 500 videos accounting for 1 400 hours of video material, including videos mostly in Spanish, English and Catalan, and also some videos in other languages.

Initially, 106 hours of poliMèdia video lectures in Spanish had manual transcriptions available.

From this transcribed dataset, a speaker-independent[3] partition was defined for training, development and test of Spanish ASR systems, as shown in Table 3.1.

Table 3.1: poliMèdia Spanish monolingual dataset partition for ASR.

	Length (h)	Sentences	Words	Vocabulary
Train	99.0	41500	96.8K	28.0K
Dev	3.5	1401	37.8K	3.5K
Test	3.8	1139	32.1K	4.1K

For MT, 15 hours of the manually transcribed poliMèdia video lectures in Spanish where manually translated from Spanish into English (accounting for a total of ∼4 000

[3] A *speaker-independent* partition means that the speakers of the videos in the development and test partitions do not overlap, and do not appear in the training partition either. This is important in order to make sure that the system is not overfitting to work well with some specific speakers; instead, we test the system against speakers it has not seen in training, and so we can expect it to generalize better when we apply it to videos with other new speakers.

sentence pairs). From this, the partition for MT was defined keeping the same videos for dev and test as in the partition for ASR, as shown in Table 3.2.

Table 3.2: poliMèdia Spanish→English dataset partition for MT.

	Length (h)	Sentence pairs ES↔EN	Words		Vocabulary	
			ES	EN	ES	EN
Train	7.7	1529	40.3K	40.2K	4.7K	3.9K
Dev	3.5	1401	37.8K	38.7K	3.5K	3.7K
Test	3.8	1139	32.1K	32.1K	4.1K	3.3K

3.1.4 Automatic transcription and translation systems

In this section, we describe the main points of the automatic speech recognition and machine translation systems used in transLectures and their evolution during the project. The output of these systems is what was evaluated in the evaluation procedures described in the following sections in this chapter.

3.1.4.1 Automatic Speech Recognition systems

At the time of transLectures (2011–2014), automatic speech recognition (ASR) was a mature research field [32, 38], having experienced a major boost over the previous two decades. Typically, ASR systems were made up of two statistical models, the acoustic model and the language model (n-gram model). For model training, maximum likelihood and other discriminative techniques were applied to automatically estimate the parameters of these models using large corpora of audio and text. As for inference, the search process provided the most probable transcription hypotheses given the acoustic data. The performance of ASR systems was most frequently assessed using the Word Error Rate (WER) automatic metric.

Acoustic models

Taking into account recent progress on ASR, it was decided to focus on neural networks for acoustic modelling throughout the course of transLectures.

UPV replaced their original Gaussian HMM models with DNN-HMM hybrid models, leading to huge improvements in performance, and explored the use of convolutional DNNs and their combination with standard DNNs. Furthermore, UPV investigated the use of speaker adaptation techniques. In particular, the well-known fCMLLR technique led to significant improvements in performance. More recently, the softmax layer adaptation technique allowed UPV to apply extra speaker adaptation steps using DNN-HMM hybrid models.

Multilingual DNNs were also explored during the project. Using this approach, it is possible to train the hidden layers of a DNN with speech data from several languages. Multilingual modelling proved consistently useful and resulted in strong gains in all tasks by providing robust features even on languages unseen during training, such as Catalan.

Language models

Massively adapted language models for each ASR (and MT-target) language were produced and gradually improved (in the case of poliMèdia, Spanish, Catalan and English). See the subsection "Language model" in Section 3.1.4.2 below for a description of the language modelling techniques that were used in transLectures for both ASR and MT.

3.1.4.2 Machine Translation systems

The state-of-the-art in MT at the time of transLectures (2011–2014) was dominated by statistical decision systems combining adequate probabilistic models learnt from training data. Given a text sentence to be translated from a source language into a target language, phrase-based statistical MT systems decided its most likely translation by combining language and translation models. Translation models were usually implemented in terms of phrase tables, learnt by maximum likelihood estimation, though discriminative training was being increasingly used. MT system comparison was often carried out on the basis of automatic metrics such as BLEU or TER. A gentle and accessible introduction to SMT technology at the time can be found in a then-recent monograph by P. Koehn [45].

Translation models

A number of different methods were applied to adapt translation systems to the domain of video lectures. By making use of all available data and state-of-the-art MT decoders, strong baseline systems were first created, and then improved through domain adaptation. One prominent approach that was investigated thoroughly was data selection based on two orthogonal criteria. On the one hand, the Infrequent N-gram Selection (or bilingual sentence selection, BSS) technique developed at UPV a few years earlier [30], and on the other, cross-entropy-based methods, such as LM cross-entropy, TM cross-entropy or a combination of both. It was shown that using less, but more relevant data can improve performance on all language pairs. The cross-entropy criterion was also applied for weighted phrase extraction, which can be seen as a generalization of data selection.

Other techniques that were applied successfully during the project include creating synthetic training data via reverse translation, lexical coverage or relevance features, hierarchical reordering models, word class language models, the use of a language model array and discriminative maximum expected BLEU training. Finally, neural networks were also applied for both translation and language modelling.

Language models

Massively adapted language models for each MT (and ASR) language were produced and gradually improved (in the case of poliMèdia, Spanish, Catalan and English). In particular, the adaptation of language models from time-aligned slides was explored.

In contrast to other language modelling tasks, videos in poliMèdia are usually accompanied with lecture slides which contain extremely focused domain knowledge. Making use of this additional knowledge is a valuable source of information, and several methods of integrating it into the recognition systems were investigated in project transLectures. Language models trained on the slide content were interpolated linearly with the standard in-domain and out-of-domain language models, either in a static (i.e., training a single language model on all slides) or a dynamic (i.e., training a separate language model for each lecture) fashion. Additional relevant documents mined from the web were applied in a similar way. To obtain these documents, the lecture titles, author names, lecture categories and keywords were used either in combination with a web search engine or document retrieval techniques. Manually corrected transcriptions generated during the user evaluation phases in poliMèdia and bootstrapped data taken from previous rounds of repository transcriptions were used as additional data sources for LM adaptation.

UPV also developed a vocabulary selection technique based on unigram model interpolation, which was able to reduce both WER and out-of-vocabulary (OOV) words. A weight rescoring method was tested as well which resulted in improvements in the English systems, where there are big differences between the test videos and the training data. Using cache language models directly adapted to the recognition output resulted in a strong decrease in perplexity, although the effect on word error rate (WER) remained small.

3.1.5 Types of evaluation

In transLectures, three types of evaluation were carried out periodically to monitor the progress in automatic transcription (ASR) and translation (MT) quality along the span of the project.

Here we introduce the three types of evaluation as they will appear in this chapter. First, automatic evaluations were carried out periodically on the basis of the supervised[4] data generated in the project. Second, human evaluations by language experts were made on current transcriptions and translations at major upgrades of project models and tools. Third, human evaluations by users were organized at the case study sites, both with internal and external users, including intelligent interaction approaches to post-editing.

[4]In the context of this master's thesis, "supervising" is the act of correcting an automatic transcription or an automatic translation (this comes from the fact that corrected transcriptions and translations constitute "supervised data" for a supervised learning approach to ASR and MT). Thus, "supervising" will be used as a synonym to post-editing, revising, correcting.

3.1.5.1 Automatic evaluations

Called "scientific evaluations" within the transLectures project, this refers to measuring automatic transcription and translation quality with the kind of objective metrics introduced in Section 2.2.1.

Thus, for MT systems, the BLEU and TER were measured as the systems progressed. BLEU was used as the main metric of reference in project reports, but TER was also tracked as a secondary metric to confirm the improvement of MT quality as the project progressed.

In the case of ASR, the objective metric used was the Word Error Rate (WER).

These automatic metrics were recorded in six-monthly project progress reports.

3.1.5.2 Human evaluations by language experts

Called "quality control" within the transLectures project, this refers to project experts manually supervising a small yet significant amount of transcribed and translated data after major upgrades of project models and tools.

This provided two yearly quality control reports during the project (Year 1 and Year 2), and also new supervised data to be added to the transcription and translation corpora used in the project (with which to improve the project's ASR and MT systems).

For automatic translations, the experts provided a "direct assessment" quality score at the sentence level (see Section 2.2.2.1). Post-editing time was also recorded, so that quality could be measured by the post-editing Real Time Factor (see Section 2.2.2.2). These human evaluations were compared to the automatic measurements of BLEU and TER (measured on the automatic translation against its revised version).

For automatic transcriptions, post-editing time was recorded, so that quality could be measured by the post-editing Real Time Factor. These human evaluations were compared to the automatic measurement of WER (measured on the automatic transcription against its revised version).

3.1.5.3 Human evaluations by users

Two internal user evaluations were organized at each case study site so as to evaluate models, tools and integration progress in a real-life yet controlled setting.

External evaluations were also carried out: user evaluations in a truly real-life setting by opening experimental integrated tools at each case study site to its own prosumers (lecturers and teaching staff) and consumers (casual users and students).

For this master's thesis, we focus on internal user evaluations, which were carried out during Year 2 and Year 3 of project transLectures (and reported in the corresponding yearly reports).

During these human evaluations, users revised the automatic transcriptions and translations produced by the ASR and MT systems. The time required for the revision was recorded. This is reported using the post-editing RTF measure described in Section 2.2.2.2. The expectation is that, as the quality of ASR and MT systems

grows (lower WER, TER...), the cost of post-editing the output (measured as post-editing RTF) will decrease. The post-editing RTF measurements taken will allow us to analyse how well do they correlate with the automatic metrics WER (ASR) and TER (MT).

It is in the context of these human evaluations by users that we will study an *intelligent interaction* approach to post-editing, with hybrid approaches.

3.2 Automatic evaluations

The goal of this task was to measure transLectures progress in terms of automatic transcription and translation quality with objective measures. To this end, development and test sets for scientific evaluations were defined for poliMèdia at the beginning of the project (see Section 3.1.3). Standard evaluation metrics were reported for the test sets along the project: for ASR, the WER; for MT, BLEU and TER. A detailed evolution of scientific results for Spanish ASR and Spanish→English MT can be observed in Figure 3.1.

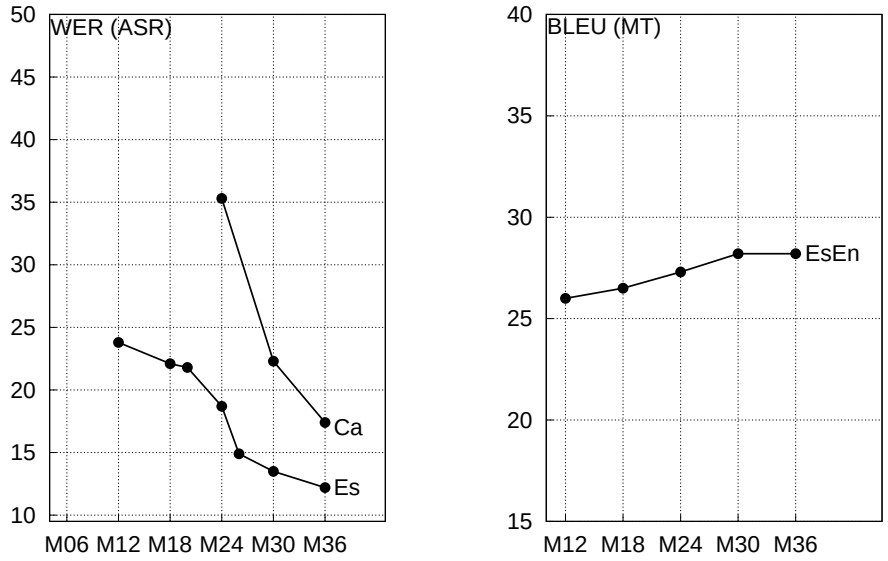

Figure 3.1: poliMèdia: Progress for Spanish and Catalan in ASR (left, in terms of WER) and for Spanish→English in MT (right, in terms of BLEU).

In general, Figure 3.1 reflects the good progress that was achieved for automatic transcription and translation accuracy over the project. For ASR, on the left side of the figure, large improvements were achieved by means of massive adaptation techniques. In absolute terms, 12.2% and 17.4% WER were achieved for Spanish and

Catalan, respectively. A WER below 20% (and even close to 10%) is a clear indication that our Spanish and Catalan systems at the end of transLectures were producing accurate enough transcriptions [57].

For MT, on the right side of the figure, also large improvements were obtained thanks to massive adaptation techniques. At the end of the project, the BLEU score for Spanish→English was 28.2.

Regarding the relationship between the MT evaluations and the ASR evaluations shown in Figure 3.1, it should be noted that these measurements reflect the improvement of ASR and MT quality results due to improvements in the corresponding ASR and MT systems used, independently from each other. What is measured is, for ASR, the WER when automatically transcribing the poliMèdia test set of video lectures; and for MT, the BLEU when automatically translating the correct transcriptions of the poliMèdia test set of video lectures. That is, these measurements are not designed to reflect the impact of improving ASR on the subsequent automatic translation of automatic transcriptions.

3.3 Human evaluations by language experts

As part of the transLectures project's ASR and MT evaluation tasks, a small but significant amount of poliMèdia transcribed and translated data was manually supervised and evaluated by language experts at UPV. These quality controls were carried out after major upgrades of project models and tools, in transLectures Year 1 and Year 2.

For the evaluation, each time a subset of automatic transcriptions (and translations) was selected from the complete set of poliMèdia transcriptions, representing a variety of transcription qualities (low, medium and high, using ASR confidence measures[5] to classify them automatically) and topics.

First, the automatic transcriptions were revised and evaluated. Then, the automatic translations were generated from the revised transcriptions, and they were in turn revised and evaluated.

Project experts were provided with common guidelines to be used in the quality control of transcriptions and translations.

For automatic transcriptions, post-editing time was recorded, so that quality could be measured by the post-editing Real Time Factor. The experts also made notes regarding transcription quality. These human evaluations were compared to the automatic measurement of WER (measured on the automatic transcription against its revised version).

For automatic translations, experts made notes, and they provided a "direct assessment" quality score at the sentence level (on a 1–5 scale; the higher, the better). Post-editing time was also recorded, so that quality could be measured by the post-editing Real Time Factor. These human evaluations were compared to the automatic measurements of BLEU and TER (measured on the automatic translation against its revised version).

[5] See Section 2.2.1.4 for a brief description of confidence measures.

Table 3.3 is a summary of the "direct assessment" translation quality scale that was applied by evaluators in transLectures. The guidelines instructed: "For every sentence, please provide a single translation quality score, using a scale from 1 to 5, following the guidelines below. Deciding which score to attribute to a translation relates to both post-editing effort, as well as the gravity of the translation errors encountered".

Table 3.3: transLectures translation quality assessment scale. Scores are intended to reflect both post-editing effort and gravity of the translation errors encountered.

Quality score	Description
1	The MT output is incomprehensible, with little or no information transferred accurately. It cannot be edited, needs to be translated from scratch.
2	About 50–70% of the MT output needs to be edited. It requires a significant editing effort in order to reach publishable level.
3	About 25–50% of the MT output needs to be edited. It contains different errors and mistranslations that need to be corrected.
4	About 10–25% of the MT output needs to be edited. It is generally clear and intelligible.
5	The MT output is perfectly clear and intelligible. It is not necessarily a perfect translation, but requires little to no editing.

3.3.1 Human evaluations by language experts, transLectures Year 1

28 poliMèdia video lectures in Spanish, totalling 3 hours in length, were selected for the manual revision and evaluation of their automatic transcriptions and Spanish↔English translations.

Regarding automatic transcriptions, the total time needed to edit the whole set of 3 hours was approximately 30 hours, that is, RTF 10. The effort required to edit the automatic transcriptions was thus similar, at that point, to the effort it would have taken to manually transcribe the lectures (taking as reference the manual transcription effort figures recorded in transLectures).

Regarding automatic translations, the average human quality score across all evaluated translations was 2.9 (from 1 to 5). The experts stated that, at this stage, automatic translation quality left ample room for improvement.

As to the MT post-editing RTF, the total time needed to edit the whole set of 3 hours was approximately 120 hours, that is, RTF 40. This means that correcting the automatic translations took a similar time as it would have taken to translate

Table 3.4: Summary of MT manual evaluation metrics vs. automatic quality metrics on a representative set of Spanish into English translations.

Lecture ID	Score	BLEU	TER
23	1.5	36.7	45.7
4	1.7	37.5	39.7
19	1.9	42.4	37.7
27	2.0	41.7	41.4
25	2.2	46.6	39.8
28	2.3	35.9	46.3
10	2.4	36.9	42.8
24	2.4	39.4	43.0
22	3.1	48.0	33.8
5	3.1	30.2	42.4
26	3.2	41.5	41.3
18	3.3	33.6	43.4
3	3.3	50.5	30.1
15	3.9	54.3	31.6
12	4.0	48.0	33.2
1	4.5	56.4	27.4
Average	2.9	42.2	39.8

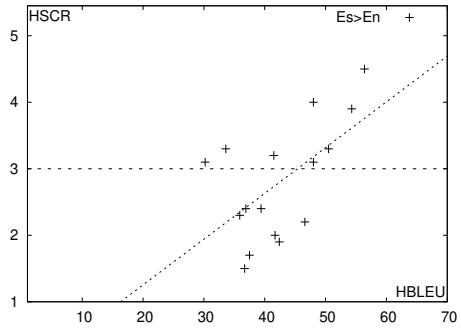

Figure 3.2: Human quality score vs. BLEU for a representative set of Spanish into English translations.

them manually (taking as reference the manual translation effort rates recorded in transLectures).

A comparison of MT manual evaluations and automatic measurements is provided in Table 3.4, sorted by human score. This is represented graphically in Figure 3.2, which shows HBLEU scores versus the human quality scores (HSCR). We can appreciate that there is a certain degree of correlation between both measurements.

In addition to the direct assessment scores, post-editing RTF measures and WER, BLEU and TER automatic measures resulting from these user evaluations, the evaluators made notes on the linguistic errors they observed that could be used to inform the improvement of the systems. As an example, these are the notes from the report on Spanish↔English automatic translations:

> Some frequent mistakes in the automatic translation process might be easily avoided or solved automatically through post-processing:
>
> - Both American English and British English spellings are used in the translations (e.g. "analyze" (AE) in Lecture ID 27, "characterises" (BE) in Lecture ID 23).

- The indefinite article "a" is not changed to "an" before words beginning with a vowel (e.g. "a illegitimate" in Lecture ID 27).
- Contractions are used at random (e.g. "we're" and "we are" in Lecture ID 27).
- The adjective plus noun order is frequently reversed (e.g. "a programme infected", "has code illegitimate" in Lecture ID 27).
- The first person pronoun "I" is written in lower-case (e.g. in Lecture ID 18).
- "That is to say" is frequently used to translate "es decir" (e.g. in Lecture ID 27), when it should appear more frequently in its simpler form, "that is".

3.3.2 Human evaluations by language experts, transLectures Year 2

20 poliMèdia video lectures in Spanish, totalling 2 hours in length, were selected for the manual revision and evaluation of their automatic transcriptions and Spanish↔English translations.

The experimental setup, in terms of transcription and translation guidelines and user interface (the TLP Player, see Section 3.4.1.1 for more details), was designed to better reproduce the conditions under which transcriptions and translations are supervised by regular users of the transLectures system. Under these conditions, user interaction and timing statistics were collected to analyse the behaviour of the experts and their performance compared to transcribing from scratch or translating from scratch (from the correct transcription).

Regarding automatic transcriptions, the average WER over the 20 video lectures was 23.2. Year 2 results showed a very significant reduction of effort, with an average RTF of 3.8. This means that the effort required to edit the automatic transcriptions was less than 50% than it would have taken to manually transcribe the lectures (taking as reference the manual transcription effort figures recorded in transLectures). Figure 3.3 confronts WER scores to RTF for automatic transcriptions, showing their correlation[6]. As can be observed, all transcriptions with WER scores below 20 were supervised with RTF lower than 3, that is, more than 3 times faster than transcribing from scratch; for transcriptions with WER higher than 20, RTF figures were above 5 but below 7 (still faster than transcribing from scratch).

Regarding automatic translations, the average quality score across all evaluated translations was 3.6 (from 1 to 5), showing a remarkable improvement in automatic translation quality over the previous year. This improvement was mimicked in BLEU and TER automatic measurements, which improved in average from 42.4 to 46.6 and from 39.8 to 36.4, respectively.

As to the MT post-editing RTF, the average was 14.8 (meaning that the process of editing one hour of transLectures translated subtitles would take 14.8 hours to

[6]It should be noted that the adjustment to the points in Figure 3.3 is linear; the adjustment curve only seems to be quadratic because the X axis is displayed in logarithmic scale.

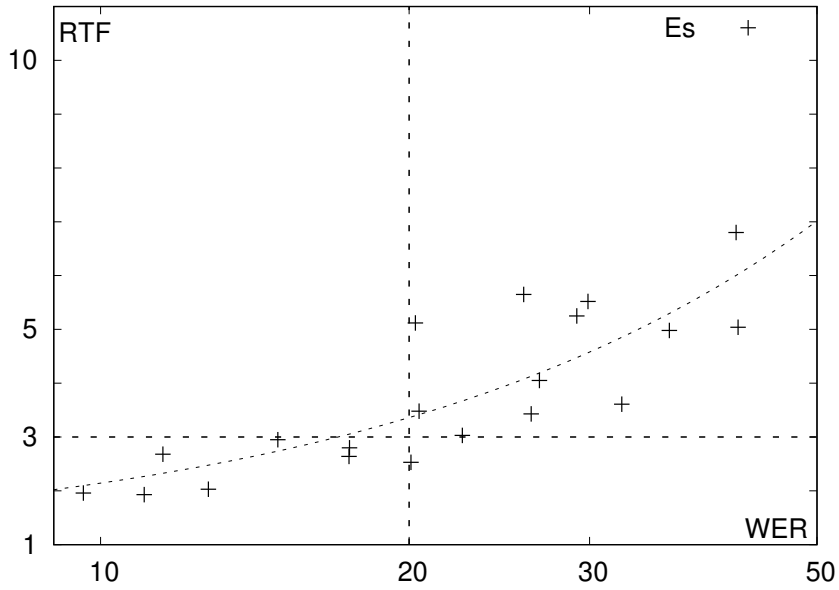

Figure 3.3: RTF versus WER (log scale) for Spanish transcription supervisions.

complete). This result is a significant improvement on previous figures of 34–40 RTF for the process of translating a video lecture from scratch (from its transcription) and means that, for this set of lectures, experts were able to produce correct translations more than 2 times faster than if translating from scratch.

Table 3.5 gives, for each lecture, lecture ID, average manual score, HBLEU, HTER and RTF (sorted by HBLEU in descending order). This is represented graphically in the following two figures.

Figure 3.4 shows HBLEU scores versus RTF. The translations from Spanish into English (poliMèdia) obtain high HBLEU scores, and present RTF figures below 20 in all cases, even below 10 in some cases.

Figure 3.5 shows HBLEU scores versus the manual quality scores (HSCR). The translations from Spanish into English (poliMèdia) obtain higher HSCR scores as their HBLEU scores increase.

Lastly, as to the notes from the evaluators on linguistic errors in the automatic translations, it was noticed that some of the recurrent errors had diminished in frequency with the improvement of the MT systems, while a few still persisted (mainly, the mix between British English and American English spellings, and the incoherent use of contractions).

Table 3.5: Summary of manual and automatic quality metrics on the selected set of Spanish into English translations (sorted by HBLEU, descending).

Lecture ID	Score	HBLEU	HTER	RTF
8600	3.8	68.4	20.5	14.7
5684	3.7	56.4	26.4	9.2
7073	3.8	55.9	32.4	9.9
7351	4.2	53.6	30.5	7.7
9156	3.2	51.7	25.9	11.1
1769	3.9	51.6	31.6	20.0
7481	3.5	50.5	35.2	15.4
1829	3.6	49.4	31.7	7.4
6489	3.9	48.8	38.1	5.0
7218	3.6	47.9	40.2	15.3
594	3.2	47.6	35.6	11.6
2300	3.7	46.2	34.5	18.4
6720	3.5	46.2	38.2	15.0
6297	3.6	45.3	38.5	18.9
6015	3.5	43.5	42.2	15.2
5013	3.4	43.0	39.3	9.5
8501	3.5	40.2	42.0	11.2
168	3.5	39.1	40.0	18.5
2430	3.3	33.6	36.3	16.0
7002	2.8	28.9	46.6	18.8
Average	3.6	46.6	36.4	14.8

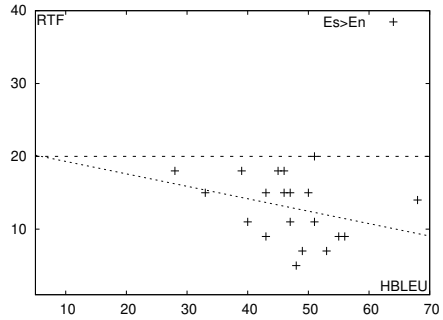

Figure 3.4: RTF versus BLEU for Spanish into English translations.

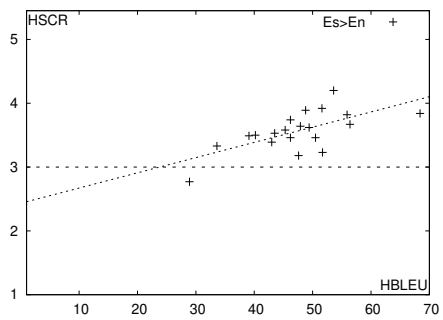

Figure 3.5: Quality score versus BLEU for Spanish into English translations.

3.3.3 Conclusions on human evaluations by language experts in transLectures

According to project experts, noticeable improvements were observed in the poliMèdia pilot as the project progressed.

Regarding ASR in poliMèdia, transcription quality was very diverse from lecture to lecture. However, sound quality was generally not a determining factor in this respect, as all poliMèdia video lectures are recorded in a dedicated recording studio, and so audio quality is similarly high for most lectures. This led to the hypothesis that the initial systems were better prepared for some topics than others, and thus these systems had to be improved in terms of language modelling. This was accomplished with extraordinary success in Y2 and Y3 (see Y3 internal evaluations in Section 3.4.2 below).

As for MT in poliMèdia, the quality of the automatic translations was again found to vary greatly from lecture to lecture. A key factor impacting the quality of the automatic translations was the competence of the speaker when presenting their subject. This has much to do with the extent to which the lecturer prepared their lecture before recording it at the studio: the more "scripted" the speech, the better the automatic translation. The impressions of the experts in this respect were that lectures in which the speaker used complete sentences, with complete grammar structures and more standard language resulted in higher-scoring translations. Lectures which had more incomplete sentences, repetitions or self-corrections, and more colloquial turns of phrase tended to result in lower-scoring translations.

Regarding the variety of MT measures gathered during these evaluations by experts, several expected correlations could be observed when looking at the results across the list of revised automatic translations and across the transLectures languages and language pairs that were part of these human evaluations (EN, ES, FR, DE, SL):

1. BLEU vs TER: A correlation could be observed between these two automatic metrics.

2. BLEU/TER vs DA score: A certain degree of correlation could be observed between the two automatic metrics and the human evaluation scores (direct assessment "correctness" score).

3. Automatic and human scores vs RTF: A certain degree of correlation could be observed between, on the one hand, the human and automatic measures, and on the other hand, the post-editing RTF. That is, the worst the measured transcription quality, the more costly it was to revise it.

All in all, in the final round of quality control (Y2), results were considered really useful in terms of productivity gains by professional editors.

3.4 Human evaluations by users

Internal and external user evaluations were also carried out in project transLectures. For internal evaluations, on the UPV's poliMèdia repository, UPV lecturers revised automatic transcriptions and translations using the TLP player (the multimedia player for subtitle post-editing developed at the UPV for transLectures); in the case of transcriptions and some translation combinations, significant user effort reductions were measured.

For external evaluations, the automatic subtitles in poliMèdia were made open for revision by any poliMèdia viewer. This functionality is still active to this day; revisions by external viewers are submitted for the lecturer's approval before being made public.

In this section, we focus on the internal user evaluations that were carried out during Year 2 and Year 3 of project transLectures (and reported in the corresponding yearly reports).

In Year 2 (Section 3.4.1), internal user evaluations were carried out for ASR, including the implementation and study of intelligent interaction approaches to post-editing, as well as hybrid approaches (referred to in this section as *computer-aided interaction*). These results were published in 2 scientific articles, one in the journal *Open Learning* (2014) and one in the conference In-Red 2014[7], on which this section is based.

In Year 3 (Section 3.4.2), the scope was widened to include internal user evaluations for both ASR and MT.

3.4.1 Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures (transLectures Y2)

In Year 2 of the transLectures project, automatic subtitles in Spanish, English and Catalan had already been made available for all videos in the poliMèdia repository and were continually being updated as technologies were improved.

At this point, the first internal user evaluations were carried out in collaboration with UPV lecturers who, having recorded material for the poliMèdia repository as part of an earlier Docència en Xarxa call, during the project trialled the transLectures transcription editing interface.

For this first internal user evaluation, it was decided to focus on ASR, which is the first step in the workflow of producing multilingual subtitles for video lectures using MT.

Having already introduced the transLectures project and poliMèdia (in Sections 3.1.1 and 3.1.2), here we discuss the three-phase user evaluation stage and the progress made based on user feedback in the area of interface design, functionality and the project's computer-aided interaction component.

[7] See Section 3.5 below for the details of these publications.

3.4.1.1 transLectures Year 2 user evaluation setup

Docència en Xarxa (Networked Teaching) [1] is an ongoing incentive-based programme to encourage university lecturers at the UPV to develop digital learning resources based on ICTs. Having already filmed material for the poliMèdia repository as part of an earlier call, in 2012/13 lecturers were invited to evaluate the computer-assisted transcription (CAT) tools being developed in transLectures and tested on the poliMèdia video repository. Lecturers signing up for this programme committed to supervising[8] the automatic transcriptions of 5 of their poliMèdia videos. These videos were transcribed using TLK, the transLectures-UPV toolkit for automatic speech recognition, which at the time of these evaluations correctly recognized 4 out of every 5 words spoken by the lecturer. For evaluation purposes they were allocated across 3 progressive evaluation stages, described below:

- *First phase*: Lecturers manually supervise the automatic transcription of the first short video presentation from start to end. In this phase, the lecturer plays the short video presentation and the automatic transcription appears, split into synchronized segments of up to 20 words. When they spot a transcription error, the lecturers click on the incorrect segment to enter their correction. The video automatically pauses. In this phase, 1 video was supervised.

- *Second phase*: In this phase, a word-level computer-aided interaction model is introduced. The CAT tool preselects a subset of words within the automatic transcription based on confidence measures (CMs), presenting the lecturer with only those words it considers least likely to be correct. The lecturer supervises these words, playing them in the context of one word before and one word after. 2 videos were transcribed in this way.

- *Third phase*: This phase is in fact split into 2 sub-phases or rounds of evaluation. The first round essentially corresponds to phase two, above, where the lecturer supervises only the sections of the transcript identified as least likely to be correct by the CAT tool. The entire video lecture is then automatically re-transcribed on the basis of the lecturer's supervision actions. In a second round, the resulting transcriptions are supervised in full by the lecturer from start to finish, as in phase one. The idea is that these new transcriptions are of a significantly higher quality than the original transcriptions [61], so much so that, even counting the time spent in round one, lecturers spend less time supervising the automatic transcriptions. In this phase, the remaining 2 short video presentations were transcribed.

Feedback was collected after each phase in the form of a brief 10-question satisfaction survey. In addition, the TLP player (described in more detail below) logged precise user interaction statistics, such as the duration for which the editor window is

[8] As explained in Section 3.1.5, "supervision" in this context should be understood as the act of reviewing the automatic transcription, and confirming or correcting the text as necessary; confirming when the suggested text is correct, and correcting when it is incorrect.

open, the number of segments (individual subtitles) edited out of the total, the display layout selected; as well as statistics at the segment level including the number of mouse clicks and key presses, editing time, and number of times a segment is played. All of this information was used to inform the design of each subsequent evaluation phase and, ultimately, of the TLP player interface itself.

As one of transLectures' main end user or "prosumer" groups, feedback from university lecturers was fundamental to the outcome of the project. The end goal was a user-friendly platform for post-editing automatic transcriptions that was cost- and time-effective [54].

The CAT tool being tested in this evaluation stage, the TLP player, consisted of an innovative multimedia web player with editing capabilities, complete with alternative display layout options and full keyboard support. It was developed as part of the transLectures project [80]. A screenshot of the TLP player (side-by-side layout) is shown in Figure 3.6.

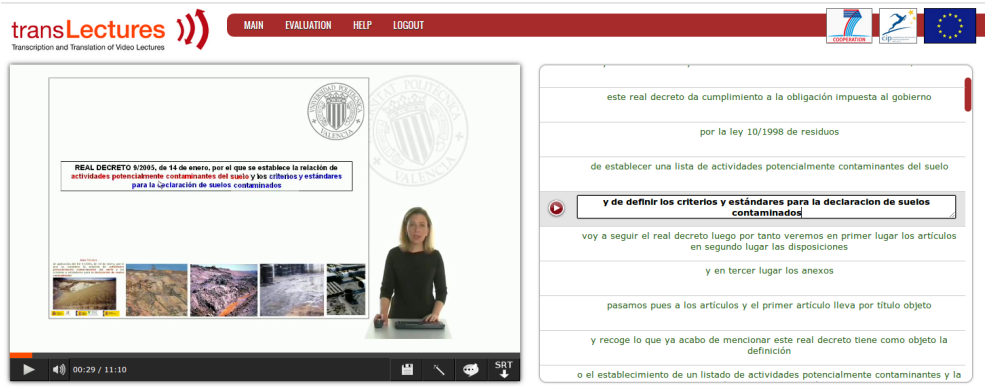

Figure 3.6: TLP player with the side-by-side layout while the lecturer edits one of the segments.

3.4.1.2 Description of user evaluations

First phase: Complete supervision

In the first phase, 20 UPV lecturers supervised the automatic transcription of the first short video presentation in its entirety. The process is straightforward: the lecturers, assigned a username and password, log into the TLP player to access a private area with their poliMèdia videos and select the video they wish to supervise. The TLP player, shown in Figure 3.6, is automatically loaded and lecturers can start supervising the transcription straight away.

The TLP player plays the short video presentation and the corresponding transcription in synchrony, allowing the user to read the transcription while watching the video. When the lecturer spots a transcription error, they press the "Enter" key or

click directly on the incorrect segment to pause the video. With the video paused, the lecturer can easily enter their changes in the text box that appears. Lecturers save their work periodically, upon which the transcription is updated and user interaction statistics dumped into a log file.

A preliminary round of phase one was carried out with just two lecturers who volunteered to trial a draft version of the TLP player. The backgrounds of these two lecturers differed (one from computer science and the other from architecture), which was important in order to avoid opinion biases on issues of interface usability. For instance, the two lecturers presented very different user interaction patterns: the computer science lecturer interacted with the CAT tool primarily using the keyboard, while the architecture lecturer showed a clear preference for the mouse.

Based on the feedback from these first two users, the TLP player was significantly improved in advance of the launch of phase one proper. Firstly, we shortened the average length of the transcription segments down to 15 words, in line with recommendations from the subtitling industry. This shorter length allows the user to more easily remember what was said in the video and therefore more efficiently correct the words incorrectly recognized by the CAT tool. Secondly, a "search and replace" function was incorporated into the TLP player, at the suggestion of our computer science lecturer. Finally, both lecturers suggested that correct transcription segments be automatically confirmed once the corresponding video segment has been played, rather than requiring manual confirmation. The remaining 18 lecturers were then asked to supervise the first of their videos using this updated version of the TLP player.

The most important finding after this first phase was the vast reduction in the time required to produce a transcription for a video lecture. The time spent by lecturers supervising the automatic transcriptions was significantly lower than if transcribing manually from scratch; just over half the time (54%). Indeed, the lecturers' performance became comparable to that of professional transcriptionists [33], rather than that expected from non-expert transcriptionists [57].

In terms of more qualitative feedback, lecturers valued the simplicity and efficiency of the interface. In the satisfaction survey they collectively scored the TLP player at 9.1 out of 10 for usability, showing a high acceptance of our CAT prototype as is. That said, one of the lecturers, who had previous professional transcription experience, was unhappy with the interface layout in that it was different to what he was used to working with.

All in all, results were largely positive and, as was the goal, lecturers were able to become familiar with the TLP player in advance of the next two phases.

Second phase: Computer-aided interaction

In the second phase, a new interaction protocol called *computer-aided interaction* [66] was introduced to find out whether this could further improve supervision times, that is, whether it was possible to make this process even more efficient for the lecturers.

This new interaction mode is based on confidence measures (CMs) [62], specifically CMs at the word level[9]. Word-level CMs provide an indicator as to the probable correctness of each word appearing in the automatic transcription. Words with low confidence values are likely to have been incorrectly recognized at the point of ASR and will, therefore, need to be corrected in order to obtain an accurate transcription. With a perfect CM system, the lecturer would only ever supervise (correct) incorrectly-recognized words. In practice they must also supervise (confirm) some correctly-recognized words incorrectly identified as errors. These false positives are unavoidable, since our systems are based on statistical models. They are, however, preferable to false negatives. The idea is that by focusing supervision actions on incorrectly-transcribed words, user interaction can be optimized to get the best possible transcription in exchange for the least amount of effort.

So in this evaluation phase, lecturers were asked to supervise a subset of words preselected by the CAT tool as low confidence, presented in order of probable incorrectness. This subset typically constituted between 10–20% of all words transcribed using the ASR system, though lecturers could modify this range at will to as low as 5% and as high as 40%, depending on the perceived accuracy of the transcription. Each word was played in the context of one word before and one word after, in order to facilitate its comprehension and resulting correction. Typically, given the starting word error rate (WER) of our automatic transcriptions (10-20%) and an average supervision rate of 15%, our CMs detected around 40% of all real transcription errors.

Figure 3.7 shows a screenshot of the transcription interface in this phase. In this example, low-confidence words are shown in red and corrected low-confidence words in green. The text box that opens for each low-confidence word can be expanded by the users in either direction in order to modify the surrounding text as required.

For this phase, the computer-aided interaction mode was activated in the TLP player by default, though lecturers could switch back to the full supervision mode tested in phase one. Analysis of the interaction statistics reveals that only 12 of the 23 lecturers stayed in the computer-aided interaction mode for the supervision of one of their poliMèdia videos in full. In the other cases, lecturers switched back to complete the supervision mode. The main reason cited for this was that only by doing so could they be sure to obtain a perfect transcription, something which they professed to value over any time-savings afforded by the computer-aided interaction mode.

In terms of supervision times, the time spent correcting automatic transcriptions in computer-aided interaction mode was reduced to 40% of the time needed for the complete supervision in phase one. However, the resulting transcriptions were not error free, unlike in phase one. That said, the error rate was as low as one in every 10 automatically-recognized words, which is not so far removed from the transcription quality delivered by commercial transcriptions services for academic video lectures [33].

Feedback from phase two was not as positive as from phase one. Lecturers showed a clear preference for obtaining perfect transcriptions, irrespective of the relative

[9]Confidence measures have already been introduced in Section 2.2.1.4; here we detail somewhat more how they apply in this case

Figure 3.7: Screenshot of the transcription interface in computer-aided interaction mode. Low-confidence words appear in red and supervised low-confidence words in green. The word being edited in this example is opened for supervision, and the text box can be expanded to the left or right by clicking on « or », respectively. Clicking the green check button to the right of the text box confirms the word as correct.

time costs, and insisted that full access be granted to both the video/audio and the transcription. The satisfaction surveys clearly reflected this dislike of the computer-aided interaction mode, preferring a protocol that gives them full control over the end quality of the transcriptions. However, they did seem to embrace the CMs, suggesting that low confidence words be indicated in red font in complete supervision mode also. Overall the system scored 7.2 out of 10 at this stage.

Third phase: Two-round supervision

As indicated, the third phase is divided into two sub-phases or rounds, and is essentially a combination of the previous two phases. First, lecturers supervise a subset of the lowest confidence words, as in phase two, for the remaining 2 poliMèdia videos. The videos are then re-transcribed on the basis of this partial supervision and, in the second round, lecturers supervise the entire re-transcription from start to end,

as in phase one. The idea is that the quality of these new transcriptions is higher than that of the original [61], sufficiently so as to allow lower overall supervision times.

In more detail, in the first round lecturers supervised isolated segments of 4 words in which the last word was the low-confidence word. These segments were presented to the lecturer for supervision in increasing order of confidence (of the last word). The segments kept on being presented to them until one of three conditions was met:

1. The total supervision time reached double the duration of the video itself; or
2. No corrections were entered for 5 consecutive segments; or
3. 20% of all words were supervised.

On average, supervision times during this first round were equal to the duration of the video being supervised, and approximately 15% of incorrectly-recognized words were corrected. These supervision actions were fed into the ASR system used to generate the re-transcriptions, which is the same system used to generate the original transcriptions, but adapted to the lecture- and lecturer-specific variables provided during round one.

Lecturers then embarked on round two, in which they supervised the retranscriptions from start to end, as in phase one. The time needed to perform this complete supervision was much lower than in phase one because, as expected [61], the quality of the re-transcription was significantly higher and, consequently, fewer corrections needed to be entered manually: approximately 35% of the incorrectly-recognized words were corrected prior to the start of round two, a combination of the lecturers' efforts in round one and the automatic re-transcription on the basis of their corrections. The end result was a perfect transcription, as in phase one.

In this computer-aided interaction mode, the aim was to blend the best outcomes of both previous phases: the perfect end transcriptions of phase one and the shorter supervision times of phase two. Ultimately, this was achieved, though only by a small margin. The complete supervision of the re-transcriptions in round two required 80% of the time needed to do the same task in phase one. However, when supervision times from round one were added, the time-saving with respect to phase one was slight (5%).

The main drawback of this model is the two-step process, since lecturers have to put time aside on two separate occasions to supervise the same video. On the whole, a preference (if not requirement) was expressed for the supervision to be carried out in a single session and the corresponding impact on user satisfaction was evident in the average satisfaction survey score for this phase: 7.8 out of 10.

3.4.1.3 Results summary

Here we summarize the average results for each phase in terms of Word Error Rate (WER), supervision Real Time Factor (RTF), and satisfaction survey score (on a scale from 0 to 10, from a satisfaction survey considering several aspects of usability and user preferences).

Table 3.6 shows, column by column, from left to right, the supervision protocol used, the average initial WER obtained by the automatic transcriptions, the average final WER after user supervision, the average time needed by the users for supervision in terms of RTF, and the average subjective evaluation of the ASR system through satisfaction surveys filled in after finishing the user evaluation process; row by row, we can see the results for each supervision protocol (complete supervision, computer-aided interaction and two-round supervision).

Table 3.6: Comparison of supervision protocols for automatic transcriptions.

Supervision protocol	Initial WER	Final WER	RTF	Survey score
1 - Complete supervision	16.9	0.0	5.6	9.1
2 - Computer-aided interaction	14.5	8.0	2.2	7.2
3 - Two-round supervision	28.4	0.0	5.3	7.8

The supervision RTF can be compared against the baseline of the time needed to create a manual transcription from scratch (without an initial automatic transcription), which is estimated to be around RTF 10 for non-experts [58, 80].

Based on the results of this three-phase user evaluation, we can say that the more simple complete supervision protocol used in phase one was preferred by the users, and it allowed them to reduce total transcription time to almost a half compared to manual transcription from scratch. In phase two, computer-aided interaction based on confidence measures required less effort from the users, but the result was a transcription still containing some errors, which was not acceptable for the users given that the goal of these videos is education. Finally, the two-round supervision in phase three offered slightly higher time savings with respect to phase one (even starting from initial automatic transcription with higher WERs), but due to the two-round process being more complex, the users did not prefer this protocol over phase one.

Correlation between automatic and human evaluations

We would expect, for each video lecture, for supervision times (RTF) to depend on the automatic transcription quality (initial WER). This would confirm that improving automatic transcription quality will imply reducing supervision times.

To verify this hypothesis, the data collected in phase one was fit to several linear regression models in order to explain RTF in terms of WER. Table 3.7 shows the two models that resulted from this study.

Model 1 revealed that WER ($beta = 0.285$, Sig $= 4.73{\times}10^{-9}$) was statistically significant as a predictor of RTF and accounted to a large extent for the variance observed in the data ($R^2 = 0.842$).

A graphical representation of our data in terms of WER versus RTF, and our prior knowledge of user behaviour (that users essentially ignore automatic transcriptions

Table 3.7: Linear regression models to explain RTF in terms of WER.

Predictor	Beta	p-Value
Model 1 ($\triangle R^2 = 0.842$, $R^2 = 0.842$, F=101.3, Sig=4.73×10^{-9})		
WER	0.285	4.73×10^{-9}
Model 2 ($\triangle R^2 = 0.075$, $R^2 = 0.917$, F=210.6, Sig=9.82×10^{-12})		
$log_e(WER)$	2.025	9.82×10^{-12}

above a certain WER threshold, preferring to transcribe from scratch) suggested that a logarithmic model might better fit our data. Consequently, the logarithmic Model 2 was proposed, resulting in a more statistically significant *beta* ($beta = 2.025$, Sig = 9.82×10^{-12}) and a minor increase in the variance explained by the model ($\Delta R^2 = 0.075$).

The correlation between WER and RTF measures in each phase can be observed in Figure 3.8, which shows RTF as a function of WER.

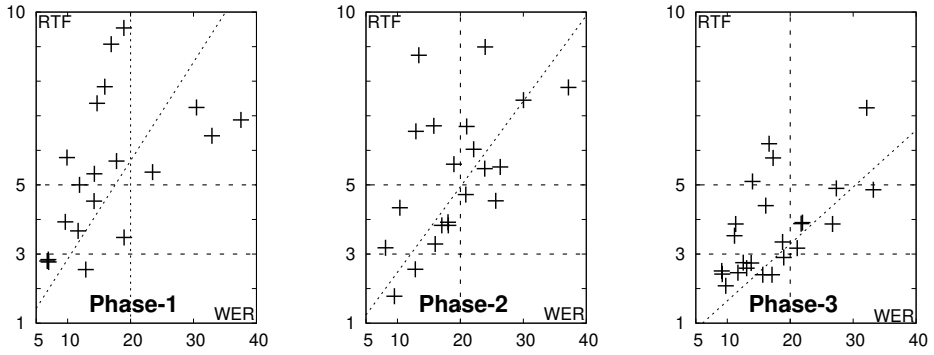

Figure 3.8: RTF as a function of WER for each of the three phases of Y2 user evaluations at poliMèdia (each point represents a supervised video).

3.4.1.4 transLectures Year 2 user evaluation conclusions

In this section we have outlined how the transcription and translation technologies developed as part of the transLectures project are adding value to the poliMèdia repository at the Universitat Politècnica de València. We paid particular attention to the three-phase internal user evaluation stage, in which transLectures tools were deployed in a real-life setting and their usefulness and usability assessed based on feedback from real-life users.

We have reported how the basic user interface trialled in phase one allowed lec-

turers full control over the quality of the transcriptions of their poliMèdia lectures. This mode of interaction scored highly in the satisfaction survey, offering considerable time-savings relative to transcribing from scratch (reducing the time required to 54%). We then presented the computer-aided interaction mode tested in phase two which, despite offering even greater time-savings (reducing the time required to 22% compared with transcribing from scratch), failed to capture the lecturers' interest because it did not lead to perfect end transcriptions. Then, in the third phase, the best of the previous phases was brought together in a two-round supervision process. Here, though, time-savings relative to phase one were small and the process less appealing to lecturers, who preferred the simplicity of a single-round supervision process.

However, we should point out that overall transcription accuracy was improved by 35% following an initial supervision of the least confident words, thereby confirming the validity of our use of Confidence Measures (CMs) and computer-aided interaction as a means of improving transcription quality. Ultimately, however, UPV lecturers were not overly interested in this trade off between perfect output and time costs.

These results suggest that it would be worth it to combine the full control allowed in phase one and the use of CMs as in phase two in a way that lecturers find useful and usable. For example, an interface where it was possible to switch seamlessly from complete supervision mode to computer-aided interaction mode, depending on the perceived quality of the automatic transcription, might be well received by lecturers.

Having verified through user evaluations the usefulness of ASR for cost-effective transcription of video lectures, for the next round of user evaluations in transLectures Year 3 it was planned to also evaluate transLectures automatic translation solutions. For these trials, the user interface would have to be redesigned to allow side-by-side, synchronized visualization of the video, the transcription and the corresponding translation. Additionally, in the light of the feedback received during this round, it was decided to reconsider the supervision protocols, keeping in mind the profile of university lecturer "prosumers" in so far as their needs and skill sets are not necessarily aligned with those of professional translators, the typical target users of translation technologies. In the next Section 3.4.2 we will see how this was implemented.

3.4.2 Human evaluations of automatic transcriptions and translations by users (transLectures Y3)

From the lessons learnt in Year 2 of project transLectures, a new internal user evaluation campaign was carried out for the 2013-2014 course under the UPV action plan Docència en Xarxa. As we will see, the main new element with regard to the topic of this master's thesis was that, this time, automatic translations were also evaluated.

3.4.2.1 transLectures Year 3 user evaluation setup

For this second round, the focus of the internal user evaluations was extended to assess also automatic translations (from Spanish into English), using again the TLP player, which was redesigned to support displaying and editing translated subtitles. More languages were also covered this time, not only including Spanish as in Section3.4.1,

but also English and Catalan (the other main languages available in the poliMèdia repository).

Also, to increase the scale of this new user evaluation round, this time not only the lecturers who had recorded video lectures participated in supervising their automatic transcriptions and translations, but also other lecturers (when such permission was provided by the video lecture's authors).

Participating lecturers committed to supervising the automatic transcriptions of 5 poliMèdia videos or the automatic translations of 3 poliMèdia videos. These videos had been transcribed using TLK, the transLectures-UPV toolkit for ASR [2], and the transcriptions had been translated into the other languages (Spanish, English and Catalan) using the transLectures project's PBMT systems.

Taking into account the results of Y2 user evaluations, the post-editing protocol was adopted for both transcription and translation supervisions. The automatic transcriptions and translations to be supervised were those generated by the UPV's transLectures systems at M24 (Spanish: ∼15 WER, English: ∼25 WER, Catalan: ∼35 WER; Spanish to English: ∼30 BLEU). The English transcription system had been trained on non-domain-specific data (mainly VideoLectures.NET data), so the acoustic conditions are rather different than those of poliMèdia recordings.

The TLP Player for transcription supervision was the same as in Y2 (see Figure 3.9). However, for translation supervision, a new interface was designed in order to show the original text segment and the corresponding translated text segment in synchrony with the video (see Figure 3.10). Feedback from lecturers was also collected using a satisfaction survey as in Y2.

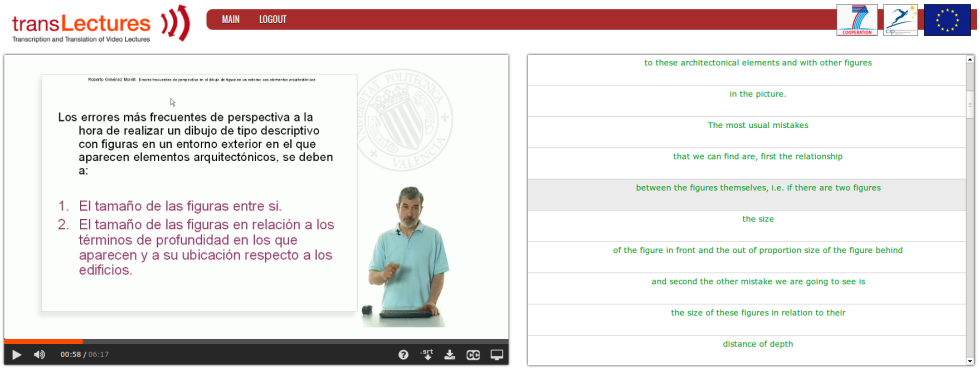

Figure 3.9: TLP Player for automatic transcription supervision (transLectures Y3).

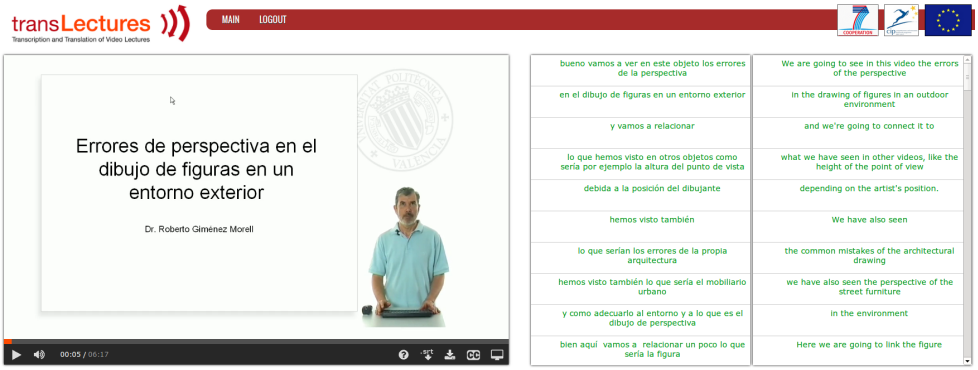

Figure 3.10: TLP Player for automatic translation supervision (transLectures Y3).

3.4.2.2 Supervision of automatic transcriptions

Table 3.8 summarizes the results in terms of average WER, RTF and the satisfaction survey (SS) score[10] on the supervision of automatic transcriptions.

Table 3.8: WER, RTF and satisfaction survey (SS) scores for automatic transcription supervisions in poliMèdia.

Language	Participants	Supervisions	Hours	WER	RTF	SS score
Spanish	39	135	18.3	12.0	2.7	8.5
English	12	57	7.9	36.0	6.2	-
Catalan	5	19	1.5	40.2	5.6	8.6

The WER figures achieved at M24 by the UPV's transLectures automatic transcription systems provided automatic transcriptions the supervision of which was already more efficient than transcribing from scratch, with significant effort reductions about 70%, 40% and 35% for Spanish, English and Catalan, respectively (the effort reduction being lower for languages with a higher WER at the time of these evaluations).

Correlation between automatic and human evaluations

As expected, supervision times seem to depend on the automatic transcription quality (as well as on user expertise). In all three cases, Spanish, English and Catalan, a statistically significant linear dependency was found between WER and RTF.

[10] Satisfaction surveys were not reported for the supervision of English automatic transcriptions, since it was performed by UPV transLectures team members.

Simple linear regression models were fitted to all the data collected from the automatic transcription evaluations in Y3. The results for Spanish and English are shown in Table 3.9.

Table 3.9: Linear regression models to explain the WER-RTF dependency for automatic transcription evaluations in Y3.

Model	Beta	p-Value
Spanish ($R^2 = 0.811$, F=134.0, Sig=$< 2.2 \times 10^{-16}$)		
$RTF = Beta \times WER$	0.184	$< 2.2 \times 10^{-16}$
English ($R^2 = 0.917$, F=583.5, Sig=$< 2.2 \times 10^{-16}$)		
$RTF = Beta \times WER$	0.168	$< 2.2 \times 10^{-16}$

As we can see, the adjustment of the model to the data is statistically significant (p-Value $< 2.2 \times 10^{-16}$), and so it can be used to infer RTF in terms of WER in poliMèdia.

The same clearly statistically significant linear dependency between WER and RTF (p-Value $< 2.2 \times 10^{-16}$) was found for Catalan.

As a graphical representation of this, Figure 3.11 represents Spanish-language video transcription supervisions in transLectures Y2 and Y3 together with their corresponding linear regression fitting plotted as a dotted line in the same colour. The X axis and Y axis show WER and RTF for each instance of supervision, respectively. As can be observed, there is a clear linear dependency between RTF and WER, with most of the Y3 supervisions concentrated below 10 WER (high transcription quality) and 3 RTF (low supervision effort). This is consistent with the fact that the WER of the ASR system on the poliMèdia test set improved from Y2 to Y3 by approximately 5 WER points.

The linear correlation for English is in turn reflected in Figure 3.12, in which English-language supervisions are represented in terms of WER versus RTF.

3.4.2.3 Supervision of automatic translations

Ten lecturers took part in the supervision of Spanish into English translations accounting for 13 video translations fully reviewed (about 2.1 hours of video). As mentioned above, the TLP Player was modified to show the video with the source and target segments in synchrony, as shown in Figure 3.10.

A summary of the main figures for automatic translation supervisions (TER, RTF and satisfaction survey (SS) score) is shown in Table 3.10.

The average RTF was 12.2, while the average BLEU was 49.8 and TER was 41.9. If we compare this RTF figure to the RTF expected for manual translations (about 30–40 RTF), we observe a significant relative decrease in user effort. Thus, the Spanish-English automatic translation supervision task proved to save time compared to translating from scratch.

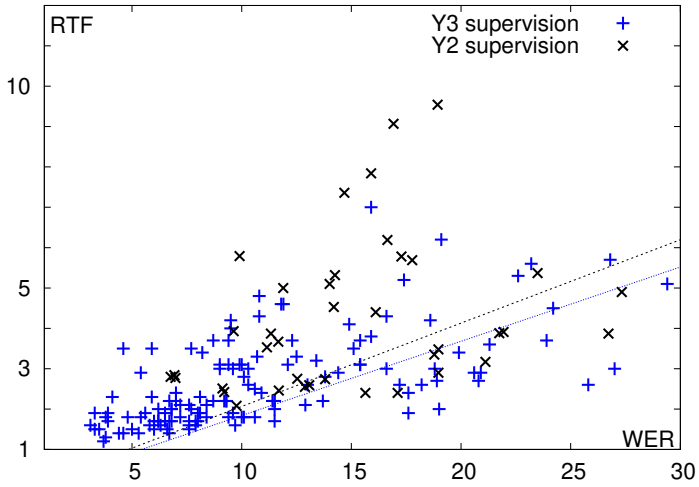

Figure 3.11: RTF versus WER for Spanish-language transcriptions supervised (transLectures Year 2 and Year 3).

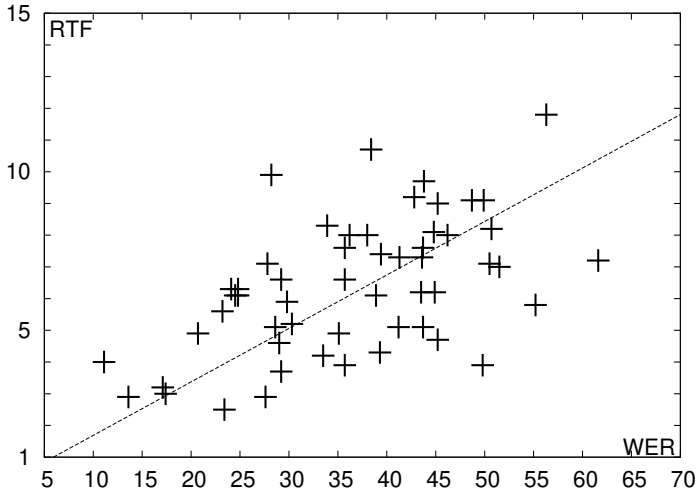

Figure 3.12: RTF versus WER for English-language transcriptions supervised (transLectures Year 3).

Generally speaking, lecturers were satisfied with the interface, however they requested that the TLP Player stopped at the end of each segment to have more time to review the translation. This request was not necessary in automatic transcription supervision, since the cognitive load is notably lower than in the case of translation.

Table 3.10: TER, RTF and SS scores for automatic translation supervisions in Y3.

Language pair	Lecturers	Supervisions	Hours	TER	RTF	SS score
Spanish-English	10	13	2.1	41.9	12.2	8.1

Finally, lecturers demanded higher automatic translation quality.

Correlation between automatic and human evaluations

As expected, supervision times seem to depend on the automatic translation quality (as well as on user expertise).

The linear dependency between automatic evaluation metrics and RTF was not statistically significant for BLEU scores, but it was for TER (p-Value = 3.73×10^{-7}), as we can see in Table 3.11.

Table 3.11: Linear regression model to explain RTF in terms of TER for automatic translation evaluations in Y3.

Model	Beta	p-Value
Es-En ($R^2 = 0.892$, $F = 99.25$, Sig = 3.73×10^{-7})		
$RTF = Beta \times TER$	0.255	3.73×10^{-7}

This is an example of how BLEU and TER, while they usually correlate closely to each other as metrics of automatic translation quality, can diverge sometimes, and so it can be useful to keep track of both for a more robust evaluation.

In this case, this correlation study is probably affected negatively by a not large enough sample of automatic translation supervisions (as seen in Table 3.10).

We can hypothesize that, with a larger sample of automatic translation supervisions, we would find a statistically significant linear dependency between RTF and both BLEU and TER, which would constitute a stronger proof of correlation between automatic and human evaluations. Nevertheless, we have found an indication of this correlation with the statistically significant linear dependency between our RTF and TER measures.

3.4.2.4 transLectures Year 3 user evaluation conclusions

In this section, we have described our findings in Y3 user evaluations at poliMèdia under real life conditions. These evaluations allowed us to test transLectures tools with final users in a real life setting and adjust system functionalities to those demanded by users.

User evaluations at poliMèdia reflected that the supervision time devoted by users measured in RTF can be adequately estimated as a function of WER (for transcriptions) or TER (for translations). In other words, as expected, supervision time (user effort) depends on the automatic transcription and translation quality. However, there are other factors such as user expertise and interface usability that are also involved in determining the supervision time needed to post-edit automatic transcriptions and translations. Regarding transcriptions, significant user effort reductions have been achieved with respect to manually transcribing from scratch: 70%, 40% and 35% for Spanish, English and Catalan, respectively. In addition, the Spanish-English automatic translation supervision task based on the TLP Player also proved to save 60-70% of the time compared to translating from scratch, and provided us with user statistics that allowed us to estimate user effort in terms of RTF as a function of TER. Last but not least, our users in these evaluations, (lecturers, in our case) expressed a clear preference for the post-editing protocol over the more sophisticated, and even more efficient, two-phase protocol.

In practical terms, producing a transcription and translation from scratch for an average-length poliMèdia video of 10 minutes takes approximately 400–500 minutes (approx. 10 RTF for transcription, plus 30–40 RTF for translation). Using transLectures technology, the user would need about 30 minutes (2.7 RTF) to supervise the automatic transcription, plus about 120 minutes (RTF 12.2) to review the translation, that is, 150 minutes. This would mean a time saving of approximately two thirds with respect to doing it from scratch.

Nevertheless, our evaluating users requested higher quality in both automatic transcriptions and translations, to save even more time in their supervision process. This is something we have continued to work on since then, with significant improvements in ASR [40], and also in MT (as we will see in the next Chapter 4), which allow for further reductions in post-editing effort (RTF).

All in all, at the time of transLectures, user evaluations under real life conditions proved that transLectures technology could already produce automatic transcriptions and translations that were accurate enough to be useful for large video lecture repositories.

3.5 Conclusions

In this chapter, we have studied how to evaluate the quality and usefulness of machine translation (MT) systems in several ways, including intelligent interaction approaches to post-editing, by reviewing the work done in this respect within the EU research project transLectures (2011–2014) in a real scenario, the poliMèdia video lecture repository of Universitat Politècnica de València.

Adequate MT (and ASR) quality measurements are key in order to measure progress and to establish how to better proceed along the way to provide high quality multilingual subtitles for Open Educational Resources via ASR and MT, with cost-effective post-editing of the automatic output.

In these conclusions, we sum up what we have learnt about evaluating automatic

(transcriptions and) translations of Open Educational Resources and about post-editing of automatic translations with hybrid intelligent interaction approaches.

The evolution of ASR and MT in 2011–2014

The transLectures project (2011–2014) took place precisely at the time when the use of Deep Neural Networks started to revolutionize the state of the art in ASR. These are the years in which new DNN-based ASR techniques became usable and affordable for more and more research groups, even with relatively modest hardware.

This is why all the evaluations carried out during transLectures showed a continuous and significant improvement in ASR quality (this is nicely summarized in Figure 3.1(a), but the same was shown in the human evaluations in Sections 3.3 and 3.4). On the one hand, more and more quality training data (corpora for ASR) were gathered and exploited during the project. But also, the MLLP research group, like other partners of the transLectures project, started applying these new DNN-based ASR techniques in the project.

Regarding MT (the main focus of this master's thesis), at this point DNN-based neural machine translation techniques were receiving a new impulse after the recent breakthroughs in ASR, but NMT had not yet reached the point of overcoming the results of phrase-based machine translation systems. Thus, in the transLectures project, we can see that improvements in MT by itself were also continuous, but not as striking as in the case of ASR (again Figure 3.1(b) is a summary of this, and the same was shown in the human evaluations in Sections 3.3 and 3.4).

Nevertheless, we should not forget that in this application (automatic transcription and translation of video lectures), the transcriptions obtained through ASR are the first step for their translation through MT; thus, even though MT by itself did not see extraordinary improvements during the project, the transcription-translation pipeline did improve significantly. Much better automatic transcriptions via improved ASR, together with relatively smaller improvements in MT technology, meant: a) significantly better automatic translations from the improved automatic transcriptions; and b) a significant reduction of the effort (and cost) required to post-edit both the automatic transcriptions and translations to reach the expected level of quality for their intended use.

In any case, in the next chapter we will move to 2018, when (as we will see) DNN-based NMT will finally deliver on its promises.

Summary of results by type of evaluation in project transLectures

Automatic evaluations showed that large improvements were achieved by means of massive adaptation techniques in all ASR tasks and most MT tasks (as shown in Figure 3.1).

In human evaluations by language experts, significant improvements were reported in both case studies along the project. All in all, in the final round of human evaluations by language experts (Y2), results were considered really useful in terms of productivity gains by professional editors.

Human evaluations by users were also carried out. On the UPV's poliMèdia repository, UPV lecturers revised automatic transcriptions and translations using the TLP player; in the case of transcriptions and some translation combinations, significant user effort reductions were measured.

Evaluation and post-editing of machine translation in the context of the automatic transcription and translation of Open Educational Resources, with hybrid intelligent interaction approaches

Regarding the measuring of post-editing times as an evaluation method, an *intelligent interaction* approach has been proposed and evaluated within the framework of project transLectures. The challenge was to implement in real use cases this intelligent interaction approach to post-editing, as well as hybrid approaches, and to assess their contribution to cost-effectiveness.

The results we obtained confirm that the intelligent interaction approach can make post-editing automatic transcriptions and translations even more cost-effective. In a series of user evaluations where UPV lecturers post-edited the automatic transcriptions for their own lectures using different approaches (standard post-editing, intelligent interaction, hybrid two-step protocol), we showed that the use of intelligent interaction and hybrid approaches can significantly reduce post-editing times (reducing the time required to as little as 22% compared with transcribing from scratch). On the other hand, in our user evaluations the UPV lecturers expressed their preference for the simplicity and complete control afforded by standard post-editing over the intelligent interaction approaches they tried; however, the fact that the latter were measurably more cost-effective leads us to assume that there are probably other use cases where these approaches will be regarded with more interest (such as professional post-editing). Regarding applicability to MT, while the work in this area presented in this master's thesis is related to post-editing of ASR output and not MT output (mainly because confidence estimation is a more difficult problem in the case of MT and so far there are no established methods), in the context of the automatic transcription and translation of OER it must be taken into account that MT is part of a pipeline that begins with ASR; thus, improvements in the cost-effectiveness of automatic transcriptions lead to an improvement in the cost-effectiveness of automatic translations too.

The need and usefulness of applying different types of evaluation for ASR and MT

In this chapter, we have seen how different types of evaluation can be applied to measure the quality and usefulness of ASR and MT systems. As we have seen, different types of evaluation were used in the transLectures project so as not to rely only on automatic evaluations, which are very important and the first point of reference during research and development, but need to be complemented with human evaluations in order to comprehensively assess the performance of ASR and MT systems in the real world.

Indeed, as ASR and MT systems improved along the transLectures project, the automatic evaluations (Section 3.2) and the human evaluations by language experts (Section 3.3) and users (Section 3.4) confirmed the relevance of the improvements, and the different types of evaluation supported and complemented each other.

In the follow-up EU project EMMA (2014–2016), studies were carried out which confirmed that adding multilinguality to OER can have a strong impact [82]. During this project, 12 EU higher-education partners delivered more than 30 multilingual MOOCs on diverse topics via the EMMA MOOC platform. For the courses not originally in English[11], offering the course in other languages (in English in all cases, plus in other additional languages in some cases) resulted in an average increase of 80% in the number of enrolled students[12].

The multilingual versions of the EMMA MOOCs were created cost-effectively thanks to the use of ASR and MT with post-editing to produce high-quality multilingual subtitles and text documents. Human evaluations allowed for the measurement of post-editing RTF as a measure of the cost of post-editing. Thus, it was possible to show that the high-quality ASR and MT systems used, together with post-editing, allowed for an average decrease of over 50% in transcription and translation effort (compared to doing it from scratch).

3.5.1 Publications and contributions

Publications

The research work covered in this chapter resulted in 3 publications by the author of this master's thesis:

- J. D. Valor Miró, R. N. Spencer, A. Pérez González de Martos, G. Garcés Díaz-Munío, C. Turró, J. Civera, and A. Juan. Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures. *Open Learning*, 29(1):72–85, 2014.

  URL: `http://dx.doi.org/10.1080/02680513.2014.909722`

  Indexed international journal: Q2 in Scimago SJR ("Education", 2014).

- J. D. Valor Miró, R. N. Spencer, A. Pérez González de Martos, G. Garcés Díaz-Munío, C. Turró, J. Civera, and A. Juan. Evaluación del proceso de revisión de transcripciones automáticas para vídeos Polimedia. In *Proc. of the I Jornadas de Innovación Educativa y Docencia en Red (IN-RED 2014)*, pages 272–278, València (Spain), 2014.

  URL: `http://hdl.handle.net/10251/54397`

  National conference on educational innovation and online teaching, organized yearly by Universitat Politècnica de València in Spain.

[11] The other original languages being Spanish, French, Italian and Dutch.

[12] In the case of courses originally in English, while the increase in enrolment because of translations to other languages was lower, it was still noticeable at an average of an 8% increase.

- J. A. Silvestre-Cerdà, M. Á. Del Agua, G. Garcés, G. Gascó, A. Giménez, A. Martínez, A. Pérez, I. Sánchez, N. Serrano, R. Spencer, J. D. Valor, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan. transLectures. In *Proc. of the VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop (IberSpeech 2012)*, pages 345–351, Madrid (Spain), 2012.

  URL: `http://hdl.handle.net/10251/37290`

  International conference on speech and language technologies, organized biennially by the International Speech Communication Association Special Interest Group on Iberian Languages (ISCA SIG-IL) and the Thematic Network on Speech Technologies of Spain (RTTH).

Contributions by the author

The author of this master's thesis was employed as a researcher at Universitat Politècnica de València to work on EU project transLectures (2011–2014) from 2012 until the end of the project in 2014. He worked on the creation and maintenance of corpora for ASR and MT (both for model training and evaluation), and on human evaluations on the poliMèdia pilot: he carried out "quality control" human evaluations as a language expert, tested and evaluated the TLP Player post-editing interface that was developed for user evaluations, participated in compiling and analysing internal and external user evaluations, and participated in drafting the corresponding project reports. His contributions to the publications listed above were on these areas of work.

This work by the author of this master's thesis has contributed: to obtaining new results on evaluation methods for machine translation (and automatic speech recognition) applied in real Open Educational Resource scenarios; to showing that high-quality machine translation (and automatic speech recognition) with post-editing constitutes a cost-effective way of obtaining high-quality transcriptions and translations; and to demonstrating that an intelligent interaction (or hybrid) approach to post-editing can make post-editing automatic transcriptions and translations even more cost-effective.

Chapter 4

The transition from Phrase-Based to Neural Machine Translation

In this chapter, we present work on developing state-of-the-art neural machine translation systems to transition from the previous phrase-based machine translation models, with the aim of making machine translation even more usable and cost-effective for Open Educational Resources[1].

While phrase-based machine translation systems were the state of the art in the field since the 2000s, neural machine translation has made great advances over the last few years, and in particular it came to outperform phrase-based machine translation and PBMT-NMT combinations in the news translation shared tasks of the Second Conference on Machine Translation (WMT17).

In line with scientific progress, until 2017 the machine translation systems developed by the MLLP research group of Universitat Politècnica de València were phrase-based. In 2018, after NMT systems conclusively showed their advantage over PBMT systems, the Third Conference on Machine Translation (WMT18) was an excellent occasion to get up to speed by developing NMT systems to replace the previous PBMT systems, and to compare their accuracy against state-of-the-art NMT systems by other research groups worldwide.

The structure of this chapter is as follows: In Section 4.1, we introduce the context for the work presented in this chapter. In Section 4.2, we describe the datasets that we

[1] Part of the contents of this chapter (in particular, Section 4.3) are adapted from a publication by the author of this master's thesis:

- Javier Iranzo-Sánchez, Pau Baquero-Arnal, Gonçal V. Garcés Díaz-Munío, Adrià Martínez-Villaronga, Jorge Civera, and Alfons Juan. The MLLP-UPV German-English Machine Translation System for WMT18. In *Proc. of the 3rd Conf. on Machine Translation (WMT18)*, pages 422–428, Brussels (Belgium), 2018.

  URL: `http://dx.doi.org/10.18653/v1/W18-6414`

used to train and evaluate PBMT and NMT systems. In Section 4.3, we describe how we developed the MLLP's German→English NMT system for WMT18. In Section 4.4, we conduct a comparison between the MLLP's previous PBMT systems and new NMT system (in the German→English translation combination). In Section 4.5, we describe the impact that these new NMT systems have in real OER scenarios within the framework of EU project X5gon. Finally, in Section 4.6, we bring this chapter to a close with some conclusions.

4.1 Introduction

In 2017, EU Horizon 2020 project X5gon begun with the aim of consolidating innovative technology elements converging currently scattered Open Educational Resources available in various modalities across Europe and the globe, deploying open technologies for recommendation, learning analytics and learning personalization services that work across various OER sites, independent of languages, modalities, scientific domains, and socio-cultural contexts.

The Universitat Politècnica de València, represented by the MLLP research group, is one of the X5gon partners. The UPV's main line of research and development within X5gon is cross-lingual communication, continuing and expanding the work on MT, ASR and TTS for Open Educational Resources that begun with EU project transLectures (as we saw in Chapter 3; see Section 3.1 for more details on transLectures).

Within this context, the MLLP research group developed its first neural machine translation systems, submitting an NMT system to the conference WMT18 in order to assess its performance amongst the best current systems at an international level.

In this section, we briefly introduce the context for the work presented in this chapter, including EU project X5gon and the yearly international conference WMT.

4.1.1 The EU project X5gon

The EU H2020 project X5gon: Cross-Modal, Cross-Cultural, Cross-Lingual, Cross-Domain, and Cross-Site Global OER Network (2017–2021) proposes easily implemented freely available innovative technology elements converging currently scattered Open Educational Resources available in various modalities across Europe and the globe.

X5gon combines content understanding, user modelling quality assurance methods and tools to boost a homogeneous network of (OER) sites and to provide users (teachers, learners) with a common learning experience. X5gon deploys open technologies for recommendation, learning analytics and learning personalization services that work across various OER sites, independent of languages, modalities, scientific domains, and socio-cultural contexts.

Fivefold solutions are offered to OER sites:

- Cross-modal: technologies for multimodal content understanding.
- Cross-site: technologies to transparently accompany and analyse users across sites.

- Cross-domain: technologies for cross domain content analytics.
- Cross-language: technologies for cross-lingual content recommendation.
- Cross-cultural: technologies for cross-cultural learning personalization.

X5gon collects and indexes OER resources, tracks data of user progress and feeds an analytics engine driven by state-of-the-art machine learning, improving recommendations via user understanding and matching it with knowledge resources of all types.

The project will create three services X5oerfeed, X5analytics and X5recommend, and run a series of studies on three main pilot cases, each of them a large multilingual, multimedia OER repository:

- VideoLectures.NET (Institut Jožef Stefan, EU/Slovenia)
- poliMèdia[2] (Universitat Politècnica de València, Spain)
- virtUOS (Universität Osnabrück, Germany)

These pilot case studies will enable the measurement of the broader goals of delivering a useful and enjoyable educational experience to learners in different domains, at different levels and from different cultures. Two exploitation scenarios are planned: 1) free use of services for OER; and 2) commercial exploitation of the multimodal, big data, real-time analytics pipeline.

The X5gon consortium is made up of: University College London (UK; coordinating partner), Institut Jožef Stefan (Slovenia), Knowledge 4 All Foundation (UK), Universitat Politècnica de València (Spain), Université de Nantes (France), Universität Osnabrück (Germany), Post of Slovenia D. O. O., and the Ministry of Education and Science of Slovenia.

The UPV's main line of research and development within X5gon is cross-lingual communication, continuing and expanding the work on MT, ASR and TTS for Open Educational Resources that begun with EU project transLectures.

4.1.1.1 Project languages and language pairs

4 ASR languages and 14 MT language pairs are being covered (by UPV) in the X5gon project:

ASR English, Spanish, German, Slovenian.

MT English↔{Spanish, French, German, Italian, Slovenian},
French↔German, Spanish↔Portuguese.

[2]The UPV's poliMèdia video lecture repository has already been introduced in Section 3.1.2.

4.1.2 WMT: The Conference on Machine Translation

The international Conference on Machine Translation (WMT, from its former name "Workshop on Machine Translation") is one of the main yearly events in the field of MT research, organized in collaboration with the Association of Computational Linguistics. Since its first instance in 2006, every year it has been organized as a series of MT evaluation campaigns (a. k. a. competitions) focused on different aspects of MT research, with the repeated participation of the most important research groups on MT (from both academia and industry), which take it as an occasion to show their advances every year.

It could be said that its main event is the yearly Shared Task on Machine Translation of News. This task allows for participation in several different language combinations, including high-resource and low-resource languages, restricting participants to use only parallel and monolingual corpora provided by the WMT organizers. This "data-constrained" format means that participants compete on the merits of their MT technology, without having to worry about having access to less training data than other participants. The results and ranking of the WMT News Translation Shared Task provide an excellent yearly comparable reference of the state of the MT field.

As already mentioned earlier, the WMT17 edition (2017) is where NMT systems conclusively showed their advantage over PBMT systems [18] (although this was already indicated by the results of WMT16 [19]). The best system in the German↔English tracks was an NMT system based on attentional encoder-decoder networks, using BPE subword segmentation and back-translated monolingual data.

The WMT18 edition (2018) showed a consolidation of NMT architectures and techniques developed in recent years [17]. The best-performing systems in the German↔English tracks were all based on the Transformer multi-headed self-attentional DNN architecture (which had been just been published in 2017).

The MLLP research group of Universitat Politècnica de València took the Third Conference on Machine Translation (WMT18) as an excellent occasion to get up to speed by developing NMT systems to replace their previous PBMT systems, and to compare their accuracy against state-of-the-art NMT systems by other research groups worldwide.

4.2 Datasets

In this section, we describe and compare the German↔English datasets for the WMT17 and WMT18 News Translation Shared Tasks that we used to train and evaluate the PBMT and NMT systems in the following Sections 4.3 and 4.4 in this chapter.

The WMT datasets are some of the largest and most important reference datasets in the field of MT at an international level, so comparing PBMT and NMT systems trained and evaluated on WMT data allows us to obtain robust results.

4.2.1 WMT17 bilingual and monolingual datasets

The German↔English dataset for the WMT17 News Translation Shared Task contains German→English bilingual parallel corpora, and German and English monolingual corpora, which we will all analyse in this section.

Bilingual dataset partition

The German↔English dataset for the WMT17 News Translation Shared Task contains German→English bilingual parallel corpora are summarized in Table 4.1, partitioned in training, dev/validation and test sets.

Table 4.1: WMT17 German→English bilingual dataset partition statistics

	Corpus	Sent. pairs	Words		Vocabulary	
		DE↔EN	DE	EN	DE	EN
Train	News Commentary v12	0.3 M	7.2 M	7.2 M	0.3 M	0.2 M
	Rapid Press Rel. 2016	1.3 M	22.1 M	23.0 M	0.6 M	0.3 M
	Europarl v7	1.9 M	50.5 M	53.0 M	0.4 M	0.1 M
	Common Crawl (2013)	2.4 M	54.6 M	58.9 M	1.6 M	0.8 M
	WMT17 total	5.9 M	134.4 M	142.1 M		
Dev	newstest2015	2.2 K	44.1 K	46.8 K	9.7 K	7.5 K
Test[3]	newstest2017	3.0 K	61.0 K	64.8 K	12.5 K	9.0 K
	newstest2018	3.0 K	54.9 K	58.6 K	16.0 K	13.4 K

The WMT17 bilingual corpora for MT system training add up to a total of 5.9 million sentence pairs, which was relatively large as a publicly available dataset by 2017 standards.

For system testing (the *test* set), we evaluated our MT systems on the hidden test sets of both WMT17 (newstest2017) and WMT18 (newstest2018), as a way to confirm that our MT systems perform consistently on different test sets.

For system validation (the *development* or *dev* set), the test sets of previous editions of WMT (WMT16 and earlier) are available (from newstest2008 to newstest2016). It is usually a good idea to use for validation a test set from recent years. In our case, we have used for validation the set newstest2015 (as we will see later on in Sections 4.3.3 and 4.4.1).

[3]newstest2017 was the blind test set for WMT17, thus it was only made available after WMT17. newstest2018 was the blind test set for WMT18, so it did not exist for WMT17. In this master's thesis these were the test sets used for evaluation of the PBMT systems trained on WMT17 data (Section 4.4), in order to allow for comparison with the NMT experiments on WMT18 data (results are reported on two test sets as a way to confirm that the MT systems perform consistently on different test sets).

Monolingual dataset

The German and English monolingual corpora for the WMT17 News Translation Shared Task are summarized in Table 4.2.

Table 4.2: WMT17 German→English monolingual dataset statistics

	Corpus	Sentences		Words		Vocabulary	
		DE	EN	DE	EN	DE	EN
Train	News Commentary v12	0.3 M	0.4 M	6.9 M	9.4 M	0.3 M	0.3 M
	Europarl v7	2.2 M	2.2 M	53.5 M	59.8 M	0.4 M	0.1 M
	News Crawl 2007–2016	0.2 G	0.2 G	3.9 G	3.8 G	0.02 G	0.01 G
	News Discussions v1–v2	–	0.2 G	–	3.6 G	–	0.03 G
	Common Crawl (2016)	2.9 G	3.1 G	65.2 G	65.1 G	0.3 G	0.3 G
	WMT17 total	3.1 G	3.5 G	69.2 G	72.6 G		

As we can see, the monolingual dataset is much larger than the bilingual dataset, adding up to a total of 3100 million sentences in German and 3500 million sentences in English. This is as expected, since it is much easier to gather monolingual data than parallel bilingual data.

In-domain versus non-domain-specific data

Regarding the domains of the WMT17 bilingual and monolingual corpora, Table 4.3 summarizes the sizes of the in-domain and non-domain-specific data subsets.

Table 4.3: Sizes of in-domain and non-domain-specific corpora in the WMT17 German→English bilingual and monolingual datasets

	Domain	Sentences	
		DE	EN
Bilingual	In-domain	1.3 M	
	Non-domain-specific	4.6 M	
Monolingual	In-domain	221.8 M	166.2 M
	Non-domain-specific	2875.3 M	3325.9 M

Among the WMT17 bilingual corpora, only one of the training corpora is in-domain with respect to the test set (made up of news texts): Rapid EU Press Releases 2016 (with 1.3 million sentence pairs). The rest of the WMT17 bilingual corpora (totalling 4.6 million sentence pairs) are non-domain-specific. As to the WMT17 monolingual corpora, again, only one of the training corpora is in-domain with respect to the test set (made up of news texts), but it is one of considerable size: News Crawl

2007–2016 (DE: 221.8 million sentences; EN: 166.2 million sentences). The rest of the WMT17 monolingual corpora (DE: 2 875.3 million sentences; EN: 3 325.9 million sentences) are non-domain-specific.

Given the large amount of in-domain data that we can obtain from monolingual corpora compared to the much smaller amount of in-domain data from bilingual corpora, it is important to make the most of the monolingual data (via the language model, in PBMT, or via backtranslations, in NMT).

4.2.2 WMT18 bilingual and monolingual datasets

The German↔English dataset for the WMT18 News Translation Shared Task contains German→English bilingual parallel corpora, and German and English monolingual corpora.

In this section, we will first compare the WMT18 datasets against the previous year's WMT17 datasets, and then we will move to an analysis of the WMT18 bilingual dataset partition and monolingual dataset, ending with a brief analysis of the in-domain and non-domain-specific data subsets.

Comparison with the WMT17 datasets

Corpus preprocessing and filtering acquired a new relevance in WMT18, due to the addition of the new parallel corpus ParaCrawl v1.0-filtered[4], which sextuplicated the amount of German↔English parallel data that was available in WMT17 and previous editions: there were approx. 36 million sentence pairs in ParaCrawl v1.0, versus approx. 6 million in the rest of the WMT18 parallel corpora (for a total sum of approx. 42 million sentence pairs in the full WMT18 training data). This is illustrated below in Table 4.4, which summarizes the number of sentences of each corpus in the WMT18 bilingual corpora.

While the large size of the ParaCrawl v1.0 parallel corpus made it a valuable resource for system training in WMT18, it was much noisier than the rest of the WMT corpora. By noise here we mean misaligned sentences, wrong languages, meaningless sentences...; that is, sentence pairs which hinder system training for the purpose of German→English translation. In our experiments, we observed that preprocessing and filtering the ParaCrawl corpus was necessary in order to make it useful as training data with the goal of increasing machine translation quality. In fact, using the ParaCrawl corpus "as is", we not only did not find any improvement in translation quality, but we even observed a degradation in quality metrics (as we will detail in Section 4.3.3.2).

However, if we exclude the ParaCrawl v1.0 corpus, the WMT17 and WMT18 bilingual and monolingual corpora for German↔English are very much comparable. Among the bilingual corpora (without ParaCrawl), only the smallest, non-domain-specific News Commentary corpus was updated (v12 to v13), but this only introduced a difference of 11 500 new sentence pairs in the whole 5.9 million bilingual corpus

[4] v1.0-filtered is the full version name given by the creators of ParaCrawl to the version that was used in WMT18. This is unrelated to the corpus filtering that is later applied in this master's thesis.

(without ParaCrawl). As to the monolingual corpora, 3 of the corpora were updated: News Commentary (v12 to v13), News Discussions (v1–v2 to v1–v3) and News Crawl (2007–2016 to 2007–2017), of which the last one is in-domain; this introduced a total of 39 million new German sentences and 134.6 million new English sentences, in a whole corpus of around 3 100 million German sentences and 3 500 million English sentences.

Bilingual dataset partition

The German↔English dataset for the WMT18 News Translation Shared Task contains German→English bilingual parallel corpora are summarized in Table 4.4, partitioned in training, dev/validation and test sets.

Table 4.4: WMT18 German→English bilingual dataset partition statistics

	Corpus	Sent. pairs DE↔EN	Words DE	Words EN	Vocabulary DE	Vocabulary EN
Train	News Commentary v13	0.3 M	7.2 M	7.2 M	0.2 M	0.1 M
	Rapid Press Rel. 2016	1.3 M	22.1 M	23.0 M	0.6 M	0.3 M
	Europarl v7	1.9 M	50.5 M	53.0 M	0.4 M	0.1 M
	Common Crawl (2013)	2.4 M	54.6 M	58.9 M	1.6 M	0.8 M
	ParaCrawl v1.0-filtered[5]	36.4 M	595.0 M	623.4 M	8.1 M	5.4 M
	WMT18 total	42.3 M	729.4 M	765.5 M		
Dev	newstest2015	2.2 K	44.1 K	46.8 K	9.7 K	7.5 K
Test[6]	newstest2017	3.0 K	61.0 K	64.8 K	12.5 K	9.0 K
	newstest2018	3.0 K	54.9 K	58.6 K	16.0 K	13.4 K

The WMT18 bilingual corpora for MT system training add up to a total of 42.3 million sentence pairs (5.9 million sentence pairs if we exclude the new, noisier ParaCrawl v1.0 corpus).

For system testing (the *test* set), we evaluated our MT systems on the hidden test sets of both WMT17 (newstest2017) and WMT18 (newstest2018), as a way to confirm that our MT systems perform consistently on different test sets.

For system validation (the *development* or *dev* set), the test sets of previous editions of WMT (WMT17 and earlier) are available (from newstest2008 to newstest2017). It is usually a good idea to use for validation a test set from recent years.

[5] v1.0-filtered is the full version name given by the creators of ParaCrawl to the version that was used in WMT18. This is unrelated to the corpus filtering that is later applied in this master's thesis.

[6] newstest2017 was the blind test set for the previous year's WMT17, thus for WMT18 it was available from the beginning as a reference for preliminary evaluations. newstest2018 was the blind test set for WMT18, thus it was only made available after WMT18. In this master's thesis, we report results on both test sets as a way to confirm that the MT system performs consistently on different test sets.

In our case, we have used for validation the set newstest2015 (as we will see later on in Sections 4.3.3 and 4.4.1).

Monolingual dataset

The German and English monolingual corpora for the WMT18 News Translation Shared Task are summarized in Table 4.5.

Table 4.5: WMT18 German→English monolingual dataset statistics

	Corpus	Sentences		Words		Vocabulary	
		DE	EN	DE	EN	DE	EN
Train	News Commentary v13	0.3 M	0.5 M	7.2 M	10.0 M	0.3 M	0.2 M
	Europarl v7	2.2 M	2.2 M	53.5 M	59.8 M	0.4 M	0.1 M
	News Crawl 2007–2017	0.3 G	0.2 G	4.6 G	4.4 G	0.02 G	0.01 G
	News Discussions v1–v3	–	0.4 G	–	5.1 G	–	0.03 G
	Common Crawl (2016)	2.9 G	3.1 G	65.2 G	65.1 G	0.3 G	0.3 G
	WMT18 total	3.1 G	3.6 G	69.9 G	74.7 G		

As we can see, the monolingual dataset is much larger than the bilingual dataset, adding up to a total of 3 100 million sentences in German and 3 600 million sentences in English. This is as expected, since it is much easier to gather monolingual data than parallel bilingual data.

In-domain versus non-domain-specific data

Regarding the domains of the WMT18 bilingual and monolingual corpora, Table 4.6 summarizes the sizes of the in-domain and non-domain-specific data subsets.

Table 4.6: Sizes of in-domain and non-domain-specific corpora in the WMT18 German→English bilingual and monolingual datasets

	Domain	Sentences	
		DE	EN
Bilingual	In-domain	1.3 M	
	Non-domain-specific	41.0 M	
Monolingual	In-domain	260.8 M	193.1 M
	Non-domain-specific	2875.3 M	3433.6 M

Among the WMT18 bilingual corpora, only one of the training corpora is in-domain with respect to the test set (made up of news texts): Rapid EU Press Releases 2016 (with 1.3 million sentence pairs). The rest of the WMT18 bilingual corpora

(totalling 41 million sentence pairs) are non-domain-specific. As to the WMT18 monolingual corpora, again, only one of the training corpora is in-domain with respect to the test set (made up of news texts), but it is one of considerable size: News Crawl 2007–2017 (DE: 260.8 million sentences; EN: 193.1 million sentences). The rest of the WMT18 monolingual corpora (DE: 2 875.3 million sentences; EN: 3 433.6 million sentences) are non-domain-specific.

Given the large amount of in-domain data that we can obtain from monolingual corpora compared to the much smaller amount of in-domain data from bilingual corpora, it is important to make the most of the monolingual data (via the language model, in PBMT, or via backtranslations, in NMT).

4.3 The MLLP-UPV German-English Machine Translation System for WMT18

In this section we describe the development of the neural machine translation system built by researchers of the MLLP research group of Universitat Politècnica de València for the German→English news translation shared task of the EMNLP 2018 Third Conference on Machine Translation (WMT18).

As mentioned in the introduction to this chapter, NMT has made great advances over the last few years, and in particular it came to outperform PBMT and PBMT-NMT combinations in the WMT16 and WMT17 news translation shared tasks [19, 18]. Taking this into account, we decided to build an NMT system taking as a basis the Transformer architecture, which at that point had recently been shown to provide state-of-the-art SMT results while requiring relatively short times to train [84].

German→English was selected as an MT language combination that was relevant and achievable, and an important language combination for EU project X5gon pilots.

Apart from the NMT system itself, here we also describe our work on parallel-corpus preprocessing and filtering, an aspect which gained importance in WMT18 with the addition of the much larger and noisier parallel corpus ParaCrawl. Regarding data augmentation, we report as well how we extended the provided WMT18 parallel dataset with augmented data based on synthetic source sentences generated from the provided target-language monolingual corpora (in compliance with the WMT18 news translation shared task's "constrained" conditions).

This section is organized as follows: in Section 4.3.1, we outline the data preparation techniques that were used (corpus preprocessing, corpus filtering, and data augmentation with synthetic source sentences); Section 4.3.2 shows the architecture and parameters of our NMT system and our system combination; in Section 4.3.3, we report our experiments and results (including on data preparation and on final system evaluation); and we draw our conclusions for this section in Section 4.3.5.

Table 4.7: Size by corpus of the WMT18 parallel dataset

Corpus	Sentences (M)
News Commentary v13	0.3
Rapid 2016 (press releases)	1.3
Common Crawl (2013)	1.9
Europarl v7	2.4
ParaCrawl v1.0-filtered[9]	36.4
WMT18 total	42.3

4.3.1 Data preparation

Here we describe the techniques that we used to prepare the provided WMT18 German↔English data (parallel and monolingual)[7] to improve our NMT system's results: corpus preprocessing, corpus filtering and parallel data augmentation with synthetic source sentences.

Corpus preprocessing and filtering acquired a new relevance in WMT18, due to the addition of the new ParaCrawl parallel corpus (v1.0-filtered)[8], which sextuplicates the amount of German↔English parallel data that was available in WMT17 and previous editions: there are approx. 36 million sentence pairs in ParaCrawl, versus approx. 6 million in the rest of the parallel corpora (for a total sum of approx. 42 million sentence pairs in the full WMT18 training data). This is illustrated in Table 4.7, which summarizes the number of sentences of each corpus in the provided parallel dataset.

While the large size of the ParaCrawl parallel corpus made it a valuable resource for system training in WMT18, it is much noisier than the rest of the WMT corpora. By noise here we mean misaligned sentences, wrong languages, meaningless sentences...; that is, sentence pairs which hinder system training for the purpose of German→English machine translation. In our experiments, we observed that preprocessing and filtering the ParaCrawl corpus was necessary in order to make it useful as training data with the goal of increasing translation quality. In fact, using the ParaCrawl corpus "as is", we not only did not find any improvement in translation quality, but we even observed a degradation in quality metrics (as we will detail in Section 4.3.3.2).

Regarding data augmentation, the usage of relevant in-domain monolingual data has been shown to be important in order to improve NMT system results [64]. The provided WMT18 dataset contains large amounts of monolingual data which can be taken advantage of to increase system accuracy. This fact led us to use these monolingual resources to generate additional synthetic data from target-language sentences.

[7] See Section 4.2 above for the details of the WMT18 parallel and monolingual datasets.

[8] v1.0-filtered is the full version name given by the creators of ParaCrawl to the version that was used in WMT18. This is unrelated to the corpus filtering that is later applied in this master's thesis.

[9] v1.0-filtered is the full version name given by the creators of ParaCrawl to the version that was used in WMT18. This is unrelated to the corpus filtering that is later applied in this master's thesis.

4.3.1.1 Corpus preprocessing

Our preprocessing was done as suggested by the WMT18 organization [87] using the provided scripts, with punctuation normalization, tokenization, cleaning and truecasing using standard Moses scripts.

Additionally, we removed from the training corpus any sentence that contained strange characters, defined as those lying outside the Latin UTF interval (u0000-u20AC) plus the euro sign (€). This allowed us to reduce the vocabulary size by eliminating unnecessary characters belonging to languages other than German or English that are not required for the translation of news.

4.3.1.2 Corpus filtering

In regard to data filtering, we aimed to filter out noisy sentence pairs from the parallel corpora. To this end, we trained two separate 9-gram character-based language models (one for German, one for English) on the newstest2014 development set, based on which we computed the perplexity for each sentence in the full WMT18 dataset (including ParaCrawl), in a manner similar to the techniques described by [90], [28] and [10]. The software used was the SRI Language Modelling Toolkit [73].

The idea was that the lower the perplexity for a given sentence with respect to a reference news test corpus, the lower the odds of this sentence being noise (for the purpose of training a German→English MT system). At the same time, this method could be considered to provide some degree of domain adaptation, since we score the sentences with respect to an in-domain reference corpus.

To produce the final score for each sentence pair, we combined the perplexity scores (s, t) using the geometric mean ($\sqrt{s \cdot t}$). The geometric mean of two character-based perplexities can be interpreted as the character-based perplexity of the concatenation, assuming both sentences have the same number of characters. This is usually not the case exactly, but it is a good enough approximation. As the square root is a monotonic function, it does not alter the order of the scores.

We then ranked all the sentence pairs in the full WMT18 dataset according to their combined perplexity score, and selected subsets of different sizes, taking in each case the n lowest-scored (less noisy) sentence pairs.

Our experimental results are shown in Section 4.3.3.

4.3.1.3 Synthetic source sentences

We augmented the WMT18 German↔English parallel training dataset (while keeping it under "constrained" conditions) with synthetic source sentences generated from the provided target-language monolingual corpora. To this end, we followed the approach outlined by [64].

In particular, we trained an English→German NMT system based on our best system configuration for German→English. Then, we used this system to generate our synthetic source sentences (German) from a subset of the WMT18 target-language monolingual corpora (English), which provided us with a significant amount of new

sentence pairs to use as in-domain synthetic training data. For more specific details on this procedure, see Section 4.3.3.3.

4.3.2 System description

We decided to build an NMT system based on the Transformer architecture [84]. We opted for a pure NMT system due to the great advances this technology has made in the field of SMT over the last few years, which has led it to outperform systematically the more traditional PBMT systems and also PBMT-NMT combinations, as explained in the introduction to this Section 4.3. In particular, the Transformer architecture, based on self-attention mechanisms, can provide state-of-the-art SMT results while keeping training times relatively short. Regarding the software used, we used the Sockeye NMT framework [34].

We based our systems on the less complex Transformer "base" configuration, which has significantly fewer parameters than the "big" configuration (65M parameters in the former case, 213M in the latter), and is thus much quicker to train (in exchange for a relatively small decrease in translation quality, in previous experiments reported in [84]). This was important in order to complete our experiments and the final training of our primary system in time for participation in this WMT18 shared task. Thus, our models used 6 self-attentive layers both on the encoder and the decoder, with a model dimension of 512 units and a feed-forward dimension of 2048 units.

During training, we applied 0.1 dropout and 0.1 label smoothing, the Adam optimization scheme [44] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and learning rate annealing: we set an initial learning rate of 0.0001, and scaled this by a factor of 0.7 whenever the validation perplexity did not improve for 8 consecutive checkpoints (each checkpoint being equivalent to 2000 parameter updates). The system was trained with a word-based batch size of 3000, and a maximum sentence length of 75 tokens (subword units).

For our internal experiments, all systems were trained after applying 20K BPE operations [65]; but when building our final submissions, we increased this amount to 40K BPE operations (this will be detailed for each system in Section 4.3.3.4).

The final system consists of an ensemble of 4 independent training runs of our best model, based on a linear combination of the individual probabilities.

4.3.3 Experimental evaluation

Here we outline our experimental setup (Section 4.3.3.1); we report our experiments and results on corpus filtering (Section 4.3.3.2); we detail our setup for parallel data augmentation with synthetic source sentences (Section 4.3.3.3); and we discuss our final German→English NMT system evaluation and results (Section 4.3.3.4).

4.3.3.1 Experimental setup

For our experiments, we used newstest2015 as the development set and newstest2017 as the test set. We also report the results obtained on WMT18's newstest2018 (as a test set).

Table 4.8: Results in MT quality of 9-gram character-based language model data filtering, by number of parallel sentences selected for training

	Dev set		*Test set*			
	newstest2015		newstest2017		newstest2018	
Training subset (no. of parallel sents.)	BLEU	*TER*	BLEU	*TER*	BLEU	*TER*
Full WMT18 parallel dataset (42M)	20.6	*71.1*	21.3	*70.2*	26.2	*64.2*
Baseline: WMT18 w/o ParaCr. (6M)	31.1	*55.4*	32.0	*54.8*	39.1	*46.3*
Filtered corpus (5M)	30.3	*56.3*	31.4	*55.5*	38.7	*46.5*
Filtered corpus (7.5M)	32.8	*54.0*	33.7	*56.5*	41.5	*44.5*
Filtered corpus (10M)	33.0	*53.7*	**34.5**	*52.9*	**42.2**	*43.7*
Filtered corpus (15M)	33.4	*53.2*	34.3	***52.7***	**42.2**	***43.6***

We evaluated our systems using the BLEU [59] and TER [69] automatic metrics, using mteval from the Moses SMT toolkit [48] and tercom [70], respectively. All reported scores are according to the instructions on system output formatting provided by the WMT18 organization.

4.3.3.2 Results on corpus filtering

We show here the results obtained with the corpus filtering techniques explained in Section 4.3.1.2.

Table 4.8 summarizes the results in translation quality obtained when training with different subsets of the WMT18 parallel dataset. We can see that using the full WMT18 parallel dataset (42M sentence pairs), including the ParaCrawl corpus "as is", led to a significant degradation in quality metrics compared to using the WMT18 dataset excluding ParaCrawl (6M sentence pairs; our baseline system for system evaluations in Section 4.3.3.4). Furthermore, we see that if we restrict ourselves to an excessively small training dataset (5M sentence pairs) using our filtering approach, there is also a degradation in quality with respect to using the unfiltered WMT18 dataset excluding ParaCrawl (6M).

We can also see (focusing on the test set results, newstest2017) that our filtering approach is effective at selecting useful training data from ParaCrawl, in the fact that the filtered datasets with sizes over the baseline's 6M sentence pairs provide significant improvements in quality (even if we limit ourselves to the small increase in size of the 7.5M subset). At the other extreme, in our experiments, going over 15M filtered sentence pairs meant setting the threshold for noise too low, as quality metrics degraded again.

As Table 4.8 shows, we obtained our best test set results with the 10M and 15M subsets. As results were very similar in both cases, we considered that any possible improvements in quality obtained from using the larger 15M subset were too small to justify using it instead of the 33% smaller 10M subset (which has a significantly faster convergence time for system training). Thus, the 10M subset is the filtered training corpus we took as a basis for the subsequent work described in Sections 4.3.3.3 and

4.3.3.4.

As a downside, this data filtering method based on independent language models for each side of a noisy parallel corpus has the caveat of not being able to detect sentence pairs where the source and the target are valid sentences, but not actually a translation of each other. To avoid this problem, it could be useful to combine into the filtering method the score provided by a simple, quick translation model (which should provide better scores for the sentence pairs which are correctly aligned translations). While we carried out some preliminary experiments on filtering with this approach, we did not obtain conclusive improvements in time for this shared task, so we left this for future work.

We also left for future work further corpus filtering experiments with other data selection approaches, such as using the cross-entropy difference (rather than just perplexity or cross entropy) to score each sentence pair [56], or the dynamic data selection method described by [83].

4.3.3.3 Synthetic source sentence setup

Here we detail how we augmented the WMT18 German↔English parallel training dataset, based on the technique introduced in Section 4.3.1.3.

We created an English→German NMT system using our best parameters for German→English (as described in Section 4.3.2), and trained it with the 10M-sentence filtered WMT18 parallel dataset that had shown the best performance for German→English (as described in Section 4.3.3.2). For reference, the resulting English→German NMT system obtained 27.4 BLEU on newstest2017. While improving this "inverse" system with further experiments could result in better synthetic training data [64], we settled on this configuration (which obtains reasonable results with respect to the best WMT17 system scores of 27–28 BLEU) in order to be in time for participation in this shared task.

Then, using this system, we translated into German a random sample of 20M English sentences from News Crawl 2017 (the most recent in-domain corpus among the provided WMT18 monolingual corpora). This provided us with 20M sentence pairs of German→English in-domain synthetic training data.

This augmented corpus was used for the final systems the results of which are discussed in the following Section 4.3.3.4.

4.3.3.4 System evaluation and results

We will now describe the most significant results obtained with the German→English NMT models we trained for WMT18 (based on the architecture and parameters outlined in Section 4.3.2). These results are shown in Table 4.9.

Our baseline model was trained excluding the ParaCrawl corpus from the training data, since using the full WMT18 corpus (with ParaCrawl) actually led to worse results (as we saw in Section 4.3.3.2). As mentioned in Section 4.3.2, this system was trained with 20K BPE operations (as is the case with the next system we will describe).

Table 4.9: Results of German→English NMT system evaluations for WMT18

	Dev set		*Test set*			
	newstest2015		newstest2017		newstest2018	
System	BLEU	*TER*	BLEU	*TER*	BLEU	*TER*
Baseline (WMT18 w/o ParaCrawl, 6M pairs)	31.1	*55.4*	32.0	*54.8*	39.1	*46.3*
Filtered corpus (with ParaCrawl, 10M pairs)	33.0	*53.7*	34.5	*52.9*	42.2	*43.7*
+ Synthetic data (2×10M+20M pairs), 40K BPE	34.3	*52.0*	35.9	*51.2*	44.7	*41.1*
Ensemble (×4)	34.6	*51.9*	**36.2**	***51.0***	**45.1**	***40.8***

Our first step to improve these baseline results was filtering the full WMT18 corpus (including ParaCrawl), as explained in Section 4.3.3.2. In Table 4.9 we show the result obtained with a system trained on our best filtered corpus. As we already saw in Section 4.3.3.2, the 10M filtered corpus provides an improvement of 2.5 BLEU and 1.7 TER in the test set over the baseline model. This shows how our data-filtering approach has allowed us to extract useful sentences from the noisy ParaCrawl corpus and improve our system performance.

For our final systems, we added 20M synthetic sentence pairs as described in Section 4.3.3.3, and we oversampled the previous 10M filtered bilingual training set by duplicating it, which gave us a final training set with a total of 40M sentence pairs[10]. We also increased the number of BPE operations from 20K to 40K. A single system trained with this configuration obtained 35.9 BLEU and 51.2 TER in the test set. This represents a significant improvement of 1.4 BLEU and 1.7 TER over the previous model, explained by a combination of the additional training sentence pairs and the increase in vocabulary size.

As reference of the training times required, training a system with this final configuration took approx. 120 hours on a single-GPU system (Nvidia GeForce GTX 1080 Ti)[11].

Finally, our primary submission for WMT18 consists of an ensemble of 4 independent training runs with this final configuration, resulting in 36.2 BLEU and 51.0 TER in our test set, and 45.1 BLEU and 40.8 TER in newstest2018.

[10] Oversampling the 10M original parallel training set was a measure intended to keep in check the weight of the comparatively large 20M synthetic training data. We left for future work experimenting with different ratios of synthetic versus original data, such as 1:1 [64, 27], as additional comparison terms to determine the best performing configuration.

[11] While our systems were trained on single-GPU machines, multi-GPU system training with proportionally larger batch sizes (larger than the 3000 words per batch we used, as noted in Section 4.3.2) could deliver better translation quality results [84]. We left this for future work.

4.3.4 WMT18 German→English news translation shared task global evaluations and results

Here we can see how the MLLP's NMT system for WMT18 fared in the WMT18 German→English news translation shared task official results [17].

The official ranking of the systems in WMT news shared tasks is not based on automatic evaluation metrics such as BLEU, but on human evaluation. A direct assessment method[12] was followed in WMT18, were human assessors were asked to rate a given automatic translation, at the sentence level, by how adequately it expressed the meaning of the corresponding reference translation (i.e., no bilingual speakers were needed to carry out the evaluation) on an absolute 0–100 rating scale.

Table 4.10 shows the official ranking for the WMT18 German→English news task Right next to it, in Table 4.11 we show the results in BLEU, for comparison.

Some notes for reading these tables:

- In Table 4.10, *Ave. z* (by which the table is sorted) is the system's average human evaluation score standardized according to each individual human assessor's overall mean and standard deviation score, while *Ave. %* is the system's average human evaluation score without standardization.

- The system clusters separated by horizontal lines in Table 4.10, systems in the same group are officially tied, according to which systems significantly outperform all others in lower ranking clusters as determined by the Wilcoxon rank-sum test (at p-level $p < 0.05$).

- A grey background in Table 4.10 indicates commercial MT systems that are included for reference, but which are "out of competition", in that they use training data outside of the official WMT18 corpora. These systems are not included in Table 4.11, since their BLEU results are not available.

[12] See Section 2.2.2.1 for a description of direct assessment among other MT evaluation methods.

Table 4.10: WMT18 DE→EN official results (human eval.)

	German→English		
	Ave. %	Ave. z	System
1	79.9	0.413	RWTH
	79.4	0.395	UCAM
	78.2	0.359	NTT
	77.3	0.346	ONLINE-B
	77.4	0.321	MLLP-UPV
	77.0	0.317	JHU
	76.9	0.315	UBIQUS-NMT
	76.7	0.310	ONLINE-Y
	75.7	0.268	ONLINE-A
	75.4	0.261	UEDIN
11	72.5	0.162	LMU-NMT
	72.2	0.149	NJUNMT-PRIVATE
13	65.2	−0.074	ONLINE-G
14	58.5	−0.296	ONLINE-F
15	45.4	−0.752	RWTH-UNSUPER
16	42.7	−0.835	LMU-UNSUP

Table 4.11: WMT18 DE→EN automatic evaluation results

System	BLEU (newstest2018)
RWTH	48.4
UCAM	48.0
NTT	46.8
JHU	45.3
MLLP-UPV	**45.1**
UBIQUS-NMT	44.1
UEDIN	43.9
LMU-NMT	40.9
NJUNMT-PRIVATE	38.3
RWTH-UNSUPER	18.6
LMU-UNSUPER	17.9

Regarding the different evaluation methods, as we can see, both direct assessment human evaluations and BLEU automatic evaluations produced almost exactly the same ranking, the only difference between the tables being a switch between two participant systems (MLLP-UPV and JHU) which are very close in both human and automatic evaluations (so that their ranking relative to one another is not actually meaningful, since their difference in results is not statistically significant). Thus, we can see a clear correlation between the measurements produced by both evaluation methods.

Going into the ranking, our MLLP-UPV NMT system earned a place in the first rank cluster, tied with the best systems. Looking both at the human and BLEU scores, our system obtains very similar results to the one by the Johns Hopkins University (at ∼45 BLEU), and our difference to the top system by the i6 Human Language Technology and Pattern Recognition Group of RWTH Aachen University (7% relative BLEU) is noticeable, but small enough to grant us a place in their same first rank.

4.3.5 Conclusions for WMT18

The MLLP research group of the Universitat Politècnica de València participated in the WMT18 German→English news translation shared task with an ensemble of neural machine translation models based on the Transformer architecture. Our models were trained on a filtered subset of the provided WMT18 parallel training dataset, plus augmented parallel data based on synthetic source sentences generated from the provided WMT18 monolingual corpora. Our primary submission was an ensemble of four independent training runs of our best model parameters. Our system, which required a relatively short time for training, was shown to be competitive in its results,

classifying among the first rank in the WMT18 German→English news task's official results.

Our results point to the usefulness of the Transformer NMT architecture to obtain highly competitive MT results with a relatively low computational cost (which can contribute to "democratizing" access to state-of-the-art research in NMT to a higher number of research groups, even those with more modest computational equipment). We also showed the importance of adequate corpus filtering to make the most of larger, noisier parallel corpora, employing a simple and quick, yet effective, approach to filtering using character-based language models that resulted in significant improvements in translation quality.

4.4 Comparing Phrase-Based and Neural MT systems

In this section, we compare, for the translation combination German→English, the most advanced phrase-based machine translation system developed by the MLLP research group until 2017, against the 2018 neural machine translation system developed as shown in Section 4.3.

We will first briefly introduce the experimental setup, followed by the PBMT and NMT system descriptions, and then by a separate analysis of each system's results, finishing with the comparison of both systems' results.

4.4.1 Experimental setup

We will compare the BLEU and TER automatic scores obtained by these DE→EN PBMT and NMT systems trained on comparable training sets (WMT17 and WMT18), and applied to translate the same test sets (newstest2017 and newstest2018).

As explained in Section 4.2.2, the basic WMT17 and WMT18 bilingual corpora (excluding WMT18's new ParaCrawl parallel corpus), while not exactly equal, are almost the same, and so our baseline PBMT and NMT systems are comparable.

4.4.2 System descriptions

The PBMT system was trained for the WMT17 News Translation Shared Task. It was trained on the WMT17 corpus, with validation on the WMT newstest2015 corpus, using then state-of-the-art PBMT techniques.

As to the NMT system, it was trained for the WMT18 News Translation Shared Task (as seen in Section 4.3). It was trained on the WMT18 corpus, with validation on the WMT newstest2015 corpus, using new state-of-the-art NMT techniques (as seen in Section 4.3).

As explained in Section 4.2.2, the basic WMT17 and WMT18 bilingual corpora (excluding WMT18's new ParaCrawl parallel corpus), while not exactly equal, are almost the same, and so our baseline PBMT and NMT systems are comparable.

4.4.2.1 PBMT system description

The PBMT system is composed of two types of models: the translation models (trained on bilingual data) and the language models (trained on monolingual data). During decoding (when the trained models are applied to generate an automatic translation), the scores provided by each model are combined as a log-linear model, the weights for each component model being interpolated using MERT [15].

Language model The language model (LM) in a PBMT system models the target language (in this case, English). Thus, the LM estimates the probability of a sequence of words in the target language.

The LM is a linear mixture of n-gram models[13] trained on the textual resources available [14] for the target language (in this case, English). More precisely, this PBMT system's LM is based on n-gram models estimated for each available individual resource. The individual LMs trained for each resource were then combined using a linear mixture optimized on textual content extracted from the domain of our application, what is generally referred to as *in-domain text*; in this case, newstest2015, one of the news texts available in the WMT17 dataset.

The LM training process itself was performed in two steps: vocabulary selection and LM estimation proper. For the vocabulary, the 200,000 most probable words were selected (based on their probability in a unigram LM), keeping the vocabulary at a reasonable size while out-of-vocabulary words remained under 1%[14]. After creating the vocabulary, the LM was estimated: a 4-gram smoothed LM using modified Knesser-Ney discount [22] was trained for each individual resource; individual LMs were then combined through a linear interpolation optimized on in-domain text; finally, for the sake of efficiency, n-grams[15] below a certain probability threshold were pruned out to produce the final LM.

The training of the LM for this PBMT system was performed with the SRILM toolkit [72] (see Section 2.3.2.5 for more details on SRILM and language modelling software).

Translation model The translation model was trained using the Moses toolkit [48], which implements statistical phrase-based log-linear models among others (see Section 2.3.2.1 for more details on Moses and PBMT software). A phrase-based translation model is built by extracting bilingual phrases (understood as segments of consecutive source-target words) from sentence-aligned parallel corpora[16]. Then, several

[13] An n-gram model is a probabilistic model which estimates the probability of a sentence as the probability of consecutive groups of up to n words [22]. See also Section 2.2.1 for a brief introduction to the concept of n-gram.

[14] The higher the number of words in the vocabulary, the larger the amount of data required to reliably estimate the LM. However, if the selected vocabulary is too small, the words that are not included in it will not be considered for the MT output. These words are known as out-of-vocabulary (OOV) words. Therefore, a trade-off between the size of the vocabulary and the amount of data available to train the LM has to be reached.

[15] With $n > 1$.

[16] In a preliminary step, word alignments are established from the sentence-aligned parallel corpora; then, phrases are extracted from the word-aligned corpora

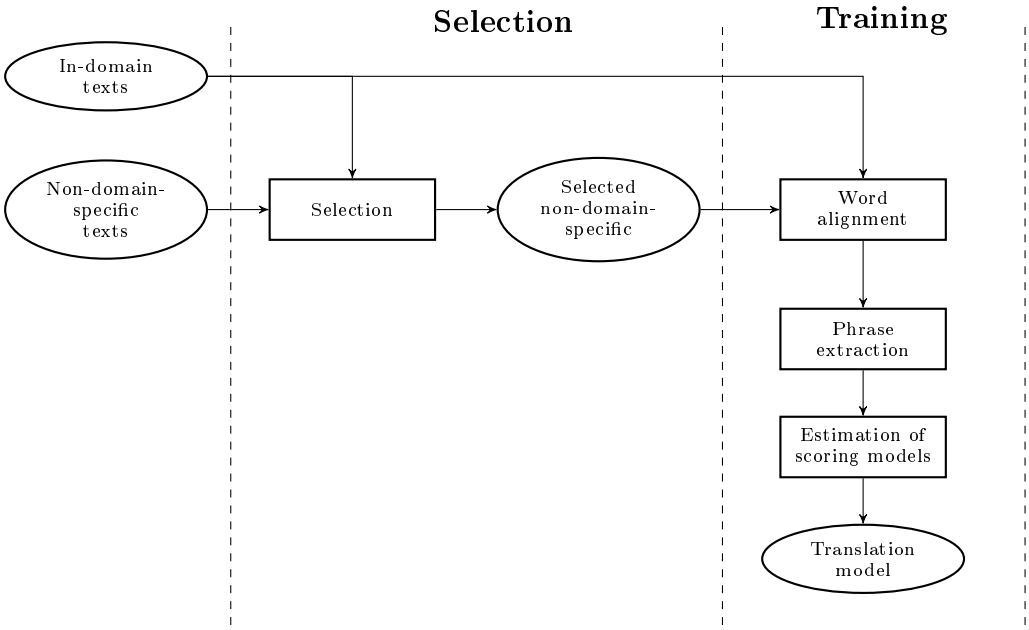

Figure 4.1: PBMT system training: Estimation of the translation model

scoring models are estimated from the extracted bilingual phrases.

Intelligent selection techniques were used to select from the non-domain-specific parallel corpora[17] bilingual sentences that are useful to train an in-domain PBMT system in order to maximize MT automatic quality metrics [89]. Given a small quantity of monolingual in-domain text (news texts available in the WMT17 dataset) and a large quantity of non-domain-specific parallel text (all other bilingual corpora in the WMT17 dataset), a two-step process is followed. First, a monolingual selection technique, Infrequent N-gram Selection or INS [30], is applied using monolingual in-domain text to select representative parallel text from the non-domain-specific corpora. Next, a bilingual language model cross-entropy selection technique [10] is applied to once again select representative parallel text from non-domain-specific text using the in-domain parallel text obtained in the previous step. The selected parallel texts are then employed to train the translation model.

The estimation of the translation model is depicted in Fig. 4.1.

4.4.2.2 NMT system description

The NMT system, developed for the WMT18 News Translation Shared Task, is based on the Transformer multi-head self-attentional DNN architecture, as described in

[17] Also known as out-of-domain.

Table 4.12: Comparison of German→English PBMT and NMT system evaluations

	Dev set		*Test set*			
	newstest2015		newstest2017		newstest2018	
System	BLEU	*TER*	BLEU	*TER*	BLEU	*TER*
PBMT baseline (WMT17, only bilingual data, 6M pairs)	24.3	*56.5*	24.0	*57.1*	28.0	*51.0*
+ Monolingual data (6M+210M pairs)	25.8	*55.3*	**26.0**	***55.5***	**30.6**	***49.2***
NMT baseline (WMT18 w/o ParaCrawl, 6M pairs)	31.1	*55.4*	32.0	*54.8*	39.1	*46.3*
Filtered corpus (with ParaCrawl, 10M pairs)	33.0	*53.7*	34.5	*52.9*	42.2	*43.7*
+ Synthetic data (2×10M+20M pairs), 40K BPE	34.3	*52.0*	35.9	*51.2*	44.7	*41.1*
NMT ensemble (×4)	34.6	*51.9*	**36.2**	***51.0***	**45.1**	***40.8***

Section 4.3.2.

4.4.3 System results

In Table 4.12 we can see the results in BLEU and TER for both the PBMT and NMT system on the test sets newstest2017 and newstest2018.

As additional points of reference, in Table 4.13 we can see our best PBMT and NMT systems compared to the best PBMT and NMT systems in WMT17 and WMT18.

4.4.3.1 PBMT system results

The baseline PBMT system is trained (both its translation model and its language model) exclusively on the WMT17 bilingual corpus (approx. 6 million sentence pairs), without using the WMT17 monolingual corpora. With this we achieve a BLEU of 24.0 on newstest2017 and 28.0 on newstest2018.

The best MLLP PBMT system takes advantage not only of the WMT17 bilingual corpus, but also of the much larger monolingual corpus (210 million sentence pairs, i.e., 35 times bigger). This is a very natural procedure in PBMT system training, where monolingual corpora can be used to improve the training of the language model. In this way, we achieve a BLEU of 26.0 on newstest2017 and 30.6 on newstest2018; that is, a relative increase of 8% and 9%, respectively.

Table 4.13: Comparison of German→English PBMT and NMT system evaluations with reference to the best PBMT and NMT systems in WMT17 and WMT18

System	*Test set* newstest2017 BLEU	*Test set* newstest2018 BLEU
MLLP PBMT baseline (WMT17, only bilingual data, 6M pairs)	24.0	28.0
MLLP PBMT best system (6M biling.+210M monoling. pairs)	**26.0**	**30.6**
WMT17 PBMT best system (RWTH Aachen; Phrasal JTR deco. w/ NMT attention)	31.0	N/A
WMT18 PBMT best system (Boğaziçi Univ.; parfda Moses PBMT)	N/A	33.4
MLLP NMT baseline (WMT18 w/o ParaCrawl, 6M biling. pairs)	32.0	39.1
WMT17 NMT best system (UEdinburgh; BPE NMT ensemble)	35.1	N/A
MLLP NMT best system (w/ ParaCrawl, 2×10M+20M pairs, ×4 ensemble)	**36.2**	**45.1**
WMT18 NMT best system (RWTH Aachen; Transformer ensemble, 24M+18M)	39.9	48.4

As we just mentioned, in the case of PBMT using both bilingual and monolingual data is straightforward, since we train separate system components that we later combine, one of which is the (target) language model, where we only use monolingual target-language data. Thus, training a baseline system based solely on bilingual data is somewhat artificial, but this will allow us to compare results with the baseline NMT system, where the most straightforward thing is actually to train using only bilingual data.

Regarding comparison with the best PBMT systems in WMT17 and WMT18, from Table 4.13 we can notice that our best PBMT system obtained noticeably worse results compared to them (−8–16% in BLEU), which means there were still refinements and techniques to be tried to make the most of the available data with a PBMT system (we should notice, however, that the WMT17 RWTH system already uses some NMT-related techniques, such as NMT attention).

4.4.3.2 NMT system results

The results for the NMT system are as described in Section 4.3.3.4.

The baseline NMT system was trained only on the WMT18 bilingual corpus (approx. 6 million sentence pairs) without the new ParaCrawl corpus (since the full ParaCrawl corpus worsened results), and without using the WMT18 monolingual corpora (since, in NMT system training, training a system using bilingual corpora is the most straightforward approach). With this we obtain 32.0 BLEU on newstest2017 and 39.1 BLEU on newstest2018.

We then improved the NMT system by adding appropriately filtered data from the large, bilingual ParaCrawl corpus, almost doubling the amount of clean bilingual training data (reaching now 10 million sentence pairs). BLEU scores thus improved by around 8% relative from the NMT baseline (on both test datasets).

The NMT system was improved again by adding data from the WMT18 monolingual corpora. In the case of NMT, additional monolingual data can be used to improve the system through the technique of synthetic source sentences or *backtranslations* (see Section 4.3.1.3 for more details). 20 million sentences of monolingual training data were used (2 times the size of the bilingual training data), producing a 4–6% relative increase in BLEU scores.

Finally, 4 independent training runs of this last NMT system were combined using the technique of system ensembling. This results in an improvement in BLEU scores of 0.8–0.9% relative. While the improvement is limited, to the point of not being statistically significant, ensembling is still an interesting technique to use, since by its nature it can result in a system that can better generalize when used to translate different texts.

Regarding comparison with the best NMT systems in WMT17 and WMT18, from Table 4.13 we can notice that our best NMT system obtains better results (+3%) than the WMT17 UEdinburgh system, owing probably to our use of the WMT18 ParaCrawl bilingual data that was not available in WMT17. On the other hand, our best NMT system obtained worse results (−7–9%) than the WMT18 RWTH system, which again means there were still refinements and techniques to be tried to make the most of the available data with an NMT system.

4.4.4 System results comparison

The first thing we should compare are the baseline PBMT and NMT systems (first and third line in Table 4.12, respectively), which are trained on comparable corpora. The baseline PBMT system is trained (both its translation model and its language model) only on the WMT17 bilingual corpus (without using the WMT17 monolingual corpora). As to the baseline NMT system, it is trained only on the WMT18 bilingual corpus (without the new ParaCrawl corpus, and again ignoring the WMT18 monolingual corpora). Here we already see that the NMT system, just because of the NMT techniques applied and not because of differences in the data used, provides a very significant relative improvement of 33–40% in terms of BLEU.

As we saw above in this same section, the second line of the NMT part of Table 4.12 shows the improvement (+8% relative in BLEU) brought by adding appropriately

filtered data from the large ParaCrawl corpus. It must be taken into account that, from this point on, the PBMT and NMT scores reported are not directly comparable, since they are trained on different corpora (a larger parallel corpus in the case of the filtered WMT18+ParaCrawl NMT system).

Nevertheless, the next lines in both the PBMT and the NMT systems still provide an interesting comparison. In the case of the PBMT system, in its second line we improve it by adding to the training data the WMT17 monolingual corpora (which is 35 times bigger than the WMT17 bilingual corpus)[18], which results in an 8–9% relative increase in BLEU scores. As to the NMT system, in its third line we also improve the system by adding data from the corresponding WMT18 monolingual corpora (and by increasing the number of BPE operations from 20K to 40K). In the case of NMT, additional monolingual data are used to improve the system through the technique of backtranslations (see Section 4.3.1.3 for more details). Even though in this case we use a smaller amount of monolingual data with respect to the size of the bilingual training corpus (here the monolingual training data used is only 2 times the size of the bilingual training data), we still obtain a noticeable, and comparable in magnitude, 4–6% relative increase in BLEU scores.

This last comparison lets us see how important it was with PBMT systems the use of as much monolingual data as possible to train the language model, and how the use of monolingual data to train NMT systems, while less straightforward, has been successfully implemented through the use of the now commonplace technique of backtranslations[19].

The fourth and final line in the NMT part of Table 4.12 reflects a minor improvement in the NMT system, by using the technique of system ensembling. While in these results the improvement in BLEU scores is limited (0.8–0.9% relative), to the point of not being statistically significant, ensembling is still an interesting technique to use, since by its nature it can result in a system that can better generalize when used to translate different texts.

Finally, we can now compare our best PBMT system with our best NMT system (the results in bold in Table 4.12), taking into account that the improvement does not come only from employing NMT technology, but also from being able to exploit the new parallel training data provided by the ParaCrawl corpus. As we can see, in 2018 we obtained with NMT and newly available parallel training data a very significant relative improvement of 39–47% in BLEU scores over what was possible for the MLLP research group in 2017 with PBMT and the WMT17 training data available then.

Regarding comparison with the best PBMT and NMT systems in WMT17 and WMT18, it is interesting to notice from Table 4.13 that our baseline NMT system

[18] In PBMT, monolingual corpora are used straightforwardly to improve the training of the language model.

[19] Research is still ongoing on determining the right amount of (monolingual-data-based) backtranslations with respect to the amount of bilingual data. As we have seen, a relatively small amount of correctly crafted backtranslations can provide noticeable improvements. Contrary to what was the case with PBMT language models, in the case of NMT it seems there is such thing as using too much monolingual data, since backtranslations are not, by definition, "clean" training data, and so having too large a proportion of them can at some point stop improving or even begin degrading MT evaluation scores.

(which does not use the WMT18 ParaCrawl bilingual corpus, which means it is based on a corpus comparable to WMT17, and has the disadvantage of not using any monolingual training corpora) is not only better than our best PBMT system, but it also obtains better results (+3–17%) than the best WMT17 and WMT18 PBMT systems. This shows an improvement based solely on the application of NMT technology, even overcoming a disadvantage in the amount of training data used (no monolingual corpora in our NMT baseline system).

Looking at the big picture, we can see that the comparison of the results between PBMT and NMT is consistent with what can be seen in the scientific literature: NMT brings noticeable improvements in BLEU scores (to the point that the baseline NMT system is noticeably better, +23% in BLEU, than the best PBMT system). We can see that this is consistent in Table 4.12 for both test sets that we run our evaluations on (newstest2017 and newstest2018)[20].

As to automatic MT evaluation metrics, in Table 4.12 we can notice the consistency between BLEU and TER as indicators of MT output quality: as BLEU MT quality scores increase, the TER error rates decrease. However, the proportion of these variations is not always exactly the same in BLEU as in TER, which comes to show that there are differences between the way they measure MT output quality, and thus confirms the usefulness of computing several different MT metrics in order to confirm that their evolution is mutually consistent (or, otherwise, to detect possible issues with the MT systems).

4.5 Impact of Neural Machine Translation systems on real Open Educational Resource scenarios

As explained in Section 4.1.1, the Universitat Politècnica de València, represented by the MLLP research group, is one of the EU project X5gon partners. The UPV's main line of research and development within X5gon is cross-lingual communication, continuing and expanding the work on MT, ASR and TTS for Open Educational Resources that begun with EU project transLectures. This means developing and improving systems for the languages of the X5gon pilots (Universitat Politècnica de València's poliMèdia, Institut Jožef Stefan's VideoLectures.NET, Universität Osnabrück's virtUOS), including English, Spanish, German and Slovenian for ASR; and English↔{Spanish, French, German, Italian, Slovenian}, French↔German, and Spanish↔Portuguese for MT.

Focusing on MT, the experience gathered while developing the new MLLP NMT systems for WMT has been applied to develop new NMT systems for other language combinations within X5gon (including the UPV's poliMèdia repository); in particular,

[20] Regarding the difference in results between newstest2017 and newstest2018 for a given system, we can see that newstest2018 was an "easier" dataset for MT, given that using the same MT systems, noticeably better scores are consistently obtained for newstest2018. This is a good example of why it is important to compare results for different MT systems on the same test set. It would not be fair to compare the BLEU score of a given MT system on newstest2017 with the BLEU score of a different MT system on newstest2018.

language combinations with English (EN↔ES,FR,DE,IT,SL) and without English (ES↔PT, DE↔FR)).

This allows us to show the impact that these new neural machine translation systems have in real Open Educational Resource scenarios, within the framework of EU project X5gon.

4.5.1 X5gon NMT systems

The MT systems for Year 1 of X5gon (2018) were already NMT systems developed to replace the previous PBMT systems by the MLLP research group. The first NMT systems developed were all along the lines of the DE→EN NMT system for WMT18 (Transformer "base" trained on 1 GPU), trained on publicly available bilingual and monolingual corpora and also on some in-domain corpora (lecture subtitles) available to UPV and to other X5gon partners (using careful data filtering and backtranslations from monolingual data). This already provided significant improvements in terms of BLEU over the previous PBMT corpora (when available; in some language combinations, such as English→German, this was the first time UPV developed MT systems for them).

The next improvements over these NMT systems came, on the NMT parameters side, from using the more complex Transformer "big" model parameters, trained on multi-GPU computers with the gradient accumulation technique in order to increase training batch sizes, and on the training data side, from adding larger and better corpora (such as new versions of ParaCrawl) and applying more sophisticated data filtering techniques.

Also, from the additional experience in developing NMT systems for WMT19 (2019) [35], a training with fine-tuning technique was incorporated for domain adaptation using in-domain bilingual data, where available (such as in Spanish↔Portuguese).

4.5.2 X5gon NMT system results

To represent the progress of BLEU scores as NMT systems were improved in the X5gon project, we focus our interest in the poliMèdia pilot, with Table 4.14 showing the progress of BLEU scores provided by X5gon NMT systems from X5gon M12 to M30 on the poliMèdia task for the poliMèdia language combinations in X5gon (Spanish→English).

Table 4.14: Progress of BLEU scores provided by X5gon NMT systems from X5gon M12 to M30 on the poliMèdia task for the poliMèdia language combinations in X5gon, with relative improvement from M12 to M30 (Δ)

Language pair	pM			
	M12	M24	M30	Δ
ES-EN	24.3	30.0	34.1	+40%

Reaching a BLEU score over 30, as is the case for this poliMèdia language combination, is considered a representative threshold for MT outputs to be of practical

use for post-editing.

All X5gon NMT systems were improved several times from X5gon M12 to M30.

To summarize MT results for all X5gon MT language combinations, and for comparative reference, Table 4.15 shows BLEU scores provided by Google Translate and X5gon NMT systems, on the VideoLectures.NET (VL) and WMT tasks, for each language pair considered. Note that rows are sorted by average relative improvement for the X5gon NMT systems over Google Translate (Avg. Δ%). Table 4.15 includes the latest MT results achieved from in X5gon M30.

Table 4.15: BLEU scores provided by Google Translate and X5gon NMT systems, on the VideoLectures.NET (VL) and WMT tasks, for each language pair considered. Rows are sorted by average relative improvement for the X5gon NMT systems over Google Translate (Avg. Δ%)

Lang.	VL			WMT			Avg.
pair	Google	X5gon	Δ%	Google	X5gon	Δ%	Δ%
ES-PT	-	-	-	43.4	**70.7**	62.9	62.9
PT-ES	-	-	-	47.6	**72.4**	52.1	52.1
SL-EN	15.0	**26.4**	76.0	29.2	**34.3**	17.4	46.7
EN-SL	16.5	**22.9**	38.8	23.6	**29.4**	24.6	31.7
DE-EN	25.7	**27.0**	5.1	43.9	**48.0**	9.4	7.3
ES-EN	37.8	**40.3**	6.6	34.4	**35.9**	-2.0	2.3
EN-ES	41.3	**44.4**	7.5	**35.3**	34.6	-2.0	2.8
FR-EN	**30.3**	30.1	-0.7	38.6	**39.7**	2.8	1.1
DE-FR	**19.6**	18.6	-5.1	32.2	**34.4**	6.8	0.9
EN-FR	**29.4**	29.2	-0.7	40.4	**41.1**	1.7	0.5
IT-EN	-	-	-	**35.7**	35.2	-1.4	-1.4
FR-DE	**18.6**	17.2	-7.5	26.6	**26.9**	1.1	-3.2
EN-IT	-	-	-	**32.1**	29.8	-7.2	-7.2
EN-DE	**24.7**	21.5	-13.0	**47.0**	45.7	-2.8	-7.9

As shown in Table 4.15, X5gon MT systems for Spanish↔Portuguese and Slovenian↔English provide far better results than those provided by Google Translate. On the other hand, the results for all other language pairs are more or less on par with those by Google Translate.

The progress made in X5gon Y2 and completed during the first half of Y3, allowed the MLLP research group to catch up with the high quality Google Translate is delivering in recent years. Moreover, this made clear that it is certainly possible to go far beyond Google Translate's quality by *adapting* MT systems to the X5gon domain, especially for language pairs with comparatively less support from Google Translate. This is the case of the Slovenian↔English pairs, which are very relevant to X5gon.

In line with the comparison to Google Translate, the NMT systems developed in the framework of X5gon were also thoroughly assessed in international competitions such as WMT19 [13, 35]. In this regard, these MT systems were ranked on average among the top five systems including those from strong academia (UCambridge,

RWTH Aachen, DFKI, JHU, UEDIN, LIUM) and industry (Microsoft, Facebook) representatives.

4.5.3 Conclusions on impact of NMT systems on real OER scenarios

We have shown the impact that these new neural machine translation systems have in real Open Educational Resource scenarios within the framework of EU research project X5gon. The experience gathered while developing the new MLLP NMT systems for WMT has been applied to develop new NMT systems for other language combinations within X5gon (including the UPV's poliMèdia repository) and other MLLP projects (including combinations with English (EN↔ES,FR,DE,IT,SL) and without English (ES↔PT, DE↔FR)). In the case of poliMèdia, improvements in the ES↔EN NMT system have increased its BLEU results on poliMèdia videos by over 40% from 2018 to 2019. Furthermore, a comparison with Google Translate on several datasets showed that our results are on par with the high quality of recent Google Translate results (and also showed that it is possible to go beyond Google Translate's quality through domain adaptation of MT systems, especially for language pairs with comparatively less support from Google Translate).

4.6 Conclusions

This chapter has revolved around the development of state-of-the-art neural machine translation systems to transition from the previous phrase-based machine translation models, with the aim of making machine translation even more usable and cost-effective for Open Educational Resources.

We have presented how we developed a German→English NMT system based on the multi-head self-attention Transformer DNN architecture, in the framework of the Third Conference on Machine Translation (WMT18). This system obtained a relative improvement of 39–47% (in terms of BLEU score on the reference corpora newstest2017 and newstest2018) over the previous MLLP PBMT system, and a relative improvement of 3% over the best performing NMT system in WMT17 (University of Edinburgh), taking advantage of the additional bilingual training data available in WMT18 (ParaCrawl corpus). In terms of the performance of our new NMT system among the WMT18 news translation shared task systems, its BLEU score was 7% relative below the best system at WMT18 (RWTH Aachen), which was close enough to classify our NMT system among the first-rank cluster of participants[21].

These research results were published in an article in the international Third Conference on Machine Translation WMT18 (see the publication's details below in this same section).

We have also carried out a more detailed comparison between the new NMT system developed and the previous MLLP PBMT system. This has allowed us to

[21] See Tables 4.10 and 4.11 in Section 4.3.4 for the full classification in terms of automatic evaluation and human evaluation.

appreciate the significant improvement in quality metrics obtained just by the application of NMT technology (+33–40% BLEU in the baseline system), the ways in which monolingual corpora can be exploited to improve baseline systems based only on parallel corpora (by way of the language model in the case of PBMT, by way of backtranslations in the case of NMT), and the final improvement obtained by our best WMT18 NMT system over the MLLP's best WMT17 PBMT system (+39–47% BLEU, as mentioned above).

Finally, we have shown the impact that these new NMT systems had in real Open Educational Resource scenarios within the framework of EU research project X5gon. The experience gathered while developing the new MLLP NMT systems for WMT has been applied to develop new NMT systems for other language combinations within X5gon (including the UPV's poliMèdia repository) and other MLLP projects; in particular, language combinations with English (EN↔ES,FR,DE,IT,SL) and without English (ES↔PT, DE↔FR)). In the case of poliMèdia, improvements in the ES↔EN NMT system have increased its BLEU results on poliMèdia videos by over 40% from 2018 to 2019. Furthermore, a comparison with Google Translate on several datasets showed that our results are on par with the high quality of recent Google Translate results (and also showed that it is possible to go beyond Google Translate's quality through domain adaptation of MT systems, especially for language pairs with comparatively less support from Google Translate).

4.6.1 Publications and contributions

Publications

The research work covered in this chapter resulted in a publication by the author of this master's thesis:

- Javier Iranzo-Sánchez, Pau Baquero-Arnal, Gonçal V. Garcés Díaz-Munío, Adrià Martínez-Villaronga, Jorge Civera, and Alfons Juan. The MLLP-UPV German-English Machine Translation System for WMT18. In *Proc. of the 3rd Conf. on Machine Translation (WMT18)*, pages 422–428, Brussels (Belgium), 2018.

  URL: `http://dx.doi.org/10.18653/v1/W18-6414`

  International conference on machine translation, organized yearly by the Association for Computational Linguistics (ACL).

Contributions by the author

The work in this chapter was carried out while the author of this master's thesis held a predoctoral research grant at Universitat Politècnica de València (2018–2021). He contributed to the publication listed above as part of his predoctoral research plan "Application of Deep Neural Networks to Machine Translation for Open Educational Resources". He participated in reviewing previous publications to identify the state-of-the-art practices for developing neural machine translation systems; in testing available frameworks for neural machine translation in order to select the most

appropriate one for this publication's work; in training neural machine translation systems and running experiments; and he coordinated the writing of the publication.

Section 4.3 in this chapter is based on this publication; Section 4.4 adds additional work by the author on comparing the new NMT system developed against the previous PBMT system for the same language combination (in order to obtain a better perspective on what NMT techniques bring to the table). In the sections before we have provided the context for this work, and in the sections after we have described the impact that these new NMT systems have in real OER scenarios, evaluating the NMT systems developed based on the first NMT system presented.

Chapter 5

Achievements and conclusions

5.1 Achievements

1. **Contributions to evaluation and post-editing of machine translation in the context of the automatic transcription and translation of Open Educational Resources, with hybrid intelligent interaction approaches**

   In this master's thesis, we have presented a review of work on evaluation of machine translation in the context of automatic transcription and translation of Open Educational Resources in a real scenario, the EU research project transLectures and its pilot cases. We have analysed how different methods of evaluation have different applications, how they complement each other to provide a clearer and more complete assessment, and we have seen how they are used to evaluate the cost-effectiveness of automatic transcriptions and translations.

   Regarding the measuring of post-editing times as an evaluation method, an *intelligent interaction* approach has been proposed and evaluated within the framework of project transLectures. The challenge was to implement in real use cases this intelligent interaction approach to post-editing, as well as hybrid approaches, and to assess their contribution to cost-effectiveness.

   The results we obtained confirm that the intelligent interaction approach can make post-editing automatic transcriptions and translations even more cost-effective. In a series of user evaluations where UPV lecturers post-edited the automatic transcriptions for their own lectures using different approaches (standard post-editing, intelligent interaction, hybrid two-step protocol), we showed that the use of intelligent interaction and hybrid approaches can significantly reduce post-editing times. On the other hand, in our user evaluations the UPV lecturers expressed their preference for the simplicity and complete control afforded by standard post-editing over the intelligent interaction approaches they

tried; however, the fact that the latter were measurably more cost-effective leads us to assume that there are probably other use cases where these approaches will be regarded with more interest (such as professional post-editing). Regarding applicability to MT, while the work in this area presented in this master's thesis is related to post-editing of ASR output and not MT output (mainly because confidence estimation is a more difficult problem in the case of MT and so far there are no established methods), in the context of the automatic transcription and translation of OER it must be taken into account that MT is part of a pipeline that begins with ASR; thus, improvements in the cost-effectiveness of automatic transcriptions lead to an improvement in the cost-effectiveness of automatic translations too.

The work covered for this objective resulted in several scientific publications, including an article in the indexed international journal *Open Learning* (2014) and an article in the conference In-Red 2014 (see the details for these publications below).

Publications

The research work covered in this objective resulted in 3 publications by the author of this master's thesis:

- J. D. Valor Miró, R. N. Spencer, A. Pérez González de Martos, G. Garcés Díaz-Munío, C. Turró, J. Civera, and A. Juan. Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures. *Open Learning*, 29(1):72–85, 2014.

  Indexed international journal: Q2 in Scimago SJR ("Education", 2014).

- J. D. Valor Miró, R. N. Spencer, A. Pérez González de Martos, G. Garcés Díaz-Munío, C. Turró, J. Civera, and A. Juan. Evaluación del proceso de revisión de transcripciones automáticas para vídeos Polimedia. In *Proc. of the I Jornadas de Innovación Educativa y Docencia en Red (IN-RED 2014)*, pages 272–278, València (Spain), 2014.

  National conference on educational innovation and online teaching, organized yearly by Universitat Politècnica de València in Spain.

- J. A. Silvestre-Cerdà, M. Á. Del Agua, G. Garcés, G. Gascó, A. Giménez, A. Martínez, A. Pérez, I. Sánchez, N. Serrano, R. Spencer, J. D. Valor, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan. transLectures. In *Proc. of the VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop (IberSpeech 2012)*, pages 345–351, Madrid (Spain), 2012.

  International conference on speech and language technologies, organized biennially by the International Speech Communication Association Special Interest Group on Iberian Languages (ISCA SIG-IL) and the Thematic Network on Speech Technologies of Spain (RTTH).

2. Developing state-of-the-art neural machine translation systems for their application in the context of Open Educational Resources

In this master's thesis, we have presented work on developing state-of-the-art neural machine translation systems to transition from the previous phrase-based models.

The challenge was to develop neural machine translation systems based on state-of-the-art neural machine translation architectures, making the most of available multilingual and monolingual corpora, and applying appropriately techniques which are essential to improve the quality of NMT systems: corpus filtering, data augmentation through backtranslations, fine-tuning for domain adaptation.

We have presented how we developed a German→English NMT system based on the multi-head self-attentional Transformer DNN architecture, in the framework of the Third Conference on Machine Translation (WMT18). Our NMT system, which required a relatively short time for training, was shown to be competitive in its results, classifying among the first rank in the WMT18 German→English News Translation Shared Task's official results and providing significant improvements over previous PBMT systems (with which we have presented a comparison). In order to improve our system, we used data augmentation through backtranslations (4–6% increase in BLEU), and we also showed the importance of adequate corpus filtering to make the most of large, noisy parallel corpora, employing a simple and quick, yet effective, approach to filtering using character-based language models that resulted in significant improvements in MT quality (8–9% increase in BLEU).

These research results were published in an article in the WMT18 conference (see the publication's details below).

Additionally, we have shown the impact that these new neural machine translation systems had in real Open Educational Resource scenarios within the framework of EU research project X5gon. The experience gathered while developing the new MLLP NMT systems for WMT has been applied to develop new NMT systems for other language combinations within X5gon (including the UPV's poliMèdia repository) and other MLLP projects; in particular, language combinations with English (EN↔ES,FR,DE,IT,SL) and without English (ES↔PT, DE↔FR)). In the case of poliMèdia, improvements in the ES↔EN NMT system have increased its BLEU results on poliMèdia videos by over 40% from 2018 to 2019. Furthermore, a comparison with Google Translate on several datasets showed that our results are on par with the high quality of recent Google Translate results (and also showed that it is possible to go beyond Google Translate's quality through domain adaptation of MT systems, especially for language pairs with comparatively less support from Google Translate).

Publications

The research work covered in this objective resulted in a publication by the author of this master's thesis:

- Javier Iranzo-Sánchez, Pau Baquero-Arnal, Gonçal V. Garcés Díaz-Munío, Adrià Martínez-Villaronga, Jorge Civera, and Alfons Juan. The MLLP-UPV German-English Machine Translation System for WMT18. In *Proc. of the 3rd Conf. on Machine Translation (WMT18)*, pages 422–428, Brussels (Belgium), 2018.

  International conference on machine translation, organized yearly by the Association for Computational Linguistics (ACL).

5.2 Final conclusions

The international education community is every year more aware that accessible multimedia Open Educational Resources for everyone are a necessity for the advancement of education and for the fulfilment of UN Sustainable Development Goal 4 "Quality Education" ("ensure inclusive and equitable quality education and promote lifelong learning opportunities for all"). This awareness is reflected in the UNESCO-sponsored 2012 Paris OER Declaration and 2017 Ljubljana OER Action Plan. In the EU, the 2013 "Opening up Education" agenda recognized the lack of a critical mass of good quality educational content in multiple languages.

Recent events have made this necessity even more immediate in many countries in the world, whether more or less economically advanced, as the COVID-19 epidemic has meant that millions of students and teachers have had to move from face-to-face education to distance education. In this situation, OER for everyone have become and invaluable resource for learning and teaching.

However, in order to achieve their full potential, OER cannot be only available in their original language, and audio/video OER should be accompanied with their transcription. This is where natural language processing technologies (ASR, MT, TTS) play a key role. Quality machine translation is a very useful tool to facilitate cross-lingual access to multimedia OER, not only through gisting, but also through the production of high-quality translations from the original language (with adequate post-editing). At the same time, automatic speech recognition is fundamental in reducing the effort to obtain high-quality transcriptions (to make multimedia OER accessible in their original language for persons with hearing disabilities, and also as the first step for multilingual subtitles via machine translation).

In this work, we have focused on MT for OER, describing the state of the art in MT, and reviewing various techniques for MT evaluation.

In Chapter 3, we have studied MT (and ASR) evaluation techniques for OER in the framework of EU project transLectures (2011–2014). Our study confirms the importance of objective automatic evaluations to quickly measure progress when researching and developing MT systems and trying new techniques. However, we have also described some of the basic limitations of automatic evaluations, which justify the con-

tinued need for well-planned human evaluations as a complement. We have observed the correlations between different automatic MT metrics, as well as between automatic and manual MT evaluation methods. The results shown prove that transLectures MT technology could already produce automatic transcriptions and translations that were accurate enough to be useful for large video lecture repositories, although these results will be clearly surpassed in the following chapter.

From the results presented in this chapter in relation with post-editing with intelligent interaction approaches, there are several aspects that should be taken into account for future research: 1) while in our user evaluations the UPV lecturers expressed their preference for the simplicity and complete control afforded by standard post-editing over the intelligent interaction approaches they tried, the fact that the latter were measurably more cost-effective means that there are probably other use cases where these approaches will be regarded with more interest (such as professional post-editing); 2) different hybrid approaches with different protocols and interfaces can be useful depending on each specific use case; 3) as ASR and MT technologies, as well as the confidence estimation techniques that intelligent interaction is based on, have kept improving with new deep learning techniques, we can imagine the results and the application possibilities for intelligent interaction approaches will also improve in future research.

Regarding Chapter 4, in it we have developed a state-of-the-art NMT system for German→English, replacing previously existing PBMT systems. We have also made a comparison between the results of our best PBMT and NMT systems for the period covered by this master's thesis (until 2018). Thus, we have confirmed, based on relevant real-life examples, how NMT systems have come to outperform the previously reigning PBMT systems.

Additionally, we have shown the impact that these new neural machine translation systems had in real Open Educational Resource scenarios within the framework of EU research project X5gon. A comparison with Google Translate showed that our results are on par with the high quality of recent Google Translate results, and also showed that it is possible to go beyond Google Translate's quality through domain adaptation of MT systems, especially for language pairs with comparatively less support from Google Translate.

UN Sustainable Development Goals: Quality Education for all

These techniques and advancements have proved to be very important, at Universitat Politècnica de València and in the international consortia of EU projects transLectures and X5gon, with regard to moving towards UN Sustainable Development Goal 4 "Quality Education", which aims to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all".

In particular, the work developed an presented in this master's thesis contributes to these three SDG 4 targets:

- 4.3 By 2030, ensure equal access for all women and men to affordable and quality technical, vocational and tertiary education, including university.

- 4.4 By 2030, substantially increase the number of youth and adults who have relevant skills, including technical and vocational skills, for employment, decent jobs and entrepreneurship.
- 4.5 By 2030, eliminate gender disparities in education and ensure equal access to all levels of education and vocational training for the vulnerable, including persons with disabilities, indigenous peoples and children in vulnerable situations.

The work of this master's thesis contributes to increasing accessibility to Open Educational Resources for everyone (target 4.3), including persons with hearing disabilities (target 4.5), speakers of minority or minoritized languages, and persons with fewer resources or without access to formal education systems (target 4.5), for whom access to Open Educational Resources (at the level of primary, secondary and higher education) in a language they understand constitutes an essential source for learning and professional renewal (target 4.4).

While in this master's thesis we have focused on work on machine translation for widely spoken languages (English, Spanish, German), the improvements in NMT described also contribute to making it possible to obtain higher quality machine translation systems for minority or minoritized languages (with fewer available linguistic resources for machine translation system training and testing). Indeed, within the EU projects transLectures and X5gon that constitute the framework for the work of this master's thesis, MLLP group researchers have successfully applied these techniques to make machine translation useful for languages with relatively few speakers, such as Catalan (over 10 million speakers) and Slovenian (2.5 million speakers).

Chapter 6

Bibliography

[1] Docència en Xarxa. https://www.upv.es/contenidos/DOCENRED/.

[2] UN SDG 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all. https://sdgs.un.org/goals/goal4.

[3] UPV Mèdia. http://media.upv.es.

[4] ¿Qué es Polimedia? http://www.upv.es/entidades/ASIC/catalogo/522359normali.html.

[5] 2012 Paris OER Declaration. https://en.unesco.org/oer/paris-declaration, 2012.

[6] Opening up Education agenda: Commission launches 'Opening up Education' to boost innovation and digital skills in schools and universities. https://ec.europa.eu/commission/presscorner/detail/en/IP_13_859, 2013.

[7] 2017 Ljubljana OER Action Plan. https://www.oercongress.org/woerc-actionplan/, 2017.

[8] Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. *CoRR*, abs/1907.05019, 2019.

[9] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proc. of the 3rd International Conference for Learning Representations ICLR 2015*, San Diego, California, USA, 2015.

[12] Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[13] Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August 2019. Association for Computational Linguistics.

[14] Jerome R Bellegarda. Statistical language model adaptation: review and perspectives. *Speech Communication*, 42(1):93 – 108, 2004. Adaptation Methods for Speech Recognition.

[15] Nicola Bertoldi, Barry Haddow, and Jean-Baptiste Fouet. Improved Minimum Error Rate Training in Moses. *The Prague Bulletin of Mathematical Linguistics*, 91(2009):7–16, 2009.

[16] John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence Estimation for Machine Translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland, aug 23–aug 27 2004. COLING.

[17] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[18] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark, September 2017.

[19] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany, August 2016.

[20] Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. Cost Weighting for Neural Machine Translation Domain Adaptation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 40–46, Vancouver, August 2017. Association for Computational Linguistics.

[21] Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[22] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318, 1996.

[23] Colin Cherry, George Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey. Revisiting Character-Based Neural Machine Translation with Capacity and Compression. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4295–4305, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[24] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[25] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42, Jan 2012.

[26] Miguel Ángel Del-Agua, Adrià Giménez, Alberto Sanchis, Jorge Civera, and Alfons Juan. Speaker-Adapted Confidence Measures for ASR using Deep Bidirectional Recurrent Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(7):1194–1202, 2018.

[27] Marzie Fadaee, Aarianna Bisazza, and Christof Monz. Data Augmentation for Low-Resource Neural Machine Translation. *ArXiv e-prints (arXiv:1705.00440)*, May 2017.

[28] George Foster, Cyril Goutte, and Roland Kuhn. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, Cambridge, Massachussets, USA, October 2010.

[29] Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. Lodem: A system for on-demand video lectures. *Speech Communication*, 48(5):516 – 531, 2006.

[30] Guillem Gascó, Martha-Alicia Rocha, Germán Sanchis-Trilles, Jesús Andrés-Ferrer, and Francisco Casacuberta. Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 152–161, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[31] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional Sequence to Sequence Learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[32] M. Gilbert, K. Knight, and S. Young. Spoken language technology [from the guest editors]. *IEEE Signal Processing Magazine*, 25(3):15–16, 2008.

[33] Timothy J. Hazen. Automatic alignment and error correction of human generated transcripts for long speech recordings. In *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006 — ICSLP)*, 2006.

[34] Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. Sockeye: A Toolkit for Neural Machine Translation. *arXiv e-prints (arXiv:1712.05690)*, December 2017.

[35] Javier Iranzo-Sánchez, Gonçal Garcés Díaz-Munío, Jorge Civera, and Alfons Juan. The MLLP-UPV supervised machine translation systems for WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 218–224, Florence, Italy, August 2019. Association for Computational Linguistics.

[36] Javier Iranzo-Sánchez, Pau Baquero-Arnal, Gonçal V. Garcés Díaz-Munío, Adrià Martínez-Villaronga, Jorge Civera, and Alfons Juan. The MLLP-UPV German-English Machine Translation System for WMT18. In *Proc. of the 3rd Conf. on Machine Translation (WMT18)*, pages 422–428, Brussels (Belgium), 2018.

[37] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On Using Very Large Target Vocabulary for Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China, July 2015. Association for Computational Linguistics.

[38] Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[39] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[40] Javier Jorge, Adrià Giménez, Javier Iranzo-Sánchez, Jorge Civera, Albert Sanchis, and Alfons Juan. Real-Time One-Pass Decoder for Speech Recognition Using LSTM Language Models. In *Proc. Interspeech 2019*, pages 3820–3824, 2019.

[41] Marcin Junczys-Dowmunt. Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[42] Marcin Junczys-Dowmunt. Microsoft Translator at WMT 2019: Towards Large-Scale Document-Level Neural Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy, August 2019. Association for Computational Linguistics.

[43] Nal Kalchbrenner and Phil Blunsom. Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[44] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations*, San Diego, California, USA, 2015.

[45] Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009.

[46] Philipp Koehn. Neural machine translation. *CoRR*, abs/1709.07809, 2017.

[47] Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. Findings of the WMT 2019 Shared Task on Parallel Corpus Filtering for Low-Resource Conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy, August 2019. Association for Computational Linguistics.

[48] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, Pennsylvania, USA, 2007. Association for Computational Linguistics.

[49] Samuel Läubli, Rico Sennrich, and Martin Volk. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[50] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2):171–185, 1995.

[51] Pavel Levin, Nishikant Dhanuka, Talaat Khalil, Fedor Kovalev, and Maxim Khalilov. Toward a full-scale neural machine translation in production: the Booking.com use case. *CoRR*, abs/1709.05820, 2017.

[52] Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[53] Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[54] Saturnino Luz, Masood Masoodian, and Bill Rogers. Interactive visualisation techniques for dynamic speech transcription, correction and training. In *Proceedings of the 9th ACM SIGCHI New Zealand Chapter's International Conference on Human-Computer Interaction: Design Centered HCI*, pages 9–16. ACM, 2008.

[55] Adrià Martínez-Villaronga, Miguel del Agua, Jesús Andrés-Ferrer, and Alfons Juan. Language model adaptation for video lectures transcription. In *Proceedings of ICASSP*, pages 8450–8454, 2013.

[56] Robert C. Moore and William Lewis. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden, July 2010.

[57] Cosmin Munteanu, Ron Baecker, and Gerald Penn. Collaborative editing for improved usefulness and usability of transcript-enhanced webcasts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 373–382, New York, NY, USA, 2008. ACM.

[58] Cosmin Munteanu, Gerald Penn, and Xiaodan Zhu. Improving Automatic Speech Recognition for Lectures through Transformation-based Rules Learned from Minimal Data. In *Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Intl. Joint Conf. on Natural Language Processing of the AFNLP (ACL-AFNLP)*, pages 764–772, Suntec, Singapore, 2009.

[59] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[60] S. Repp, A. Gross, and C. Meinel. Browsing within Lecture Videos Based on the Chain Index of Speech Transcription. *IEEE Transactions on Learning Technologies*, 1(3):145–156, 2008.

[61] Isaias Sanchez-Cortina, Nicolás Serrano, Alberto Sanchis, and Alfons Juan. A prototype for interactive speech transcription balancing error and supervision effort. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 325–326. ACM, 2012.

[62] Alberto Sanchis, Alfons Juan, and Enrique Vidal. A word-based naïve bayes classifier for confidence estimation in speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):565–574, 2012.

[63] Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany, August 2016. Association for Computational Linguistics.

[64] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, 2016.

[65] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, 2016.

[66] Nicolás Serrano, Adrià Giménez, Jorge Civera, Alberto Sanchis, and Alfons Juan. Interactive handwriting recognition with limited user effort. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–13, 2013.

[67] J. A. Silvestre-Cerdà, M. Á. Del Agua, G. Garcés, G. Gascó, A. Giménez, A. Martínez, A. Pérez, I. Sánchez, N. Serrano, R. Spencer, J. D. Valor, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan. transLectures. In *Proc. of the VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop (IberSpeech 2012)*, pages 345–351, Madrid (Spain), 2012.

[68] J. A. Silvestre-Cerdà, A. Pérez, M. Jiménez, C. Turró, A. Juan, and J. Civera. A System Architecture to Support Cost-Effective Transcription and Translation of Large Video Lecture Repositories. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3994–3999, 2013.

[69] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proc. of 7th Conf. of Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231, 2006.

[70] Matthew Snover, Shuguang Wang, and Spyros Matsoukas. Translation Error Rate. http://www.cs.umd.edu/~snover/tercom/, 2008. [Online; accessed 6-July-2018].

[71] Swee Kit Alan Soong, Lay Kock Chan, and Christopher Cheers. Impact of video recorded lectures among students. In *Proc. of the the 23rd Annual Ascilite Conference: Who's Learning? Whose Technology?*, pages 789–793, 2006.

[72] Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904, 2002.

[73] Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. SRILM at Sixteen: Update and Outlook. In *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, Hawaii, USA, 2011.

[74] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[75] The transLectures-UPV Team. TLK: The transLectures-UPV Toolkit for Automatic Speech Recognition. https://web.archive.org/web/20170710213839/http://www.mllp.upv.es/tlk/, 2014.

[76] The transLectures project consortium (UPVLC, XEROX, JSI-K4A, RWTH, EML and DDS). transLectures: Final publishable summary. Technical report, Universitat Politècnica de València, December 2014.

[77] Carlos Turró, Aristóteles Cañero, and Jaime Busquets. Video Learning Objects Creation with Polimedia. In *Proc. of the 2010 IEEE International Symposium on Multimedia*, pages 371–376, Singapore, 2010.

[78] Carlos Turró, Miguel Ferrando, Jaime Busquets, and Aristóteles Cañero. Polimedia: a system for successful video e-learning. In *Proc. of the Intl. Conf. EUNIS 2009*, Santiago de Compostela (Spain), 2009.

[79] J. D. Valor Miró, R. N. Spencer, A. Pérez González de Martos, G. Garcés Díaz-Munío, C. Turró, J. Civera, and A. Juan. Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures. *Open Learning*, 29(1):72–85, 2014.

[80] Juan Daniel Valor Miró, Alejandro Pérez González de Martos, Jorge Civera, and Alfons Juan. Integrating a state-of-the-art asr system into the opencast matterhorn platform. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 237–246. Springer, 2012.

[81] J. D. Valor Miró, R. N. Spencer, A. Pérez González de Martos, G. Garcés Díaz-Munío, C. Turró, J. Civera, and A. Juan. Evaluación del proceso de revisión de transcripciones automáticas para vídeos Polimedia. In *Proc. of the I Jornadas de Innovación Educativa y Docencia en Red (IN-RED 2014)*, pages 272–278, València (Spain), 2014.

[82] Juan Daniel Valor Miró, Pau Baquero-Arnal, Jorge Civera, Carlos Turró, and Alfons Juan. Multilingual videos for moocs and oer. *Journal of Educational Technology & Society*, 21(2):1–12, 2018.

[83] Marlies van der Wees, Arianna Bisazza, and Christof Monz. Dynamic Data Selection for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark, 2017.

[84] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. 2017.

[85] Mike Wald. Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. *Interactive Technology and Smart Education*, 3(2):131–141, 2006.

[86] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. Learning Deep Transformer Models for Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy, July 2019. Association for Computational Linguistics.

[87] WMT18 organizers. WMT18 Shared Task: Machine Translation of News. http://www.statmt.org/wmt18/translation-task.html, 2018. [Online; accessed 24-July-2018].

[88] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[89] Joern Wuebker, Hermann Ney, Adrià Martínez-Villaronga, Adrià Giménez, , Alfons Juan, Christophe Servan, Marc Dymetman, and Shachar Mirkin. Comparison of Data Selection Techniques for the Translation of Video Lectures. In *Proc. of the Eleventh Biennial Conf. of the Association for Machine Translation in the Americas (AMTA-2014)*, Vancouver (Canada), October 2014.

[90] Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto, and Eiichiro Sumita. Method of Selecting Training Data to Build a Compact and Efficient Translation Model. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 655–660, Hyderabad, India, 2008.

[91] Dong Yu and Li Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Springer Publishing Company, Incorporated, 2014.

[92] Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. Simultaneous Translation with Flexible Policy via Restricted Imitation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5816–5822, Florence, Italy, July 2019. Association for Computational Linguistics.

Agraïments

A Radha i Teo.

A ma mare i mon pare.

Gràcies a Alfons, Jorge i Adrià el jove.

I gràcies a la resta de companyes i companys dels laboratoris: Adrià-senpai, Albert, Álex, Germán, Guillem, Ihab, Isaías, Javi I., Javi J., Jesús núm. 1, Joan Albert, Juanda, Miwe, Nico, Pau, Rachel, Santi. I als dels altres laboratoris: Antoine, Bea, Dani MA, Dani l'alt, Dani T., Jesús núm. 2, Jordi, Jorge, Lio, Luis, Paco, Pepe, Ricardo, Vicente.

A les meues ties, oncles, cosines, cosins i nebots de Madrid i de Grècia.

A mis tías, tíos, primas, primos, sobrinas y sobrinos de Castro-Urdiales, con su diáspora nacional e internacional.

A la meua família de Xàtiva.

A les companyes i els companys del col·le del Saler: Ad., Ai., Alb., APC, Ar. M., Ar. T., Carm., Carl., Ed., G. G., G. S., Joa., Jor. Bl., Jor. Bo., Jor. N., Llu., Ma., Nat., Nach., P., R., Se., Sí., Su., T. I als de l'institut del sud de Benimaclet: An., El., Er., G. M., He., Joa. A., Jor. G., Jú., Mi. Ra., Mi. Ro., Mo., Ne., Nú.

A les companyes i els companys de la uni: An., Ar., Be., Bo., Caro., Carl., Dan., Dav., E., Ge., Gl., H., Ja., Ju., M., Nat, Naz., Ram., Raü., S., X.

A les companyes i els companys del màster: B., Ca., Cl., Co., Ga., Gui., M., Ó., R.

A les companyes i els companys de japonés del Poli: カルラさん、 ダさん、 ハさん、 フラさん、 べさん、 パブ一さん、 パブ二さん、 ラファさん。 I als del Confuci: アさん、 エちゃん、 オさん、 カルロくん、 フェくん、 ビセさん、 ビくさん、 パトさん、 マくん、 ラモさん、 リサ先生、 リタさん、 ルくん。

日本で会った友達に: Ca., Hi., Ma., Sa., Shi., Yu.

To my Luxembourgish friends: Ai., Alex, Alexa., Ana B., Ana M., And., Car., Cat., Cr., Da., Den., Des., G., Ia., If., Ir., Ka., L. Man., Maria., Marin., Màx., Mira, Mirj., Mo., R., Sa., So., T., Z.

A Kondo Koji, Kimijima Tadashi, Yamashita Kinuyo, Uematsu Nobuo, Yamane Kazunaka, Izumi Mutsuhiko, Takada You, Tonomura Hiroshige, Michael Z. Land & Peter McConnell & Clint Bajakian, Sawa Kazuo, Tamiya Junko, Tanaka Hirokazu, Chikuma Jun, Tim Wright,Shimomura Yoko, Koshiro Yuzo, Frank Klepacki, Muraoka Kazuki, Totaka Kazumi, Kawasaki Yasuhiro, Takeuchi Izuho, Yamamoto & Tomozawa & Iwai & Takehara & Horiyama, Araki Taisuke, David Wise, Alberto José González Pedraza, Kitajima Katsunari, Kuzume Masaya, Jeremy Soule, Glenn Stafford, Nagata Kenta, Mitsuda Yasunori, Sakuraba Motoi, Jake Kaufman, Toby Fox.

Funding acknowledgements

The research leading to these results has received funding from the **European Union**'s Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287755 (transLectures), Competitiveness and Innovation Framework Programme (CIP) under grant agreement no. 621030 (EMMA) and Horizon 2020 research and innovation programme under grant agreement no. 761758 (X5gon); from the **Government of Spain**'s research projects iTrans2 (ref. TIN2009-14511, MICINN/FEDER EU) and MORE (ref. TIN2015-68326-R, MINECO/FEDER EU); and from the **Universitat Politècnica de València**'s PAID-01-17 R&D support programme.

List of Figures

List of Tables