UNIVERSITAT POLITÈCNICA DE VALÈNCIA
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Different Contributions to Cost-Effective Transcription and Translation of Video Lectures

Thesis
presented by Joan Albert Silvestre Cerdà
supervised by Dr. Alfons Juan Císcar and Dr. Jorge Civera Saiz

January 27, 2016

Different Contributions to Cost-Effective Transcription and Translation of Video Lectures

Joan Albert Silvestre Cerdà

Thesis performed under the supervision of doctors
Alfons Juan Císcar and Jorge Civera Saiz
and presented at the Universitat Politècnica de València
in partial fulfilment of the requirements for the degree
Doctor en Informàtica

València,
January 27, 2016

Work supported by the Spanish Goverment under the FPU scolarship (AP2010-4349), and under the iTrans2 (TIN2009-14511), erudito.com (TSI-020110-2009-439) and Active2Trans (TIN2012-31723) projects. Also supported by the EC (FEDER/FSE) under the transLectures (FP7-ICT-2011-7-287755) and EMMA (ICT-PSP/2007-2013-621030) projects.

Per a tu, lluneta.

AGRAÏMENTS

Ara que estic en els compassos finals d'aquest projecte que vaig encetar fa quasi cinc anys, em trobe escrivint aquestes línies, probablement les més difícils de la tesi... i això que les escric en la meva llengua materna. És força complicat transformar en paraules l'entramat de pensaments, records i sentiments que em venen a la ment. Ho intentaré.

Volia començar donant les gràcies als meus directors de tesi, Alfons i Jorge, per tota la seva ajuda, implicació, proximitat, enteniment i empatia; per les facilitats que m'han donat durant tot aquest temps; per tots els coneixements, habilitats i valors que m'han transmès; per la confiança que han dipositat en la meva persona des del primer moment; però sobretot per la gran oportunitat que em van brindar en el seu dia per tal de formar part del seu grup d'investigació. Moltíssimes gràcies a tots dos.

També vull manifestar el més sincer agraïment a la resta de companys de laboratori i de grup (MLLP), ja que sense la seva ajuda i col·laboració, aquesta tesi no hauria estat possible. Parle de Miguel i els seus repositoris de xocolate; Adrià Giménez i els seus infinits assortiments de xiclets; JuanDa i les seves gominoles; Àlex i les nostres nits romàntiques al seu pis; Adrià Martínez i la seva especial sensibilitat cultural i artística; Gonçal i la seva *built-in Wikipedia* (alguns li diuen molt encertadament *Wikiçal*); Santi i la seva increïble psicomotricitat fina *smashbrossiana*; Albert Sanchís i la seva camaraderia; i NiCo i el nostre amor impossible. Un grup de persones genials amb les que m'he trobat molt còmode i amb les que espere poder compartir molts més anys d'èxits i aventures. Moltes gràcies amics!

No vull oblidar-me tampoc dels meus antics companys del MLLP i del PRHLT, dels quals guarde bons records: Ihab, Jesús Andrés, Guillem, Gemán, Rachel, Isaías, Dani Martín, Dani Ortiz, Jesús González, Luis Leiva, Joan Puigcerver, Vicent Bosh, Ricardo... d'entre els quals faig especial menció a Jesús Andrés, el nostre guia espiritual i qui fou co-director del meu PFC i treball d'investigació, al qual li dec el tercer capítol d'aquesta tesi. També mereix figurar entre aquestes línies Matjaz Rihtar, amb el qual estic molt agraït per la magnífica acollida que em va donar i per vetlar que no em faltara de res durant la meva estada de tres mesos a Ljubljana, Slovenija. Hvala Matjaz!

Hi ha altra gent que també, en certa manera, ha influït en la meva persona i en el meu estat anímic, i conseqüentment, en el procés d'elaboració d'aquesta tesi. Parle especialment,

d'una banda, dels meus companys de l'Associació Unió Musical Bocairent, amb els quals tinc el luxe de poder fer música simfònica els caps de setmana, matant així el *cuquet* musical que sovint recorre el meu cos; i d'altra banda, dels meus companys de la delegació d'Àrbitres d'Alcoi, amb els que he compartit grans moments a nombrosos estadis de futbol al llarg de la geografia del nostre País Valencià. A tots vosaltres, gràcies per compartir amb mi eixos moments de desconnexió tant necessaris.

També vull fer una breu però destacada menció als companys i companyes de viatge amb els quals he compartit de forma recurrent els 96 kilòmetres d'asfalt que separen Bocairent de València. En la majoria dels casos, gent magnífica que ha convertit aquests pesats viatges en un intercanvi de coneixements i experiències altament enriquidor.

I ara toca parlar de la gent que més ha "sofrit" aquesta tesi, començant pels meus amics i amigues del poble que tant m'han recolzat i animat durant tot aquest temps. A pesar que a aquestes altures de la vida cadascú fa la seva marxa i no ens veiem molt a sovint, hi compartim uns vincles i afectes especials que ens faran estar units per sempre.

A tota la meva família, tant de sang com política, però en especial als de casa: al meu germà Luís, a la meva cunyada Raquel, i a la meva mare Mari Carmen, moltes gràcies per tot el que haveu fet per mi durant tota aquesta etapa, especialment pel vostre suport, ajuda, interès i estima. No dubteu que sou els principals responsables i instigadors de la consecució d'aquesta fita tant important en la meva vida. Tinc motius de sobra per estar molt, molt orgullós de vosaltres. I allà on estiga el meu pare, Luís, que mentre va viure es va desviure pel benestar de la seva família, n'estic segur que hi estarà ben devanit dels seus fills. El teu esforç no es va quedar debades, pare.

Per acabar, només em queda parlar d'una persona molt especial. *Ella*. I és que després de tot el que he viscut durant aquests anys, estaré eternament agraït pel seu suport incondicional i la seva inesgotable paciència, comprensió, tendresa i estima. *Ella* ha sofrit aquesta tesi tant o més com si fos seva, i, malgrat tot, sempre ha fet tot el que ha estat al seu abast per fer-me feliç. És per això que *Ella* es mereix, més que cap altra persona, tot el meu reconeixement, admiració i estima. No saps ben bé quant afortunat em sent i quant me n'orgulleix tenir-te al meu costat, Lídia. Aquesta tesi és teva.

A tots i totes els i les que he nomenat implícita o explicitament, com també a tots aquells i aquelles que puc haver-me deixat al tinter: moltes gràcies per tot.

Joan Albert Silvestre Cerdà
Bocairent (València)
Novembre 2015

ABSTRACT

In recent years, on-line multimedia repositories have experiencied a strong growth that have made them consolidated as essential knowledge assets, especially in the area of education, where large repositories of video lectures have been built in order to complement or even replace traditional teaching methods. However, most of these video lectures are neither transcribed nor translated due to a lack of cost-effective solutions to do so in a way that gives accurate enough results. Solutions of this kind are clearly necessary in order to make these lectures accessible to speakers of different languages and to people with hearing disabilities. They would also facilitate lecture searchability and analysis functions, such as classification, recommendation or plagiarism detection, as well as the development of advanced educational functionalities like content summarisation to assist student note-taking.

For this reason, the main aim of this thesis is to develop a cost-effective solution capable of transcribing and translating video lectures to a reasonable degree of accuracy. More specifically, we address the integration of state-of-the-art techniques in Automatic Speech Recognition and Machine Translation into large video lecture repositories to generate high-quality multilingual video subtitles without human intervention and at a reduced computational cost. Also, we explore the potential benefits of the exploitation of the information that we know a priori about these repositories, that is, lecture-specific knowledge such as speaker, topic or slides, to create specialised, in-domain transcription and translation systems by means of massive adaptation techniques.

The proposed solutions have been tested in real-life scenarios by carrying out several objective and subjective evaluations, obtaining very positive results. The main outcome derived from this thesis, *The transLectures-UPV Platform*, has been publicly released as an open-source software, and, at the time of writing, it is serving automatic transcriptions and translations for several thousands of video lectures in many Spanish and European universities and institutions.

RESUMEN

Durante estos últimos años, los repositorios multimedia on-line han experimentado un gran crecimiento que les ha hecho establecerse como fuentes fundamentales de conocimiento, especialmente en el área de la educación, donde se han creado grandes repositorios de vídeo charlas educativas para complementar e incluso reemplazar los métodos de enseñanza tradicionales. No obstante, la mayoría de estas charlas no están transcritas ni traducidas debido a la ausencia de soluciones de bajo coste que sean capaces de hacerlo garantizando una calidad mínima aceptable. Soluciones de este tipo son claramente necesarias para hacer que las vídeo charlas sean más accesibles para hablantes de otras lenguas o para personas con discapacidades auditivas. Además, dichas soluciones podrían facilitar la aplicación de funciones de búsqueda y de análisis tales como clasificación, recomendación o detección de plagios, así como el desarrollo de funcionalidades educativas avanzadas, como por ejemplo la generación de resúmenes automáticos de contenidos para ayudar al estudiante a tomar apuntes.

Por este motivo, el principal objetivo de esta tesis es desarrollar una solución de bajo coste capaz de transcribir y traducir vídeo charlas con un nivel de calidad razonable. Más específicamente, abordamos la integración de técnicas estado del arte de Reconocimiento del Habla Automático y Traducción Automática en grandes repositorios de vídeo charlas educativas para la generación de subtítulos multilingües de alta calidad sin requerir intervención humana y con un reducido coste computacional. Además, también exploramos los beneficios potenciales que conllevaría la explotación de la información de la que disponemos a priori sobre estos repositorios, es decir, conocimientos específicos sobre las charlas tales como el locutor, la temática o las transparencias, para crear sistemas de transcripción y traducción especializados mediante técnicas de adaptación masiva.

Las soluciones propuestas en esta tesis han sido testeadas en escenarios reales llevando a cabo numerosas evaluaciones objetivas y subjetivas, obteniendo muy buenos resultados. El principal legado de esta tesis, *The transLectures-UPV Platform*, ha sido liberado públicamente como software de código abierto, y, en el momento de escribir estas líneas, está sirviendo transcripciones y traducciones automáticas para diversos miles de vídeo charlas educativas en diversas universidades e instituciones Españolas y Europeas.

RESUM

Durant aquests darrers anys, els repositoris multimèdia on-line han experimentat un gran creixement que els ha fet consolidar-se com a fonts fonamentals de coneixement, especialment a l'àrea de l'educació, on s'han creat grans repositoris de vídeo xarrades educatives per tal de complementar o inclús reemplaçar els mètodes d'ensenyament tradicionals. No obstant això, la majoria d'aquestes xarrades no estan transcrites ni traduïdes degut a l'absència de solucions de baix cost capaces de fer-ho garantint una qualitat mínima acceptable. Solucions d'aquest tipus són clarament necessàries per a fer que les vídeo xarres siguen més accessibles per a parlants d'altres llengües o per a persones amb discapacitats auditives. A més, aquestes solucions podrien facilitar l'aplicació de funcions de cerca i d'anàlisi tals com classificació, recomanació o detecció de plagis, així com el desenvolupament de funcionalitats educatives avançades, com per exemple la generació de resums automàtics de continguts per ajudar a l'estudiant a prendre anotacions.

Per aquest motiu, el principal objectiu d'aquesta tesi és desenvolupar una solució de baix cost capaç de transcriure i traduir vídeo xarrades amb un nivell de qualitat raonable. Més específicament, abordem la integració de tècniques estat de l'art de Reconeixement de la Parla Automàtic i Traducció Automàtica en grans repositoris de vídeo xarrades educatives per a la generació de subtítols multilingües d'alta qualitat sense requerir intervenció humana i amb un reduït cost computacional. A més, també explorem els beneficis potencials que comportaria l'explotació de la informació de la que disposem a priori sobre aquests repositoris, és a dir, coneixements específics sobre les xarrades tals com el locutor, la temàtica o les transparències, per a crear sistemes de transcripció i traducció especialitzats mitjançant tècniques d'adaptació massiva.

Les solucions proposades en aquesta tesi han estat testejades en escenaris reals duent a terme nombroses avaluacions objectives i subjectives, obtenint molt bons resultats. El principal llegat d'aquesta tesi, *The transLectures-UPV Platform*, ha sigut alliberat públicament com a programari de codi obert, i, en el moment d'escriure aquestes línies, està servint transcripcions i traduccions automàtiques per a diversos milers de vídeo xarrades educatives en nombroses universitats i institucions Espanyoles i Europees.

CONTENTS

CHAPTER 1

INTRODUCTION

Contents

1.1 Motivation

Artificial Intelligence (AI) is a very active research field whose aim is to develop systems, machines and algorithms capable to mimic the human intelligence. Specifically, AI faces the challenge of conferring reasoning, learning, natural language processing and perception capabilities to machines. Hence, AI research is divided in several highly specialised sub-fields. Two of the most sizzling AI sub-fields are the Pattern Recognition (PR) and Machine Learning (ML). Both areas, hardly divisible, study the construction of algorithms that can learn or infer specific knowledge from real-life data to make predictions or to identify patterns. This is the case of applications such as Automatic Speech Recognition (ASR) and Machine Translation (MT), in which the input data (audio signal or text) is properly pre-processed, classified according to the system knowledge, and post-processed to generate the appropriate or expected outputs (text transcription or translated text). In this thesis we will focus on a particular application of ASR and MT technologies in the areas of On-line Educational Technologies and Technology Enhanced Learning (TEL).

In recent years, the growth of the world wide web has offered a great opportunity for academic institutions to enhance the learning process of their students with digital media contents that complement and even replace conventional teaching methods such as face-to-face lectures [12]. Indeed, these digital resources are being incorporated into existing university curricula around the world with enthusiastic response from students [13].

In this sense, on-line multimedia repositories have become established as fundamental knowledge assets, specially in those specialised on serving on-line video lectures. These repositories are being built on the back of on increasingly available and standardised infrastructure [3, 4]. A well-known example of this is VideoLectures.NET [14], a free and open access web portal that has already published more than 20.000 educational videos and conference recordings given by relevant world-wide researchers and professors.

However, the utility of these audiovisual assets could be further extended by adding subtitles that can be exploited to incorporate added-value functionalities such as searchability, accessibility, and discovery of content-related videos, among others. In fact, most of the video lectures available in large university repositories are neither transcribed nor translated, despite the clear need to make their content accessible to speakers of different languages and people with disabilities [15]. Also, the subtitles can be used to develop advanced educational functionalities like content summarisation to assist student note-taking [5].

For this reason, this thesis aims to developing a cost-effective solution that can do so to a reasonable degree of accuracy. More specifically, we propose the integration of state-of-the-art techniques in ASR and MT into large video lecture repositories to generate high-quality multilingual video subtitles without human intervention and at a reduced computational cost. Of course, although it would be the most desirable scenario, we do not expect to produce error-free transcriptions and translations, and, for this reason, we also aim to create efficient and ergonomic tools to allow the review of transcription and translations under a collaborative-editing scenario.

The integration of ASR technologies into multimedia repositories is a well-known problem which has been previously and successfully explored, especially in the case of news broadcasting [6, 8, 9] and TV content in general [1], although most of these systems are mainly designed to provide subtitles in real-time. Other ASR applications can be found in

video indexing [11], or in the subtitling of Parliament sessions [10] where, in some cases, manual transcripts are synchronised with the input audio signal in order to generate the corresponding subtitles [2]. The integration of MT systems, meanwhile, has not been explored in any great depth. There is one exception to this [7], where manual transcripts are assumed to be available before the translation process starts. However, we have not found previous works probing the integration of both ASR and MT technologies into large media repositories, nor the deployment of a collaborative framework through which users can amend errors in transcriptions and translations with little effort.

The generation of multilingual subtitles for video lectures involves the consecutive application of both technologies: on a first step, ASR to generate speech transcripts from the lecturer, and on a second step, MT to translate these transcripts into other languages. Assuming that recognition errors are likely to arise on the first step, and that these errors are propagated to the second step, we need to ensure that our MT technology yields good quality translations regardless the input source language text. In this line, Chapter 3 discusses how length information is modelled in state-of-the-art Statistical MT (SMT) systems, proposing a novel approach in which length variability of word sequences among source and target languages is explicitly taken into account when translating sentences from one language to another. Empirical results show how the proposed length models significantly improve baseline state-of-the-art SMT systems.

It is important to note that ASR systems are the bottleneck of the generation of multilingual subtitles: MT systems can be parallelized in order to reduce the overall computation time, however, they cannot start generating translations until the speech transcript is available. Consequently, ASR systems must be boosted as much as possible without compromising significantly the quality of their outputs. Since the temporal cost of generating an automatic transcription strongly depends on the length of the input audio signal, a simple way to speed up the whole process is to apply a previous step in which the input audio signal is split into homogeneous acoustical regions to detect speech segments, and delivering these isolated speech segments to the ASR system. Furthermore, transcription quality may be improved due the fact that the ASR system does not have to deal with non-speech segments, which are usually but erroneously transcribed by their closest phonetic transcripts. This process of segmenting the input audio signal to detect speech regions is addressed by Audio Segmentation (AS) systems. Since their application is motivated to hasten the overall process of transcribing a video lecture, these systems must be as fast as possible. In Chapter 4, a simple yet powerful and fast AS system is presented. This system participated in the Audio Segmentation competition of the Albayzin 2012 Evaluations, achieving a worthy 2nd place in the final standings.

Despite state-of-the-art ASR and MT systems have been proved to yield accurate speech transcriptions in most cases, their outputs can be greatly improved through the application of massive adaptation techniques. Massive adaptation refers to process of exploiting the wealth of knowledge available in video lecture repositories, that is, lecture-specific knowledge, such as speaker, topic and slides, to create a specialised, in-domain transcription or translation system. A system adapted using this knowledge is therefore likely to produce a far better ASR and MT output than a general-purpose system. These techniques are reviewed and tested in Chapters 5 and 8. In addition, a novel approach to topic adaptation for ASR systems using lecture-related text documents downloaded from the internet is proposed and evaluated in Chapter 7.

As for the integration of ASR and MT technologies into large video lecture repositories, it is needed to design and develop a system architecture capable of blending the existing workflows in remote repositories with transcription and translation processes, as well as to engage users and authors into subtitle review processes. This architecture should also facilitate the incorporation of technological upgrades into ASR and MT systems to allow a progressive refinement of the overall transcription and translation quality of the repository. Indeed, Chapter 5 introduces a novel system architecture that satisfies these requirements. The implementation of this architecture, called *The transLectures Platform* (TLP), was tested under a real-life environment. Furthermore, the proposed system architecture is refined and extended in Chapter 8.

Users that visit multimedia repositories are often overwhelmed by the vast amount of choices that these sites offer. They may not have the time or knowledge to find the most suitable videos for their needs. However, having all video lectures transcribed with our solution described in Chapters 5 and 8, we can generate accurate semantic representations of every lecture that can be used to recommend lectures to users based on their interests. Hence, Chapter 6 describes a novel Recommender System (RS) that exploits lecture transcriptions plus other related text resources to provide better recommendations to users. This RS was developed, deployed and tested in the VideoLectures.NET repository web site.

1.2 Scientific and Technological Goals

All in all, the scientific and technological goals pursued in this work are the following:

- Propose an approach to explicit length modelling for SMT.
- Develop an efficient Audio Segmentation system to speed up ASR systems.
- Study how massive adaptation techniques can lead to better results in transcription and translation of video lecture repositories.
- Propose alternative topic adaptation techniques for ASR.
- Develop a system architecture capable of integrating ASR and MT technologies into video lecture repositories.
- Develop appropriate solutions to enable users to edit transcriptions and translations with ease and relatively small effort under a collaborative scenario.
- Design a Recommender System capable of exploiting speech transcriptions to provide accurate recommendations to users in video lecture on-line repositories.
- Evaluate these contributions in real-life scenarios.
- Make public releases of the software tools developed in this thesis.

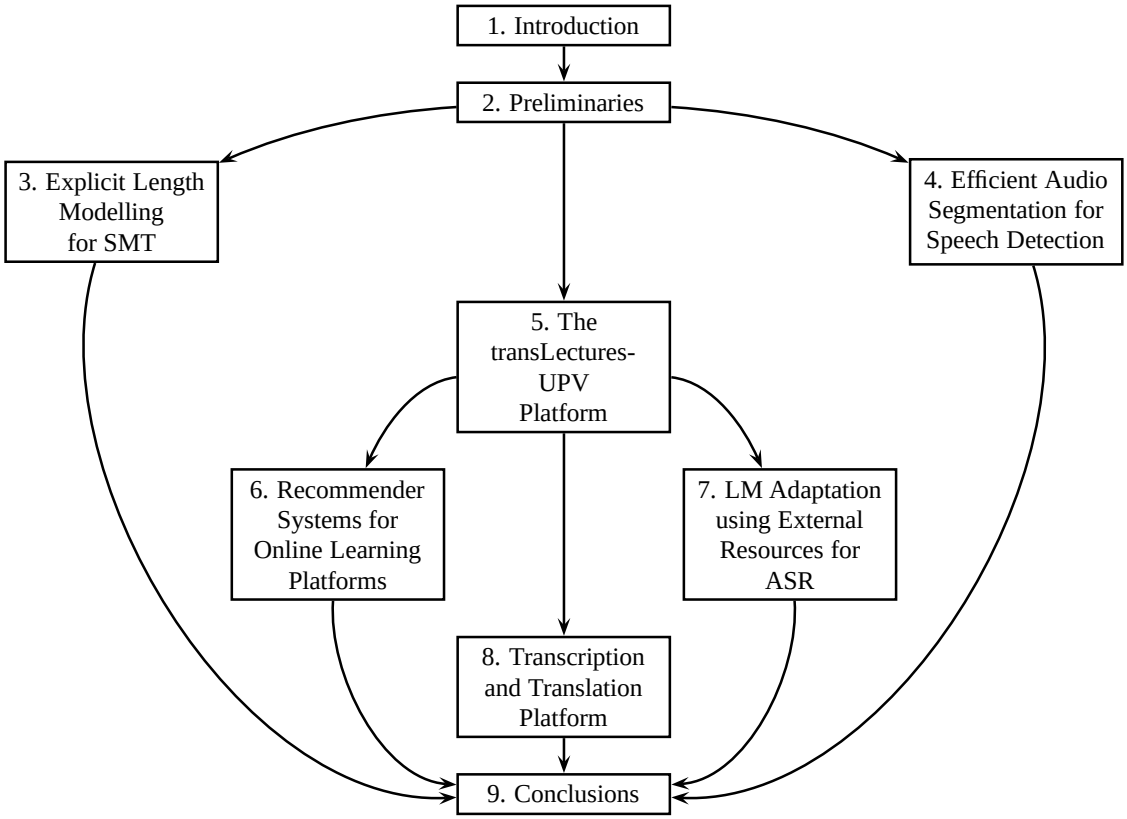

Figure 1.1: Thesis' chapter dependency graph.

1.3 Document Structure

This document is structured in nine sequential chapters that cover the topics and scientific and technological goals proposed in this thesis. First, Chapter 2 gives some preliminary concepts and background knowledge on the research fields covered by this thesis. Then, Chapter 3 addresses the problem of length modelling in SMT and proposes new models that explicitly convey length information about word sequences when translating sentences from one language to another. Next, Chapter 4 describes our proposed AS system to expedite the subsequent ASR processes. Then, Chapter 5 presents a first system architecture to support the integration of ASR and SMT technologies into video lectures, accompanied with objective and subjective evaluations of transcription and translation quality with automatic metrics and real users, respectively. Next, Chapter 6 describes a RS that takes advantage of the availability of automatic transcriptions on a video lecture repository to provide better recommendations to their users. Then, Chapter 7 proposes a new topic adaptation technique for ASR that exploits lecture-related text information extracted from documents downloaded from the web. Next, Chapter 8 can be seen as a continuation of Chapter 5, in which an enhanced version of the system architecture is described, together with an extension of the evaluation results presented in that chapter. Finally, Chapter 9 gives a brief summary of the work described

along the previous eight chapters, highlighting the scientific publications that endorse the scientific impact of the contributions of this thesis, as well as some concluding remarks and future work.

A sequential reading of the nine chapters of this document is encouraged if the reader wants to learn about the whole work, however, specific chapters can be read attending to the dependency graph shown in Figure 1.1.

Bibliography

[1] A. Álvarez, A. del Pozo, and A. Arruti. Apyca: Towards the automatic subtitling of television content in spanish. In *International Multiconference on Computer Science and Information Technology - IMCSIT 2010, Wisla, Poland, 18-20 October 2010, Proceedings*, pages 567–574, 2010.

[2] G. Bordel, S. Nieto, M. Peñagarikano, L. J. Rodríguez, and A. Varona. Automatic subtitling of the basque parliament plenary sessions videos. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, pages 1613–1616, 2011.

[3] Coursera: Take the World's Best Courses, Online, For Free. `http://www.coursera.org`.

[4] edX: Access to free education for everyone. `http://www.edx.org`.

[5] J. R. Glass, T. J. Hazen, D. S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay. Recent progress in the MIT spoken lecture processing project. In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, pages 2553–2556, 2007.

[6] H. Meinedo, M. Viveiros, and J. P. Neto. Evaluation of a live broadcast news subtitling system for portuguese. In *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, pages 508–511, 2008.

[7] M. Melero, A. Oliver, and T. Badia. Automatic multilingual subtitling in the etitle project. *Proc. of Translating and the Computer*, 28, 2006.

[8] J. P. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro. Broadcast news subtitling system in portuguese. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, March 30 - April 4, 2008, Caesars Palace, Las Vegas, Nevada, USA*, pages 1561–1564, 2008.

[9] A. Ortega, J. E. G. Laínez, A. Miguel, and E. Lleida. Real-time live broadcast news subtitling system for spanish. In *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, pages 2095–2098, 2009.

[10] A. Prazák, J. V. Psutka, J. Hoidekr, J. Kanis, L. Müller, and J. Psutka. Automatic online subtitling of the czech parliament meetings. In *Text, Speech and Dialogue, 9th International Conference, TSD 2006, Brno, Czech Republic, September 11-15, 2006, Proceedings*, pages 501–508, 2006.

[11] S. Repp, A. Groß, and C. Meinel. Browsing within lecture videos based on the chain index of speech transcription. *TLT*, 1(3):145–156, 2008.

[12] T. Ross and P. Bell. "No significant difference" only on the surface. *International Journal of Instructional Technology and Distance Learning*, 4(7):3–13, 2007.

[13] S. K. A. Soong, L. K. Chan, C. Cheers, and C. Hu. Impact of video recorded lectures among students. *Who's learning*, pages 789–793, 2006.

[14] VideoLectures.NET: Exchange ideas and share knowledge. `http://www.videolectures.net`.

[15] M. Wald. Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. *Interactive Technology and Smart Education*, 3(2):131–141, 2006.

CHAPTER 2

PRELIMINARIES

Contents

In this chapter we give the essential background information required to understand the rest of the document. Since this is a multidisciplinary thesis, we will review the basics of Automatic Speech Recognition, Machine Translation, Language Modelling and Recommender Systems. Also, we will introduce the evaluation metrics used in this thesis.

2.1 Automatic Speech Recognition

Automatic Speech Recognition (ASR) is an application of the field of Natural Language Processing whose goal is to provide an automatic transcription y of the speech utterances contained in a given input audio signal x. Provided a parametrized audio signal x, we look for the most probable transcription $\hat{y}$ so that

$$\hat{y} = \underset{y \in \mathcal{Y}^*}{\operatorname{argmax}}\, p(x \mid y)\, p(y) \tag{2.1}$$

where $p(x \mid c)$ and $p(y)$ are modelled by acoustic and language models, respectively, being $\mathcal{Y}$ the vocabulary of the system. The most widely adopted audio signal parametrisation technique for automatic speech recognition is the Mel-Frequency cepstral coefficients (MFCC) [6, 23]. State-of-the-art acoustic models use Hidden Markov Models (HMM) to convey triphoneme (phonemes with context) length variability [7, 18, 22], combined with Gaussian Mixture Models (GMM) [7, 22] or Deep Neural Networks (DNN) [5, 8, 20, 22] to model qualitative triphoneme variability (speech features such as timbre, pitch or strength). Language models are introduced in Section 2.4.

2.2 Audio Segmentation

Audio Segmentation (AS) is a task with applications in subtitling, content indexing and analysis that has received notable attention due to the increasing application of ASR systems to multimedia repositories and broadcast news [11, 12, 14, 15]. Formally, this task can be stated as the segmentation of a continuous audio stream into acoustically homogeneous regions. Audio segmentation facilitates posterior speech processing steps such as speaker diarization or speech recognition.

Audio segmentation can be viewed as a simplified case of ASR, in which the system vocabulary is constituted by a reduced set of acoustic classes [3]. For instance, a simple scenario could be the definition of two non-overlapping classes: *Speech* and *No-Speech*.

Provided an audio stream x, the segmentation problem can be stated from a statistical point of view as the search of a sequence of class labels $\hat{c}$ so that

$$\hat{c} = \underset{c \in \mathcal{C}^*}{\operatorname{argmax}}\, p(x \mid c)\, p(c) \tag{2.2}$$

where, as in ASR, $p(x \mid c)$ and $p(c)$ are modelled by acoustic and language models, respectively.

2.3 Machine Translation

Machine Translation (MT) is an application of the field of Natural Language Processing whose goal is to provide an automatic translation of a source sentence x into a target sentence y:

$$x = x_1 \dots x_j \dots x_{|x|} \quad x_j \in \mathcal{X}$$
$$y = y_1 \dots y_i \dots y_{|y|} \quad y_i \in \mathcal{Y}$$

being $\mathcal{X}$ the source language vocabulary, and $\mathcal{Y}$ the target language vocabulary. There exist different approaches to tackle this automatic process:

- Rule-based MT: in this approach a set of linguistic translation rules between two languages are defined by human experts. Therefore, to translate a source language sentence into the target language, only the most suitable existing translation rules are applied to the source sentence.
- Example-based MT: under this approach, the source sentence to be translated is compared against a database of translation examples to find the minimal set of word phrases whose combination generates a complete translation.
- Statistical MT: this approach uses statistical methods to infer probabilistic distributions from real translation examples that model how words or sequences of words from a source language are translated into words or sequences of words in the target language.
- Hybrid MT: combines both the Rule-based and Statistical MT approaches.

In this thesis we will focus on Statistical MT (SMT), which represents the state-of-the-art of the field. In SMT, we formulate the problem of translating a sentence as the search of the most probable target sentence $\hat{y}$ given the source sentence x

$$\hat{y} = \underset{y \in \mathcal{Y}^*}{\operatorname{argmax}} p(y \mid x)\,. \tag{2.3}$$

Applying the Bayes's theorem, we can reformulate Eq. 2.3 as follows:

$$\hat{y} = \underset{y \in \mathcal{Y}^*}{\operatorname{argmax}} p(x \mid y)\, p(y) \tag{2.4}$$

where $p(x \mid y)$ is a translation model, and $p(y)$ a language model. There are different approaches to model the translation model probability, which can be word-based [2] or phrase-based (i.e. that handle the problem of translating sequences of words instead of isolated words) [9].

Nevertheless, state-of-the-art SMT systems are based on log-linear models that combine a set of feature functions to directly model this posterior probability

$$p(y \mid x) = \tfrac{1}{Z(x)} \exp\Big(\textstyle\sum_i \lambda_i\, f_i(x, y)\Big)\,, \tag{2.5}$$

being λ_i, the weight for the i-th feature function $f_i(x, y)$ and $Z(x)$, a normalisation term so that the posterior probability sums up to 1. Feature weights are usually optimised according to minimum error rate training (MERT) on a development set [16]. Conventional feature functions in SMT systems range from those depending on word-based and phrase-based translation models [10], over that directly derived from an n-gram language model [4], to those inspired on word and phrase reordering models, and word and phrase penalties.

2.4 Language Modelling

Language Modelling is another application of the Natural Language Processing field in which grammatical, semantic and syntactic relations between words of a given vocabulary or language are learned by a probabilistic model. A language model (LM) gives a measure of how likely a certain sequence of words is a valid phrase in the target language. LMs are widely used in several applications such as ASR and MT. In these scenarios, the LM is used to compensate the lack of word context information that possess acoustic and translation models, assigning a probability to each possible output proposed by the acoustic or translation models.

Formally, a LM models the probability $p(y)$ of observing a given sentence y:

$$y = y_1 \dots y_i \dots y_I \quad y_i \in \mathcal{Y}$$

being I the length (number of words) of the sentence and $\mathcal{Y}$ the vocabulary of the target language. This probability can be decomposed applying the chain rule:

$$p(y) = p(y_1) \prod_{i=2}^{I} p(y_i \mid y_1^{i-1}) \,. \tag{2.6}$$

where $p(y_i \mid y_1^{i-1})$ denotes the probability of observing y_i after observing the previous subsequence of words y_1^{i-1}, commonly called word history. This means that the probability of observing the word y_i depends on the $i-1$ previous words. If we try to directly model this posterior probability, we will find that the number of parameters of the LM exponentially grows with the length of y, making unfeasible to estimate such model on a large vocabulary $\mathcal{Y}$. Instead, the word history is limited up to the latest n words [1]:

$$p(y) \simeq p(y_1) \prod_{i=2}^{I} p(y_i \mid y_{i-n+1}^{i-1}) \,. \tag{2.7}$$

This approach is called n-gram language modelling. n-gram probabilities are estimated by maximum likelihood as follows [1]:

$$p(y_i \mid y_{i-n+1}^{i-1}) = \frac{N(y_{i-n+1}, \dots, y_{i-1}, y_i)}{N(y_{i-n+1}, \dots, y_{i-1})} \,, \tag{2.8}$$

being $N(\cdot)$ the number of occurrences of the given word sequence observed in the training corpus. However, the available training data is usually insufficient to properly estimate all

model parameters, and, to alleviate this problem, smoothing techniques are applied. These smoothing techniques are based on the idea of discounting some mass probability from all the observed events in the training set and redistributing it among all unobserved events. Further information about smoothing techniques and discount methods can be found in [13].

2.5 Recommender Systems

Recommender Systems (RS) are a multidisciplinary application of Natural Language Processing, Artificial Intelligence and Data Mining that aims to provide suggestions of items that could be of the interest of a user. These suggestions are made usually to intervene in decision-making processes, such as buying a specific object or playing a particular video, specially when the user has to choose among a vast amount of alternatives.

RS are built taking into account mainly these three sources of information:

- Items: the objects that are recommended to the user. Items are usually described by metadata such as their name and their main characteristics or features.
- Users: information about the users is key to deliver good recommendations, since it can be used to infer or predict their willings and preferences. However, in some environments this information might not be accessible, and then recommendations cannot be personalised.
- Transactions: interactions between the RS and the user, that can be used to update user models.

Depending on how this information is exploited, we can distinguish between the following approaches [19]:

- Content-based: recommendations are made on basis of the similar items that the user liked before.
- Collaborative filtering: the system recommends items that were relevant in the past for other users with similar desires.
- Demographic: recommendations are based on the demographic profile of the user, that is, location, genre, age, etc. in the way that the same recommendations are made for users sharing a similar profile.
- Knowledge-based: these systems recommend items based on specific domain knowledge about how user preferences and needs can be satisfied with the recommended objects.
- Community-based: widely used in social networks, these RS recommends to people those items that were explicitly recommended by their friends.
- Hybrid: a combination of any of the previous approaches.

RS predictions typically rely on statistical classifiers [19] such as bayesian networks, support vector machines, or artificial neural networks. Training data is collected from real-life logs and databases. This data is preprocessed to generate parametric representations that can be used to train these classifiers. Finally, the RS evaluates all possible recommendations, returning the best n ranked items.

2.6 Evaluation Metrics

2.6.1 Segmentation Error Rate

The Segmentation Error Rate (SER) is an error metric defined as the fraction of class time that is not correctly attributed to that specific class:

$$\text{SER} = \frac{\sum_n T(n)\left[\max(R(n), H(n)) - C(n)\right]}{\sum_n T(n)R(n)} \tag{2.9}$$

where $T(n)$ is the duration of segment n, $R(n)$ is the number of reference classes that are present in segment n, $H(n)$ is the number of system classes that are present in segment n, and $C(n)$ is the number of reference classes in segment n correctly assigned by the segmentation system.

2.6.2 Word Error Rate

The Word Error Rate (WER) is an error metric that computes the number of edits (insertions, deletions and replacements) that are needed to correct an hypothesis transcription (the output of the ASR system) into the reference:

$$\text{WER} = \frac{I + D + R}{N} \cdot 100$$

being I the number of insertions, D the number of deletions, R the number of replacements, and N the number of words in the reference. WER can be thought of as a percentage approximation of the number of words that need to be corrected in order to achieve the reference transcription.

2.6.3 Bilingual Evaluation Understudy

The Bilingual Evaluation Understudy (BLEU) [17] is a metric that computes different n-gram order precisions between the hypothesis and one or more possible reference translations. BLEU scores can be intuitively understood as the degree of overlap between the automatic translation generated by the MT system and the reference translation provided by a professional linguist. BLEU is computed as follows:

$$\text{BLEU} = \text{BP} \cdot \left(\sum_{n=1}^{N} w_n \log p_n \right)$$

being BP the Brevity Penalty factor used to penalise short translations, N the maximum n-gram order (typically 4), p_n the n-gram precision of order n computed between both hypothesis and reference translations, and w_n the weight assigned to the corresponding n-gram precision (typically $1/n$).

The BLEU is a quality metric ranging from 0 to 100, meaning that the higher value, the better translation quality.

2.6.4 Translation Error Rate

The Translation Error Rate (TER) [21], similarly to the Word Error Rate (WER), is an error metric that computes the number of edits that are needed to correct an hypothesis translation (the output of the MT system) into the reference. The TER is a modified WER in which shifts of word sequences are counted one instead of twice in the case of WER, because it requires two different operations: a deletion and an insertion. Thus, TER is computed as follows:

$$\text{TER} = \frac{I + D + R + S}{N} \cdot 100$$

being I the number of insertions, D the number of deletions, R the number of replacements, S the number of word shifts, and N the number of words in the reference. TER can be thought of as a percentage approximation of the number of words that need to be corrected in order to achieve the reference translation.

2.6.5 Real Time Factor

Since we are tackling the problem of transcribing and translating large media repositories from the viewpoint of cost-effectiveness and efficiency, we are interested in measuring how much time is needed a) to generate a transcription or translation from scratch, either manually or automatically, and b) to manually review automatic subtitles until reaching a perfect transcription or translation.

To this end, we use the Real Time Factor (RTF) metric. RTF is computed as the time spent generating/editing a transcription/translation file divided by the duration of the corresponding video or audio file. So if, for example, a video lasts 20 minutes and the review of its automatic transcription takes, by way of example only, 60 minutes, then the RTF for this video would be 3.

Also, to compute a measure of review efficiency, for automatic transcriptions we will compute the WER reduction per RTF unit, that is, by how many WER points the transcription error is reduced for each RTF unit spent reviewing the automatic transcription. Similarly, for MT, we can compute the TER reduction per RTF unit or the BLEU increase per RTF unit.

Bibliography

[1] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479, Dec. 1992.

[2] P. F. Brown, S. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[3] F. Brugnara, D. Falavigna, and M. Omologo. Automatic segmentation and labeling of speech based on hidden markov models. *Speech Communication*, 12(4):357 – 370, 1993.

[4] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics, 1996*, pages 310–318, 1996.

[5] G. 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 (receiving 2013 IEEE SPS Best Paper Award)*, 20(1):30–42, January 2012.

[6] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, Aug 1980.

[7] M. Gales and S. Young. The application of hidden markov models in speech recognition. *Found. Trends Signal Process.*, 1(3):195–304, Jan. 2007.

[8] G. Hinton, L. Deng, D. Yu, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. G. Dahl, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012.

[9] P. Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010.

[10] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, 2003*, pages 48–54, 2003.

[11] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22(5):533 – 544, 2001.

[12] L. Lu, H. Jiang, and H. Zhang. A robust audio classification and segmentation method. In *Proc. of ACM International Conference on Multimedia*, pages 203–211, 2001.

[13] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[14] H. Meinedo and J. Neto. Audio segmentation, classification and clustering in a broadcast news task. In *Proc. of ICASSP*, volume 2, pages 5–8, 2003.

[15] T. Nwe and H. Li. Broadcast news segmentation by audio type analysis. In *Proc. of ICASSP*, volume 2, pages 1065–1068, 2005.

[16] F. J. Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, 2003*, pages 160–167, 2003.

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002*, pages 311–318, 2002.

[18] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice-Hall, 1993.

[19] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor. *Recommender Systems Handbook*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.

[20] F. Seide, G. Li, and D. Yu. Conversational speech transcription using context-dependent deep neural networks. In *Interspeech*, pages 437–440, 2011.

[21] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas, 2006*, pages 223–231, 2006.

[22] D. Yu and L. Deng. *Automatic Speech Recognition - A Deep Learning Approach*. Springer, October 2014.

[23] F. Zheng, G. Zhang, and Z. Song. Comparison of different implementations of mfcc. *Journal of Computer Science and Technology*, 16(6):582–589, 2001.

CHAPTER 3

EXPLICIT LENGTH MODELLING FOR STATISTICAL MACHINE TRANSLATION

Contents

3.1 Introduction

As stated in Chapter 1, the main objective of this thesis is to provide a cost-effective solution to generate multilingual subtitles for media repositories. In this sense, MT is the technology that enables multilingualism. Under this scenario, MT systems take as input automatic video lecture transcripts, translating them into other languages. However, these transcripts are likely to contain errors due to the inherent complexity of the ASR task. These errors are propagated to MT systems, which also add another possible source of errors that can compromise the quality of translations. Therefore, we should ensure that our MT technology yields good quality translations regardless of the input transcripts.

Given this premise, at a first stage, our work was focused on finding possible ways to improve state-of-the-art Statistical MT. We realised that state-of-the-art SMT systems do not explicitly take into account the variability in length of word sequences between languages. For this reason, in this Chapter we address the problem of explicit length modelling in SMT.

Explicit length modelling is a well-known problem in pattern recognition which is often disregarded. However, it has provided positive results in applications such as author recognition [23], handwritten text and speech recognition [27], and text classification [10], whenever it is taken into consideration.

Length modelling may be considered under two points of view. On the one hand, the so-called implicit modelling in which the information about the length of the sequence is indirectly captured by the model structure. This is often the case of handwritten text and speech recognition [11], language modelling [5] and machine translation [18], which often include additional states to convey length information.

On the other hand, we may perform an explicit modelling by incorporating a probability distribution in the model to represent length variability in our data sample [22]. Explicit modelling can be found in language modelling [12, 19], and bilingual sentence alignment and segmentation [3, 9], among others.

Explicit length modelling in SMT has received little attention since Brown's seminal paper [4] until recently. Nowadays state-of-the-art SMT systems are grounded on the paradigm of phrase-based translation [18], in which sentences are translated as segments of consecutive words. Thereby, most recent work related to explicit length modelling has been performed at the phrase level with a notable exception [25]. Explicit phrase length modelling was initially presented in [24] where the difference ratio between source and target phrase length is employed to phrase extraction and scoring with promising results. Zhao and Vogel [26] discussed the estimation of a phrase length model from a word fertility model [4], using this model as an additional score in their SMT system. In [8], a word-to-phrase model is proposed which includes a word-to-phrase length model. Finally, [1] describes the derivation and estimation of a phrase-to-phrase model including a model for the source and target phrase lengths.

However, none of the previous works report results on how explicit phrase length modelling contributes to the performance of a state-of-the-art phrase-based SMT system. Furthermore, phrase-length models proposed so far depend on their underlying model or phrase extraction algorithm, which differ from those employed in state-of-the-art SMT systems. The current work is inspired on the explicit phrase length model proposed in [1], but applied to a state-of-the-art phrase-based SMT system [17] and assessed on diverse language pairs in

order to systematically evaluate the contribution of explicit phrase length modelling in SMT.

The organisation of this chapter is as follows. The next section describes how state-of-the-art SMT systems implicitly model length information. Section 3.3 explains the proposed conditional phrase length models. Experimental results are reported in Section 3.4. Finally, conclusions and future work are discussed in Section 3.5.

3.2 Implicit Length Modelling

As stated in Section 2.3, state-of-the-art SMT systems are implemented by log-linear models that combine several feature functions based on phrase-based models.

Phrase tables do not model conditional phrase length correlation between corresponding phrase translations, that is, the probability of translating a source phrase made up of l words by a target phrase of m words, even though conditional phrase length models seamlessly emerge in the generative process of a bilingual phrase-based segmentation [1].

Nevertheless, phrase-based models do implicitly model sentence length information through some of these features, such as word and phrase penalty, that controls the number of words and phrases in the resulting translation. As discussed in more detail below, the word penalty compensates for the bias towards short sentences [4] or prevents the generation of spurious words [17], while the phrase penalty avoids the bias towards long phrases.

In general, n-gram language models incorporate the special end-of-sentence symbol that implicitly models sentence length information, even though it is not able to incorporate long-term constraints. This limitation produces that ill-formed sentences receive an exponentially growing probability mass depending on their length [4]. Hence, the probability of well-formed sentences exponentially decays with their length. In order to alleviate this bias towards short sentences, the word penalty feature introduces a constant bonus for each new word added to the translation. However, in phrase-based SMT systems, the word penalty avoids the generation of spurious words [17]. In any case, the word penalty feature aims at implicitly modelling sentence length information, not phrase length information, as the models proposed in this work do.

On the other hand, phrase tables suffer from a bias towards long phrases due to a similar modelling deficiency. Indeed, the phrase penalty adds a constant bonus for each additional phrase incorporated into the translation. In fact, as shown in Section 3.3, the phrase penalty is complementary to the proposed conditional phrase length models.

In this chapter, we address the problem of explicit conditional length modelling at the phrase level. In addition to the conventional features mentioned above, additional features derived from conditional phrase length models [1] are introduced. These additional features are presented in the next section.

3.3 Explicit length modelling

In the phrase-based approach to SMT, the translation model considers that the source sentence x is generated by segments of consecutive words defined over the target sentence y. As in [1],

in order to define these segments we introduce two hidden segmentation variables

$$p(x \mid y) = \sum_T \sum_{l_1^T} \sum_{m_1^T} p(x, l_1^T, m_1^T \mid y)\,, \tag{3.1}$$

being T the number of phrases into which both sentences are to be segmented, and being l_1^T and m_1^T the source and target segmentation variables, respectively. Thus, we can factor Eq. (3.1) as follows

$$p(x, l_1^T, m_1^T \mid y) = p(m_1^T \mid y)\, p(l_1^T \mid m_1^T, y)\, p(x \mid l_1^T, m_1^T, y)\,, \tag{3.2}$$

where $p(m_1^T \mid y)$ and $p(l_1^T \mid m_1^T, y)$ are phrase length models, whilst $p(x \mid l_1^T, m_1^T, y)$ constitutes the phrase-based translation model. We can independently factorise terms in Eq. (3.2) from left to right,

$$p(m_1^T \mid y) = \prod_t\ p(m_t \mid m_1^{t-1}, y)\,, \tag{3.3}$$

$$p(l_1^T \mid m_1^T, y) = \prod_t\ p(l_t \mid l_1^{t-1}, m_1^T, y)\,, \tag{3.4}$$

$$p(x \mid l_1^T, m_1^T, y) = \prod_t\ p(x(t) \mid x(1), \ldots, x(t-1), l_1^T, m_1^T, y)\,, \tag{3.5}$$

where t ranges over the possible segmentation positions of the target sentence, l_t and m_t are the length of the t-th source and target phrase, respectively, and $x(t)$ is the t-th source phrase.

In state-of-the-art systems, the model in Eq. (3.3) is approximated by the phrase penalty, which is intended to control the number of phrases involved in the construction of a translation, as previously discussed. Eq. (3.5) is simplified by conditioning only on the t-th target phrase to obtain the conventional phrase table, which is used as another feature,

$$p(x(t) \mid x(1), \ldots, x(t-1), l_1^T, m_1^T, y) := p(x(t) \mid y(t))\,, \tag{3.6}$$

with parameter set, $\theta = \{p(u \mid v)\}$, for each source, u, and target, v, phrase. Finally, Eq. (3.4) is used to derive conditional phrase length models that become new feature functions of our log-linear model, and the corresponding phrase-based SMT system.

Next sections present two conditional phrase length models, namely, *standard* and *specific*, as a result of different assumptions on Eq. (3.4). In addition, two alternative parametrisations will be considered for each of these models, referred to as *parametric* and *non-parametric*.

3.3.1 Standard length models

The standard length model is derived from Eq. (3.4) by taking the assumption that the source length l_t only depends on the corresponding target phrase length m_t as follows

$$p(l_t \mid l_1^{t-1}, m_1^T, y) \approx p(l_t \mid m_t)\,. \tag{3.7}$$

The parametric model further assumes that the rightmost probability in Eq. (3.7) follows a Poisson distribution

$$p_{\gamma_{m_t}}(l_t \mid m_t) \propto \gamma_{m_t}^{l_t} \exp(-\gamma_{m_t}) \tag{3.8}$$

where the mass probability function is renormalised to sum 1, if a maximum phrase length is specified. Therefore, the parameter set is $\gamma = \{\gamma_m\}$ for each target phrase length m.

On the contrary, in the non-parametric model, each $p(l_t \mid m_t)$ term in Eq. (3.7) plays the role of a parameter, and, consequently, the parameter set is given by $\gamma = \{p(l \mid m)\}$ for each source, l, and target, m, lengths. This model is more sparse than the parametric model and it is smoothed to alleviate this problem as follows

$$\tilde{p}(l \mid m) := (1 - \varepsilon) \cdot p(l \mid m) + \varepsilon \cdot \frac{1}{M}\,, \tag{3.9}$$

where M stands for the maximum phrase length.

For a given maximum phrase length M, say 7, the parametric standard model requires M parameters, i.e. $\{\gamma_1, \gamma_2, \ldots, \gamma_M\}$, while the non-parametric model needs M^2 parameters, i. e. $\{p(1 \mid 1), p(2 \mid 1), \ldots, p(M \mid 1), p(1 \mid 2), \ldots, p(M \mid M)\}$.

3.3.2 Specific length models

In the specific model, we take a more *specific* assumption for Eq. (3.4) than that of Eq. (3.7) by considering the dependency on the actual phrase $y(t)$, instead of its length,

$$p(l_t \mid l_1^{t-1}, m_1^T, y) \approx p(l_t \mid y(t))\,, \tag{3.10}$$

being $p(l_t \mid y(t))$, a source phrase-length model conditioned on the t-th target phrase. This latter probability $p(l_t \mid y(t))$ in Eq. (3.10) can be regarded as a parameter itself, yielding the non-parametric model. In this case, the parameter set is defined by $\gamma = \{p(l \mid v)\}$, for any target phrase v. In practice, v is any target phrase observed in the training set.

Similarly to the standard length model, the parametric model will assume that the probability in Eq. (3.10) follows a Poisson distribution

$$p_{\gamma_{y(t)}}(l_t \mid y(t)) \propto \gamma_{y(t)}^{l_t} \exp(-\gamma_{y(t)})\,, \tag{3.11}$$

where the probability mass function is renormalised so that it sums up to 1 if a maximum phrase length is specified. Hence, the parameter set is $\gamma = \{\gamma_v\}$ for each target phrase v. It is worth noting the difference between Eq. (3.8) and Eq. (3.11). In the former, a Poisson distribution is considered for each *target phrase length*, while in the latter a Poisson distribution is assumed for each *target phrase*.

Specific length models, both parametric and non-parametric, are considerably more sparse than those of the standard model. In order to alleviate overfitting problems, the specific parameters are smoothed with the standard parameters as follows

$$\tilde{p}(l \mid v) := (1 - \varepsilon) \cdot p(l \mid v) + \varepsilon \cdot \tilde{p}(l \mid |v|)\,, \tag{3.12}$$

denoting by $|\cdot|$ the length of the corresponding phrase. The interpolation parameter ε is adjusted on a validation set in order to maximise BLEU.

Given a maximum phrase length M, and the set of all unique target phrases that have been extracted from the training data $\mathcal{V} = \{v_1, \ldots, v_n\}$, the parametric specific model requires one Poisson parameter for each phrase, i.e., $\{\gamma_{v_1}, \gamma_{v_2}, \ldots, \gamma_{v_n}\}$. On the other hand, the non-parametric specific model requires M parameters for each target phrase, $\{p(1 \mid v_1), \ldots, p(M \mid v_1), p(1 \mid v_2), \ldots, p(M \mid v_2), \ldots, p(1 \mid v_n), \ldots, p(M \mid v_n)\}$.

As usual in statistical modelling, there is a trade-off between the complexity of the model, basically the number of parameters to be learnt, and the number of data samples available. The more parameters to be learnt, the more data samples are needed to properly train the corresponding model. If we compare the specific model with the standard model, the former possesses a significantly greater number of parameters with respect to the latter, and hence, smoothing becomes critical. For instance, most of the target phrases occur only once, and then, the ratio of samples to parameter for the non-parametric specific model is less than 1, which requires a strong smoothing. However, the standard model possesses fewer parameters, and hence, it does not suffer from severe overfitting problems, i.e., the ratio will always be much larger than 1.

Another problem that arises in the context of data sparsity is the excessive generalisation of parametric models. Our Poisson model makes a stronger assumption regarding the probability length distribution than that of the non-parametric model. However, it could be the case that a target phrase occurs only once with its corresponding source phrase translation. If we assume that source and target phrase share the same length, say $m = l$, then the maximum likelihood estimation of a non-parametric model gives $(1 - \varepsilon)$ probability mass to the single observed hypothesis. In contrast, the parametric model, which follows a Poisson distribution, would smoothly decrease this probability according to $(1 - \varepsilon)\, l^m \, \exp(-l)$ for source lengths m different from l. Depending on the language pairs involved, either the parametric or the non-parametric model would be a better hypothesis. These trade-offs are experimentally analysed in Section 3.4.

3.3.3 Estimation of phrase-length models

The parameters of the models introduced in the previous section could be estimated by maximum likelihood criterion using the EM algorithm [7]. As shown in [1], the phrase-based translation model is estimated as

$$p(u \mid v) = \frac{N(u, v)}{\sum_{u'} N(u', v)}\,, \tag{3.13}$$

being $N(u, v)$, the expected counts for the bilingual phrase (u, v). The estimation of $p(l \mid m)$ is computed as

$$p(l \mid m) = \frac{N(l, m)}{\sum_{l'} N(l', m)}\,, \tag{3.14}$$

where

$$N(l, m) = \sum_{u,v} \delta(l, |u|)\, \delta(m, |v|)\, N(u, v)\,, \tag{3.15}$$

being δ, the Kronecker delta. Conversely, for the Poisson model, the estimation of the parameter γ_m is similar to that of Eq. (3.14), and is given by

$$\gamma_m = \frac{\sum_l l \cdot N(l, m)}{\sum_l N(l, m)}\,. \tag{3.16}$$

The parameters for the specific models, $p(l \,|\, v)$, are estimated analogously.

Although expected counts $N(u,v)$ were exactly computed in [1], this was done at the expense of limitating the expressiveness of the model to only consider monotonic bilingual segmentations, otherwise the computational cost of the expected counts would require exponential time [13].

However, the parameter estimation of conventional log-linear phrase-based system approximate expected counts $N(u,v)$ with approximated counts $N^*(u,v)$, derived from a heuristic phrase-extraction algorithm [16]. Similarly, our first approach is also to approximate $N(l,m)$ in Eq. (3.14) as follows

$$N^*(l,m) = \sum_{u,v} \delta(l,|u|)\,\delta(m,|v|)\,N^*(u,v)\,. \tag{3.17}$$

This approach is referred to as *phrase-extract* estimation, or simply *extract* estimation. The estimation of the proposed models with the phrase-extract estimation can be implemented adding a constant time to the the phrase-extract algorithm for each phrase extracted. But for simplicity reasons, it has been implemented as an additional pass over the extracted phrases.

A second approach to the estimation of phrase-length parameters is based on the idea of a Viterbi approximation to Eq. (3.1). This approach considers only the source and target segmentation that maximises Eq. (3.1)

$$\hat{l_1^T}, \hat{m_1^T} = \underset{l_1^T, m_1^T}{\operatorname{argmax}} \{Pr(x, l_1^T, m_1^T \mid y)\}\,. \tag{3.18}$$

So, the hidden segmentation variables are uncovered and the counts in Eq. (3.13) are not expected fractional counts, but integer counts approximated by the Viterbi segmentation.

The search denoted by Eq. (3.18) is performed using a conventional log-linear phrase-based system which is based on a A^* search algorithm. It must be noted that the source and target sentences are available during the training phase, so this search becomes a guided search in which the target sentence is known.

In terms of computational complexity, the Viterbi-based estimation introduces a large additional computational cost to the standard training phase. First, the Viterbi segmentation of each training sample needs to be computed, which is an NP-hard problem approximated by a A^* search algorithm. Then, counts are collected from Viterbi segmentations in order to estimate phrase length parameters.

Regarding the estimation method, it is not clear whether Viterbi or phrase-extract counts better approximate actual expected counts, or even more important, which counts yield a better estimation. On the one hand, Viterbi counts are more sparse than extract counts since they are obtained only from a single segmentation, that is, the most probable segmentation. On the other hand, phrase-extract counts are extracted from several "heuristic segmentations". For example, let $y = (y_1, y_2)$ a target sentence, and $x = (x_1, x_2)$ its source counterpart. Despite its simplicity, this example allows us to illustrate the sparseness of the different estimation methods. The heuristic extraction is based on word alignments between source and target words. For this example, we further assume that x_1 is aligned with y_1, and x_2 with y_2. Provided this example, the Viterbi approximation would probably consider the full sentence as a phrase, (x_1x_2, y_1y_2), counting it once. In contrast, the phrase-extraction heuristic, would produce 3 phrases: $(x_1x_2, y_1y_2), (x_1, y_1)$, and (x_2, y_2), and each of them would be counted once.

Table 3.1: Basic statistics for Europarl-v3.

Training sets	Monolingual			Bilingual			
Language pairs	En	Es	De	En	Es	En	De
Bilingual sentences	1.4M			965K		995K	
Vocabulary size	115.7K	167.6K	327.2K	81.8K	113.0K	74.6K	226.9K
Running words	38.3M	40.3M	36.7M	20.3M	20.9M	21.5M	20.4M

	Development						Test		
	dev2006			devtest2006			test2007		
Language	En	Es	De	En	Es	De	En	Es	De
Sentences	2K			2K			2K		
Vocab. size	6.1K	7.7K	8.8K	6.1K	7.8K	8.7K	6.0K	7.8K	8.8K
Run. words	58.8K	60.5K	55.1K	58.1K	60.2K	54.2K	59.2K	61.3K	55.6K
Perplexity	74	75	119	73	76	118	71	76	121

3.4 Experimental results

In this section, a systematic evaluation is performed to elucidate the benefits of explicit phrase length modelling in phrase-based SMT. To this purpose, three language pairs were involved in the experiments: English-Spanish (En-Es), Spanish-English (Es-En), English-German (En-De), German-English (De-En) and Chinese-English (Zh-En). Experiments on Spanish and German were carried out using the Europarl-v3 parallel corpora [15], which is a reference task in the SMT field, while the Chinese-English experiments were performed using the BTEC parallel corpora provided in the evaluation campaign for the IWSLT09 [21]. Basic statistics for both corpora, Europarl and BTEC, are shown in Tables 3.1 and 3.2, respectively.

The experimental setup for the Europarl-v3 corpora provides three separate sets for the purpose of evaluation campaigns: training, development, and test. The training set consists of two datasets. The first of them is a monolingual dataset, that is devoted to train language models, while the second dataset is a parallel corpus to train translation models. Also, two development sets, known as *dev2006* and *devtest2006*, are provided. On the one hand, the dataset *dev2006* is used to perform Minimum Error Rate Training (MERT) of the weights involved in the log-linear SMT model [20]. On the other hand, the development set *devtest2006* is dedicated to adjust the interpolation smoothing parameter ε. Finally, final performance results are reported on the *test2007* set.

Similarly, the BTEC corpora was also divided in three sets. The training set consists in an unique parallel dataset which is used to train both language and translation models. The development set *devset6* was divided into two datasets. The first dataset, referred to as *dev-mert*, is devoted to perform Minimum Error Rate Training, and the second dataset, *dev-smooth*, is used to optimise the interpolation smoothing parameter, ε. Finally, *devset7* is the test set on which final results are reported.

The performance of phrase length models was assessed on the freely available Moses toolkit [17]. Basically, we compare the performance of the Moses baseline system (including word and phrase penalties) to that of an augmented version of the Moses system incorporating the phrase length models as additional features. More precisely, the phrase-length augmented system includes two additional features, a source-conditioned and a target-conditioned phrase

Table 3.2: Basic statistics for BTEC (IWSLT09).

	Training		Development (devset6)				Test (devset7)	
			dev-mert		dev-smooth			
Language pairs	Zh	En	Zh	En	Zh	En	Zh	En
Bilingual sentences	20K		389		100		511	
Number of references	1	1	1	6	1	6	1	10
Vocabulary size	8.4K	7.1K	726	724	307	321	888	872
Running words	171.5K	188.9K	2.5K	3K	682	871	3.3K	4.2K
Perplexity	-	-	47	26	60	31	51	33

length models.

A selection of the most relevant and representative experiments are shown in this section. In order to gauge the translation quality of the different systems, the well-known BLEU and TER metrics were used. In all cases, we reported the performance of the system on case-insensitive translations.

First, BLEU scores as a function of maximum phrase length are plotted for each language direction to provide an initial performance analysis of phrase length models (standard vs. specific), parametrisations (Poisson vs. contingency table) and estimation methods (extract vs. Viterbi). In these plots confidence intervals are not reflected for the sake of clarity. Afterwards, we report BLEU and TER for maximum phrase length (maxPL=7) including confidence intervals and pairwise statistical significance tests.

To be more precise, two bootstrapping methods, referred to as standard [14] and pairwise [2], were applied to all experiments in order to verify the statistically significance of our results. The standard bootstrapping method computes absolute confidence intervals for BLEU or TER for each system [14], while the second performs pairwise system comparison [2]. Although the standard bootstrapping method yields trustful confidence intervals, it ignores the current bootstrapping sample complexity and thereby, typically produces large confidence intervals. In other words, it does not consider that there are sentences which are more difficult to translate than others, such as long sentences. In contrast, the pairwise bootstrapping method provides smaller variances, since it takes into account the sample variance by computing the difference with respect to a baseline system. For this reason, we also report the so-called *probability of improvement* (PI) [2] that aims at minimising the variety of bootstrapping sample complexity by simply counting the number of times a system is better than other without taking into account the absolute improvement. PI figures for BLEU and TER evaluations are reported when comparing the performance of our proposed models to that of the conventional Moses baseline system.

Figures 3.1, 3.2 and 3.3 show the evolution of the BLEU score (y-axis) as a function of the maximum phrase length (x-axis) for experiments involving Spanish, German and Chinese, respectively. In the case of Spanish and German experiments, the left-most column of plots presents BLEU trends with English as the source language, while the right-most column does the same with English as the target language. Reading Figures 3.1 and 3.2 by rows from top to bottom, first the standard (std) or specific (spc) model is set, depending which the best performing system is, leaving the other two experimental parameter (estimation method and parametrisation) free. The second row sets the estimation method (Viterbi or extract) and the

rest of experimental parameters are left free. Finally, the third row leaves the parametrisation constant, Poisson (param) or contingency table (non-param), and explores the other two experimental parameters.

In Figure 3.1, the experiments regarding Spanish are presented. In both directions, the best performing system involved the standard model with Poisson parametrisation and Viterbi parameter estimation. As shown later, the PI for BLEU proved that the improvement over the baseline is statistically significant, and not only for the best performing system. However, given the standard model, there seem not to be a clearly better estimation method or parametrisation.

Figure 3.2 shows BLEU trends involving German. Generally speaking, the best results are obtained again with the standard model, but the specific model using the extract estimation model obtains similar performance in the English-German pair. In all cases, the extract estimation method seems to perform the best on average. These results can be explained in the light of the trade-off between the number of parameters and data samples available, mentioned in Section 3.3.2. As shown in Table 3.1, the German language possesses a higher perplexity compared to Spanish, so this fact reduces the parameter to sample ratio. To compensate this effect more samples are required to train the same model (standard or specific). For this reason, the extract estimation method is preferred over the Viterbi method.

Interestingly enough, the non-parametric parametrisation based on a contingency table outperforms the Poisson parametrisation in all cases. This phenomenon is related to the stronger assumption on the probability length distribution of the Poisson compared to that of the non-parametric approach explained in Section 3.3.2.

Figure 3.3 presents BLEU score trends in the Chinese-English BTEC task. The leftmost plot sets the specific model (best performing model for maximum phrase length equal to 7) to analyse the influence of the estimation method and the parametrisation, while the rightmost plot sets the non-parametric approach and compares the performance of the standard and specific models, and estimation methods. As shown, the non-parametric approach supersedes the Poisson parametrisation given the specific model. This phenomenon is the same than that observed in the experimental results with German. However, the specific model outperforms the standard model in all cases, although an estimation method is not clearly preferred over the other.

Tables 3.3, 3.4 and 3.5 show comparative performance results achieved by the baseline system and the different proposed phrase length models for maximum phrase length equal to 7. Furthermore, confidence intervals at 95% computed according to the bootstrapping method proposed in [14] are reported just below the corresponding measure, while PI in percentage for BLEU and TER were calculated according to [2].

For all the experiments, we analyse the behaviour of the BP, and observed that it does not varies significantly. For instance, for the German to English task it is 0.995 ± 0.006 for all models including the baseline. The only exception in which the BP could have some effect in BLEU scores is the English-Spanish pair. In this pair, it is observed the greatest variability in BP 0.987 ± 0.009, but the BP of the baseline is similar to that of some of our proposed models. For this reason, we do not report a systematic evaluation in terms of BP values. Note that this result is in accordance with the discussion in Section 3.3, in which we concluded that the sentence length prediction is not expected to improve as a direct consequence of applying phrase length models.

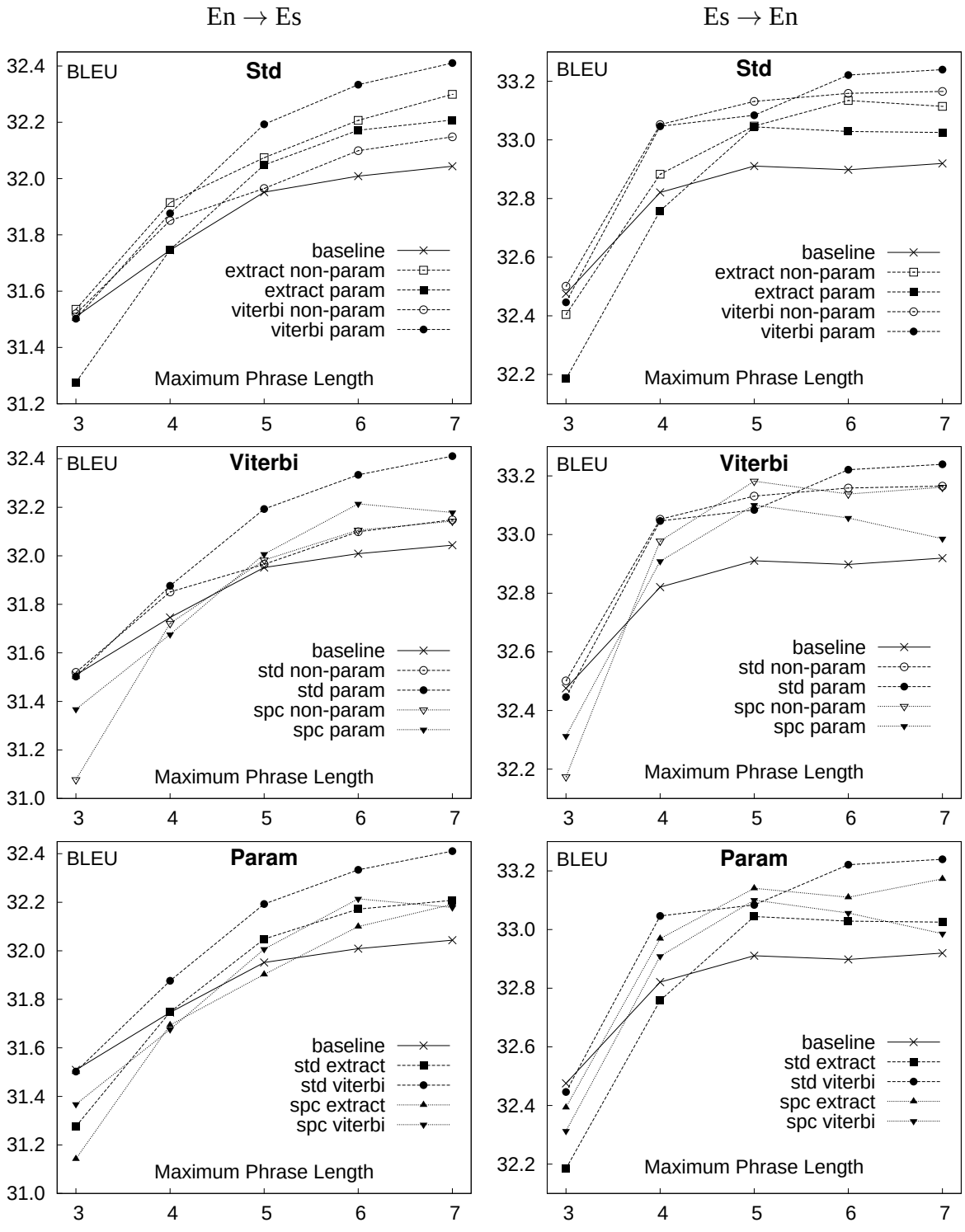

Figure 3.1: BLEU scores as a function of the maximum phrase length in English-Spanish and Spanish-English.

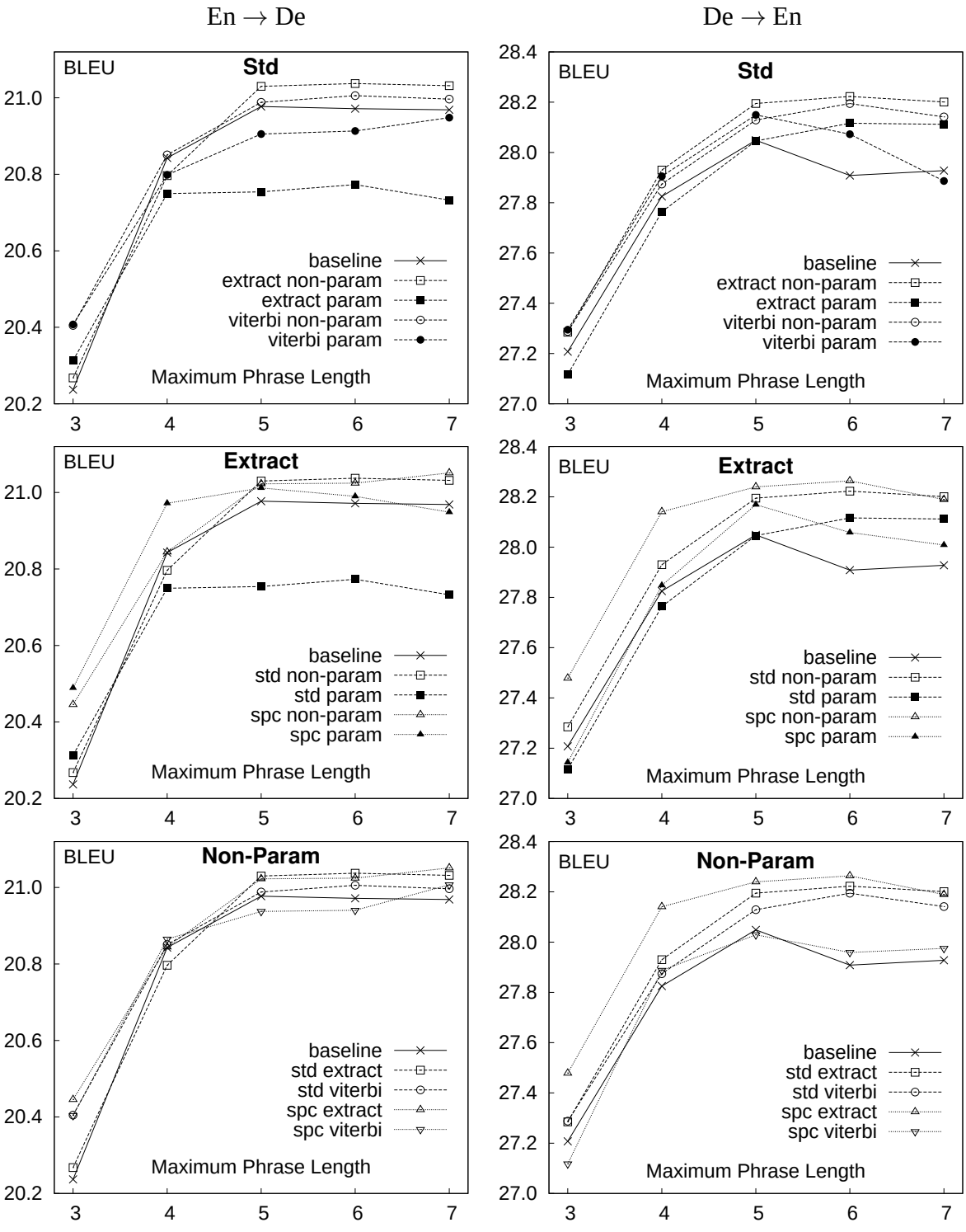

Figure 3.2: BLEU scores as a function of the maximum phrase length in English-German and German-English.

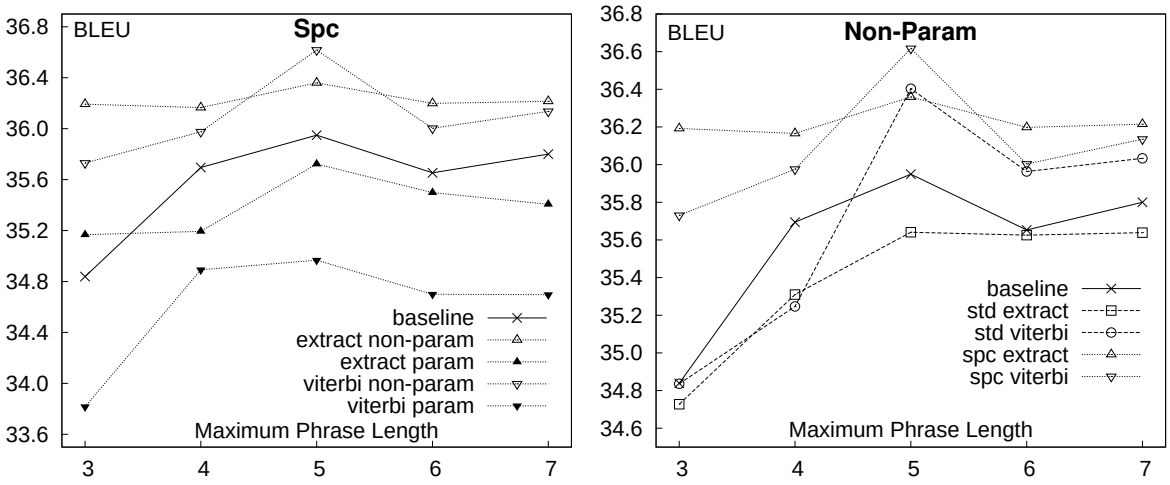

Figure 3.3: BLEU scores as a function of the maximum phrase length for Chinese-English BTEC task setting the specific model (left) and the non-parametric approach (right), while the other two experimental parameters are left free.

Table 3.3: Evaluation results in terms of BLEU, Probability of Improvement (PI) for BLEU, TER and PI for TER on English-Spanish (En-Es) and Spanish-English (Es-En) pairs.

System	En-Es				Es-En			
	BLEU $\pm$ 0.9	BLEU PI	TER $\pm$ 1.0	TER PI	BLEU $\pm$ 1.0	BLEU PI	TER $\pm$ 1.1	TER PI
baseline	32.0	-	54.2	-	32.9	-	52.6	-
std extr non-par	32.3	99.4	54.2	62.1	33.1	97.9	52.3	100.0
std extr param	32.2	90.8	54.3	34.3	33.0	78.1	52.6	55.0
std vite non-par	32.2	87.5	53.9	100.0	33.2	97.7	52.4	99.0
std vite param	32.4	100.0	54.2	57.9	33.2	98.7	52.3	99.6
spc extr non-par	32.2	93.9	54.1	76.8	33.2	98.6	52.5	94.6
spc extr param	32.2	91.7	54.3	26.7	33.2	98.4	52.4	96.1
spc vite non-par	32.1	84.9	54.3	15.1	33.2	97.3	52.4	98.9
spc vite param	32.2	92.1	54.1	87.8	33.0	76.4	52.4	98.9

Table 3.3 presents the results for the English-Spanish (En-Es) and Spanish-English (Es-En) pairs. As observed, confidence intervals overlap in all cases for TER and BLEU measures. However, PI for BLEU figures reflect that most of the systems proposed supersedes the baseline system in more than 90% of the bootstrapping rounds. This is not so clear in PI for TER on the English-Spanish pair, but again on the Spanish-English we observe the superiority of the phrase length models proposed.

Table 3.4 provides experimental results on English-German and German-English pairs in a similar fashion to Table 3.3. Again, we observed that the confidence intervals for BLEU and TER between the baseline system and the proposed systems overlap. Indeed, PI for BLEU on English-German do not reflect a notable superiority of phrase length systems over the baseline, but PI for TER clearly does for at least four of our models. Nevertheless, the analysis of PI on the German-English for BLEU and TER provides statistically significance

Table 3.4: Evaluation results in terms of BLEU, Probability of Improvement (PI) for BLEU, TER and PI for TER on English-German (En-De) and German-English (De-En).

System	En-De				De-En			
	BLEU $\pm$ 0.8	BLEU PI	TER $\pm$ 0.9	TER PI	BLEU $\pm$ 1.0	BLEU PI	TER $\pm$ 1.0	TER PI
baseline	21.0	-	65.8	-	27.9	-	58.6	-
std extr non-par	21.0	72.4	65.8	40.0	28.2	99.7	58.4	93.1
std extr param	20.7	4.7	65.8	46.7	28.1	92.9	58.2	100.0
std vite non-par	21.0	67.5	65.5	98.5	28.1	99.1	58.4	96.2
std vite param	20.9	47.7	65.5	98.0	27.9	53.5	58.7	19.2
spc extr non-par	21.1	73.2	65.6	95.5	28.2	99.5	58.3	99.4
spc extr param	21.0	52.2	65.5	99.5	28.0	89.4	58.3	98.8
spc vite non-par	21.0	55.9	65.3	100.0	28.0	74.7	58.5	74.8
spc vite param	20.9	21.8	65.8	61.2	27.9	71.7	58.5	77.8

Table 3.5: Evaluation results in terms of BLEU, Probability of Improvement (PI) for BLEU, TER and PI for TER on Chinese-English (Zh-En).

System	Zh-En			
	BLEU	BLEU PI	TER	TER PI
baseline	35.8 $\pm$ 2.8	-	46.2 $\pm$ 2.3	-
std extract non-param	35.6 $\pm$ 2.9	42.3	44.0 $\pm$ 2.2	100.0
std extract param	35.7 $\pm$ 2.8	43.8	46.9 $\pm$ 2.4	11.3
std viterbi non-param	36.0 $\pm$ 2.9	64.7	45.4 $\pm$ 2.3	92.7
std viterbi param	35.1 $\pm$ 2.8	9.0	47.9 $\pm$ 2.3	0.0
spc extract non-param	36.2 $\pm$ 3.0	73.2	44.0 $\pm$ 2.2	100.0
spc extract param	35.4 $\pm$ 2.8	31.8	44.6 $\pm$ 2.2	99.5
spc viterbi non-param	36.1 $\pm$ 2.9	66.7	43.6 $\pm$ 2.1	100.0
spc viterbi param	34.7 $\pm$ 3.0	11.8	43.4 $\pm$ 2.0	100.0

evidence of the superiority of some of our proposed models.

Experimental results on the BTEC Chinese-English task are displayed on Table 3.5. As happened in the other language pairs, confidence intervals overlap, though some of the proposed models obtain a higher (not statistically significant) average performance than the baseline system. Nonetheless, four of the proposed models outperform in terms of TER the baseline system in all bootstrapping rounds. However, the same cannot be claimed when analysing the PI for BLEU.

As mentioned above, there are language pairs for which the Poisson distribution (parametric) is a better parametrisation than a contingency table (non-parametric). For instance, the Poisson distribution obtains surprisingly good results in the English-Spanish pair compared with its non-parametric counterpart. However, in the German-English pair, we observed that the non-parametric models performs better. A possible explanation is a mismatch between the underlying probability distribution and the Poisson distribution. In order to verify this hypothesis, Figure 3.4 plots the standard model probabilities learnt with the parametric and non-parametric parametrisation for a fixed target length of 7 on the aforementioned language pairs. In the English-Spanish pair, the non-parametric models seem to approximate the learnt Poisson distribution. However, this is not the case in the German-English pair, in which the non-parametric model learns a completely different probability distribution. Note that we are

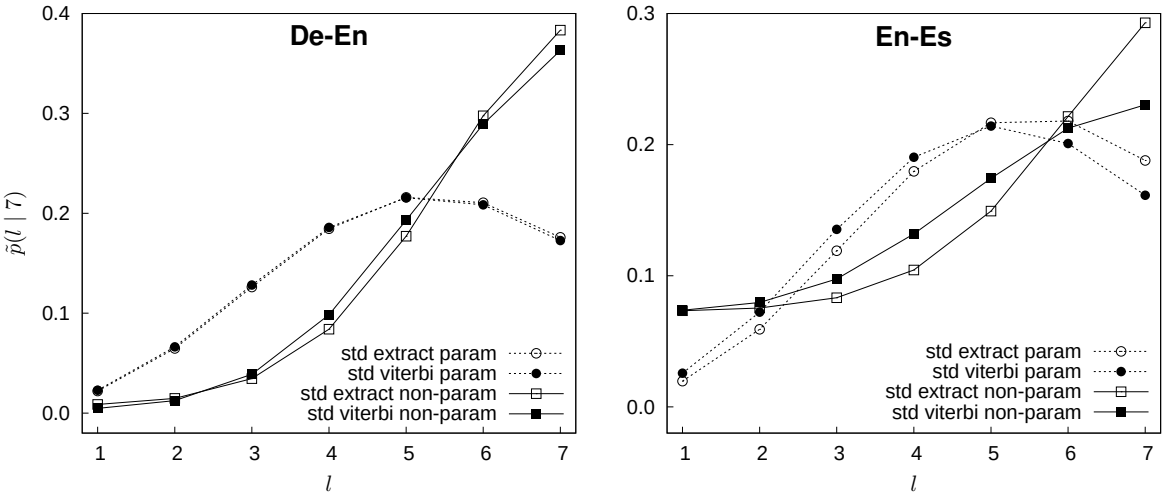

Figure 3.4: Probabilities learnt with standard model for both parametrisation and estimation algorithms. Vertical axis plots the learnt probability for a target phrase length of 7 as a function of the source phrase length in the horizontal axis.

already comparing smoothed models in order to avoid overfitting on the training data. Furthermore, we have also analysed this behaviour in the case of specific models, in which the disagreement is more accentuated.

In Table 3.6, positive and negative translation examples were selected to illustrate the behaviour of phrase length models on the Spanish-English task. Each example shows the source sentence, the reference translation and the translation provided by the baseline system followed by translations generated by a system augmented with phrase length features. Those phrase length systems providing the same translation are referred to as *others* and common suffixes are replaced by "...". In the first example, as a side effect of a better phrase length model, the standard parametric model improves both, the quality of the translation and the sentence length as a byproduct. In Table 3.3, the standard model approximated by a Poisson distribution obtains the best results when trained using Viterbi counts. Indeed, it is the only system able to balance the implicit length modelling features and the conditional phrase length in this sentence. A similar result is observed in the second example, where both parametric models trained with Viterbi counts produce a better translation. Finally, the last example is a difficult sentence for which no system obtains a good translation. The proposed length models in this case decrease the performance of the system in terms of TER. However, the translation is similar for many of the length models. It is interesting to analyse in this case how the Spanish word "desde" which means "from" is not translated by the baseline system, but some of the length models introduce it, even at the expense of other word. In general, the appropriate length model yields similar or better translations than the baseline system both in terms of TER and BLEU, as shown in Table 3.3.

Table 3.6: Translation examples on the Spanish-English pair. Phrase length systems providing the same translation are referred to as *others* and common suffixes are replaced by "…".

Length models improve evaluation measures	
source	nosotros hemos votado en contra .
reference	we voted against it .
baseline	we voted against .
std vite param	we voted against *it* .
others	we voted against .
source	estos documentos sumamente secretos nos proporcionan una extraña mirada entre bastidores de …
reference	these extremely secret documents give us a rare look behind the scenes of …
baseline	these documents secret extremely strange, give us a hidden behind the scenes of …
std vite param	these *highly secret documents* give us a *strange* look behind the scenes of the policy of …
spc vite param	these *highly secret documents* give us a *strange* look behind the scenes of the policy of …
others	these documents extremely secrets provide us with a strange look behind the scenes of …
Length models degrade evaluation measures	
source	desde el grupo socialista estimamos que el actual funcionamiento de la administración pública comunitaria es …
reference	the socialist group considers that the current functioning of the community's public administration is …
baseline	we in the socialist group believe that the current functioning of the european public service is …
std vite param	from the group of the party of european socialists, we believe that the current functioning of
spc extr non-par	from the socialist group , we believe that the current functioning of the european public service is …
spc vite non-par	from the socialist group we believe that the current functioning of the european public service is …
others	we in the socialist group believe that the current functioning of the european public service is …

3.5 Conclusions

In contrast to the conventional implicit length modelling features present in state-of-the-art SMT systems, we propose two novel explicit conditional phrase length models: the standard length model and the specific length model. These two models can be seamlessly derived from a generative bilingual segmentation process as shown in Section 3.3. Although previous work [1] addresses explicit phrase length modelling for a hidden semi-Markov model, no systematic evaluation in a state-of-the-art system has been performed to the best of our knowledge.

The proposed models have been parametrised in two different ways, using a contingency table or assuming a Poisson distribution. In addition, two alternative parameter estimation methods have been also presented: a heuristic algorithm based on the well-known phrase-extract algorithm, and a maximum likelihood estimation method based on the Viterbi segmentation.

These phrase-length models have been integrated in a state-of-the-art log-linear SMT system as additional feature functions, providing in most cases a systematic boost of translation quality on unrelated language pairs. This improvement, albeit not being large, has been proved to be statistically significant for several language pairs: English to/from Spanish, English to/from German and Chinese to English.

From the comparison of phrase-length models and parameter estimation approaches it has been observed that, as theoretically expected, there is a trade-off between model complexity and data sparseness. While for the Spanish pairs, a simple but properly estimated model (standard model) suffices, other languages require a more complex and flexible model (specific model). Regarding the estimation procedures a similar behaviour is observed. On the one hand, whenever a simple model is enough to model bilingual phrase length correlations, the Viterbi approach obtains a reliable and accurate estimation making the most out of the model. On the other hand, for complex models, the phrase-extract produce a better estimation since more approximated counts are generated to better estimate the parameters.

In the light of the results, as future work, we plan to perform a full Viterbi-like iterative training algorithm that may improve the quality obtained by the proposed Viterbi-based estimation method, for example using n-best segmentation lists instead of one simple segmentation. Moreover, we would also study as a smoothing technique, the combination of Viterbi extracted counts with those heuristically extracted. Finally, alternative optimisation methods to MERT, such as MIRA [6], will also be explored.

Bibliography

[1] J. Andrés-Ferrer and A. Juan. A phrase-based hidden semi-markov approach to machine translation. In *Procedings of European Association for Machine Translation, 2009*, pages 168–175, 2009.

[2] M. Bisani and H. Ney. Bootstrap estimates for confidence intervals in asr performance evaluation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004*, pages 409–412, 2004.

[3] P. F. Brown, J. C. Lai, and R. L. Mercer. Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics, 1991*, pages 169–176, 1991.

[4] P. F. Brown, S. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[5] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics, 1996*, pages 310–318, 1996.

[6] D. Chiang, Y. Marton, and P. Resnik. Online large-margin training of syntactic and structural translation features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 224–233, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statistical Society. Series B*, 39(1):1–38, 1977.

[8] Y. Deng and W. Byrne. HMM word and phrase alignment for statistical machine translation. *IEEE Trans. Audio, Speech, and Language Processing*, 16(3):494–507, 2008.

[9] W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, 1996*, pages 177–184, 1991.

[10] A. Giménez, J. Andrés, and A. Juan. Modelizado de la longitud para la clasificación de textos. In *Actas del I Workshop de Reconocimiento de Formas y Análisis de Imágenes, 2005*, pages 21–28, 2005.

[11] S. Günter and H. Bunke. HMM-based handwritten word recognition: on the optimization of the number of states, training iterations and gaussian components. *Pattern Recognition*, 37(10):2069–2079, 2004.

[12] R. Kneser. Statistical language modeling using a variable context length. In *Proceedings of the 4th International Conference on Spoken Language, 1996*, pages 494–497, 1996.

[13] K. Knight. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25:607–615, 1999.

[14] P. Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2004*, pages 388–395, 2004.

[15] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit, 2005*, pages 79–86, 2005.

[16] P. Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010.

[17] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*, 2007.

[18] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, 2003*, pages 48–54, 2003.

[19] E. Matusov, A. Mauser, and H. Ney. Automatic sentence segmentation and punctuation prediction for spoken language translation. In *Proceedings of the 3rd International Workshop on Spoken Language Translation, 2006*, pages 158–165, Kyoto, Japan, 2006.

[20] F. J. Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, 2003*, pages 160–167, 2003.

[21] M. Paul. Overview of the IWSLT 2009 Evaluation Campaign. In *Proc. of the International Workshop on Spoken Language Translation, 2009*, pages 1–18, Tokyo, Japan, 2009.

[22] H. S. Sichel. On a distribution representing sentence-length in written prose. *J. Roy. Statistical Society. Series A*, 137(1):25–34, 1974.

[23] O. Uzuner and B. Katz. A comparative study of language models for book and author recognition. *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, pages 969–980, 2005.

[24] A. Venugopal, S. Vogel, and A. Waibel. Effective phrase translation extraction from alignment models. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 7-12 July 2003, Sapporo Convention Center, Sapporo, Japan.*, pages 319–326, 2003.

[25] R. Zens and H. Ney. N-gram posterior probabilities for statistical machine translation. In *Human Language Technology Conference, North American Chapter of the Association for Computational Linguistics Annual Meeting, Workshop on Statistical Machine Translation, 2006*, pages 72–77, New York City, 2006.

[26] B. Zhao and S. Vogel. A generalized alignment-free phrase extraction. In *Proceedings of ACL Workshop on Building and Using Parallel Texts, 1995*, pages 141–144, 1995.

[27] M. Zimmermann and H. Bunke. Hidden markov model length optimization for handwriting recognition systems. In *Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition, 2002*, pages 369–374, 2002.

CHAPTER 4

EFFICIENT AUDIO SEGMENTATION FOR SPEECH DETECTION

Contents

4.1 Introduction

Obtaining a full set of transcriptions for an entire video lecture repository is a high time-consuming task, even if it is automated by taking profit of ASR systems. The temporal cost of generating an automatic transcription strongly depends on the length of the input audio signal. In order to make this process more efficient, audio streams are split into homogeneous acoustic regions which identify different acoustic classes, such as speech, silence, noise or music segments. This task is carried out by Audio Segmentation systems. Then, the detected speech segments are delivered to the ASR system, while all other non-speech classes are discarded.

It is important to note that the audio segmentation process must be as fast as possible, since its application is mainly motivated by making the whole process of generating automatic transcriptions more efficient. However, a prior segmentation of the audio signal can also provide a better transcription quality. For example, if we feed an ASR system with an audio signal containing both speech and non-speech segments, the ASR system will try to generate the closest transcription also for the non-speech regions, producing undesired outputs. Even more, automatic audio segmentations can lead better transcription outputs in comparison with manual segmentations, as proved in [8].

This chapter describes an efficient audio segmentation system that was developed to be part of our proposed transcription and translation platform (see Chapter 5). This system also participated in the Albayzin Audio Segmentation Evaluation 2012.

Previous work on audio segmentation can be classified into those tackling this task at the feature extraction level [1, 4, 5, 7], and those approximations working at the classification level [2, 6]. This latter approximation is adopted in our audio segmentation system.

The rest of the chapter is organised as follows. Section 4.2 describes the corpora used to train, tune and evaluate the system. Next, a complete system description is provided in Section 4.3. Experimental results are presented in Section 4.4. Then, Section 4.5 summarises the participation of this system in the Albayzin Audio Segmentation Evaluation 2012. Finally, conclusions and future work are drawn in Section 4.6.

4.2 Corpora

The corpora used to train, tune and test the audio segmentation system was the Albayzin 2012 corpus, consisting of a Catalan broadcast news database from the 3/24 TV channel, which comprises 87 hours of acoustic data for training purposes. In this dataset, speech (sp) can be found in a 92% of the segments, music (mu) is present a 20% of the time and noise (no) in the background a 43%. Regarding the overlapped classes, 40% of the time speech can be found along with noise and 15% of the time speech along with music.

In addition, two sets, *dev1* and *dev2*, from the Aragón Radio database of the Corporación Aragonesa de Radio y Televisión (CARTV), are used for developing and internal testing purposes, respectively. Both sets sums up to 4 hours of acoustic data. All audio signals are provided in PCM format, mono, little endian 16 bit resolution, and sampling frequency of 16 kHz.

Table 4.1 shows basic statistics for the train, dev1 and dev2 sets. In addition, Table 4.2

Table 4.1: Basic statistics of the Albayzin Corpus

	train	dev1	dev2
Time (h)	87.5	2.8	2.5

Table 4.2: Audio time distribution of all overlapping classes for the training set.

Class	Time (h)	Time (%)
sp	31.9	38.2
sp+no	31.4	37.6
sp+mu	12.6	15.1
mu	4.9	5.9
$\varnothing$	4.2	4.8
sp+no+mu	1.7	2.0
no	0.9	1.1
no+mu	0.1	0.1
Total	87.5	100

shows the audio time distribution over all overlapping acoustic classes as disjoint sets for the training set.

4.3 System description

As mentioned in section 2.2, audio segmentation takes into consideration a reduced set of acoustic classes, being in this case the power set of all segment classes found in the Albayzin database: Speech (sp), music (mu) and noise (no), plus a silence (si) class used to denote that none of the three classes is present in a given time instant. Thus, the system vocabulary is defined as

$$\mathcal{C} = \{sp, mu, no, sp + mu, sp + no, mu + no, sp + mu + no, si\} \tag{4.1}$$

In our case, it should be noted that each word/class is composed by a single phoneme.

According to Eq. 2.2, the audio segmentation system is composed by an acoustic model and a language model.

On the one hand, acoustic models were trained on MFCC feature vectors computed from acoustic samples. We used a 0.97 coefficient pre-emphasis filter and a 25 ms Hamming window that moves every 10 ms over the acoustic signal. From each 10ms frame, a feature vector of 12 MFCC coefficients is obtained using a 26 channel filter bank. Finally, the energy coefficient and the first and second time derivatives of the cepstral coefficients are added to the feature vector.

Each segment class is represented by a single-state HMM without loops, and its emission probability is modelled by a GMM. Acoustic HMM-GMM models were trained using

Table 4.3: SER figures for the three acoustic classes (speech, music, noise) in isolation and overall SER, computed over the *dev1* and *dev2* sets.

	Speech	Music	Noise	Overall
dev1	1.2	25.3	71.4	24.9
dev2	2.2	20.2	71.2	26.4

the TLK toolkit [3], which implements the conventional Baum-Welch algorithm. For each segment class, the number of mixture components per state was tuned on the development set.

On the other hand, a 5-gram back-off language model with constant discount was trained on the sequence of class labels at the speech frame level using the SRILM toolkit [9]. Constant discounts for each order were optimised on the development set. The segmentation process (search) was also carried out by the TLK toolkit.

4.4 Experimental results

This section is devoted to the description of the experimental setup and results performed before submitting our final audio segmentation system. For these experiments, acoustic and language models were trained on the *training* set, while acoustic and language model parameters were tuned on the *dev1* set. Table 4.3 shows SER figures computed on the *dev1* and *dev2* sets. In addition to the overall SER, SER values are provided for each acoustic class (speech, noise, music) in isolation.

As observed in Table 4.3, our audio segmentation system offers an excellent performance in speech detection, so it could be successfully employed to provide speech segments to speech recognition systems, which is our main aim.

However, the system provides low performance at detecting non-speech classes, specially the noise class. This fact can be explained by two reasons. First, we are using a feature representation of the acoustic signal that is focused on highlighting human voice characteristics, and conversely to abate acoustic features from music and noise. Secondly, few music and noise data samples appear in isolation (5% and 1%, respectively) to make feasible to robustly estimate acoustic models for these classes. For this reason, the global classifier suffers from a bias towards the isolated speech class. For instance, the posterior probability of an *sp+no* segment, given the isolated speech model parameter set is expected to be larger than that of the isolated noise model parameter set.

Regarding speed, our system is extremely fast, with an average RTF near to zero (RTF $= 0.001$, i.e. segmenting an audio signal of 2 hours takes approximately 7 seconds).

4.5 Albayzin 2012 Evaluation

The AS system described here participated in the Albayzin 2012 Evaluation. In this edition, 6 different systems from five Spanish research groups participated in this evaluation campaign.

Table 4.4: Segmentation error rate (SER) for the three acoustic classes (speech, music, noise) in isolation and overall SER, computed over blind *test* set of the Albayzin 2012 evaluation.

	Speech	Music	Noise	Overall
test	1.9	36.8	46.5	26.5

Table 4.5: Final standings for the Audio Segmentation Competition of the Albayzin 2012 Evaluations.

Pos.	System	SER
1	AHOLAB-EHU	26.3
2	**PRHLT-UPV**	**26.5**
3	GTM-UVIGO	28.1
4	CAIAC-UAB	33.3
5	GTTS-EHU-2	39.6
6	GTTS-EHU-1	40.0

All participants were committed to detect overlapping speech, music and noise segments on a blind test. This test set comprises 18 hours length from the Corporación Aragonesa de Radio y Televisión (CARTV), as the dev1 and dev2 sets from the Albayzin database.

Table 4.4 shows the performance of our system in terms of SER for the three acoustic classes, including the overall SER, over the blind test. These results are consistent with the ones shown in Table 4.3.

Finally, Table 4.5 shows the competition's final standings. Our system achieved the 2nd position, very close to the winner [10]. However, it must be noted that our system (RTF $= 0.001$) was considerably faster than the winner (RTF $= 1.6$). For example, in absolute terms, our system processed the blind test (18 hours length) in 64 seconds, while the winner system required about 29 hours, as inferred from their results [10].

4.6 Conclusions

In this chapter we have described the Audio Segmentation system that was integrated into our transcription and translation platform. The system tackles the task of audio segmentation from the viewpoint of an ASR system with a reduced vocabulary set. Experimental results show, on the one hand, that our system provides excellent performance detecting speech segments, and in the other hand, that it is extremely fast, providing real-time audio segmentations. Also, this system participated in the Audio Segmentation Competition of the Albayzin 2012 Evaluations, achieving the 2nd place, very close to the winner system in terms of SER.

As future work, we will study a hybrid DNN-HMM approach for audio segmentation, that is, the replacement of the emission probabilities of the HMMs by DNNs instead of GMMs.

Bibliography

[1] J. Ajmera, I. McCowan, and H. Bourlard. Speech/music segmentation using entropy and dynamism features in a hmm classification framework. *Speech Communication*, 40(3):351–363, 2003.

[2] A. Bugatti, A. Flammini, and P. Migliorati. Audio classification in speech and music: A comparison between a statistical and a neural approach. *EURASIP J. Audio Speech Music Process.*, pages 372–378, 2002.

[3] M. A. del Agua, A. Giménez, N. Serrano, J. Andrés-Ferrer, J. Civera, A. Sanchís, and A. Juan. The translectures-upv toolkit. In *Advances in Speech and Language Technologies for Iberian Languages - Second International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain, November 19-21, 2014. Proceedings*, pages 269–278, 2014.

[4] A. Gallardo-Antolín and J. M. Montero. Histogram equalization-based features for speech, music, and song discrimination. *IEEE Signal processing letters*, 17(7):659–662, 2010.

[5] T. Izumitani, R. Mukai, and K. Kashino. A background music detection method based on robust feature extraction. In *Proc. of ICASSP*, pages 13–16.

[6] Y. Lavner and D. Ruinskiy. A decision-tree-based algorithm for speech/music classification and segmentation. *EURASIP J. Audio Speech Music Process.*, 2009:2:1–2:14, 2009.

[7] C. Panagiotakis and G. Tziritas. A speech/music discriminator based on rms and zero-crossings. *IEEE Transactions on Multimedia*, 7(1):155–166, 2005.

[8] B. Ramabhadran, J. Huang, U. V. Chaudhari, G. Iyengar, and H. J. Nock. Impact of audio segmentation and segment clustering on automated transcription accuracy of large spoken archives. In *INTERSPEECH*, 2003.

[9] A. Stolcke. SRILM - an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*, 2002.

[10] D. Tavarez, E. Navas, D. Erro, and I. Saratxaga. Audio segmentation system by aholab for albayzin 2012 evaluation campaign. In *Proceedings of IberSPEECH 2012*, pages 577–584, Madrid (Spain), 2012.

CHAPTER 5

THE TRANSLECTURES-UPV PLATFORM

Contents

5.1 Introduction

In this chapter, an initial system architecture is described to support cost-effective transcription and translation of large video lecture repositories, which is the main goal of this thesis. This architecture was adopted in the EU project transLectures, whose main aim was to achieve just this through the use of advanced ASR and MT technologies. The starting hypothesis in transLectures was that the gap that must be bridged by these technologies in order to achieve acceptable results for the kind of audiovisual collections being considered is relatively small. In transLectures, two key lines of research were pursued: massive adaptation and intelligent (user) interaction [39].

Massive adaptation refers to process of exploiting the wealth of knowledge available about these video lecture repositories (lecture-specific knowledge, such as speaker, topic and slides) to create a specialised, in-domain transcription and translation system. A system adapted using this knowledge is therefore likely to produce a far better ASR and MT output than a general-purpose system.

Intelligent interaction is the process of human-computer interaction whereby we can exploit feedback from the user or prosumer community of a given video lecture repository. For instance, we can more or less count on a university lecturer being willing to devote a few minutes of his time to correcting any errors in the automatic transcript generated for a lecture he has recently recorded. The intelligent part comes in when deciding exactly which segments of the transcript the lecturer should be asked to interact with. For example, the system should not simply present the first minute of a lecture for review, since this section may well be perfectly transcribed and need no manual corrections. This would be a waste of user effort. Instead, an intelligent system should first identify which section(s) of the lecture contain the most errors; that is, which section(s), based on its automatic confidence measures, are most likely to contain errors, and then present these sections only to the the user for correction.

The above ideas were tested on two case studies: VideoLectures.NET [41] and poliMedia [38]. VideoLectures.NET is an online video repository with more than 19.000 talks (12.000 hours) given by top researchers in various academic settings. poliMedia is a collection of over 15.000 videos (3.000 hours) recorded by course lecturers under controlled conditions at the Universitat Politècnica de València, Valencia (Spain). Both repositories are active players in the diffusion of the open-source Opencast (formerly Matterhorn) platform [14] currently being adopted by many education institutions and organisations within the Opencast community. Indeed, a third key premise of transLectures was to use (and develop) a system architecture that works with Opencast, to allow the rapid adoption and real-life testing of transLectures technologies.

In this chapter we will focus on the description and deployment of the aforementioned system architecture over the poliMedia repository. From now, we will refer to this system as *The transLectures-UPV Platform* (TLP).

The rest of the chapter is structured as follows. First, the poliMedia media repository and the transLectures project are described in Sections 5.2 and 5.3 respectively. Next, the proposed architecture is described in detail in Section 5.4. Then, in Section 5.5, empirical results are reported on the quality of the transcriptions and translations of poliMedia lectures currently being maintained and steadily improved. Finally, some key conclusions are drawn in Section 5.6.

Table 5.1: Basic statistics of the poliMedia repository (September 2015).

Lectures	15436
Duration (hours)	3079
Avg. Lecture Length (minutes)	12
Speakers	1759
Avg. Lectures per Speaker	8

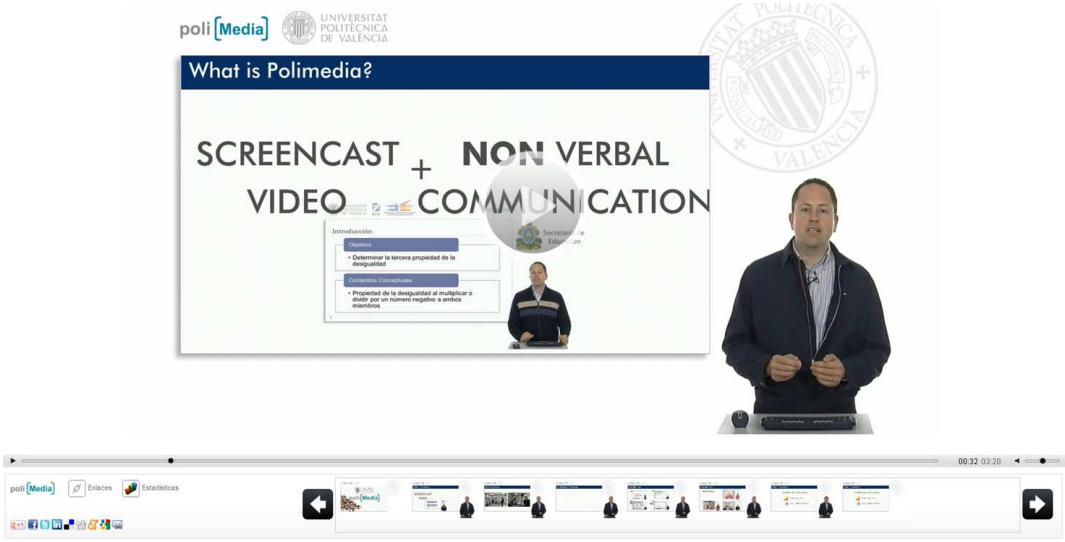

Figure 5.1: Example of a poliMedia lecture.

5.2 poliMedia

poliMedia is a service for the creation and distribution of multimedia educational content at the UPV. It was designed primarily to allow UPV professors to record their courses in videos lasting around 10 minutes and accompanied by time-aligned slides. It serves more than 36,000 students and 2800 university lecturers and researchers. poliMedia began in 2007 and has already been exported to several universities in Spain and South America. Table 5.1 shows basic statistics of the poliMedia repository. The vast majority of poliMedia lectures (88%, 2.800h) are recorded in Spanish, although we can find recordings in other languages such as English (7%, 159h) or Catalan (3%, 49h), among others.

poliMedia recordings are made up of two videos stacked horizontally: one of the slides and another of the speaker, with a resolution of 1280x720 points. See Figure 5.1 for an example of a poliMedia recording.

A reduced set of choices for the recordings can be chosen by speakers, but specific adjustments are not allowed, in order to confer more homogeneity to the recordings. The reduced set of setups for the recordings, depicted in Figure 5.2 are the following, from left to right and from top to bottom: full person shot with slide, half person shot with slide, white blackboard, and top camera view.

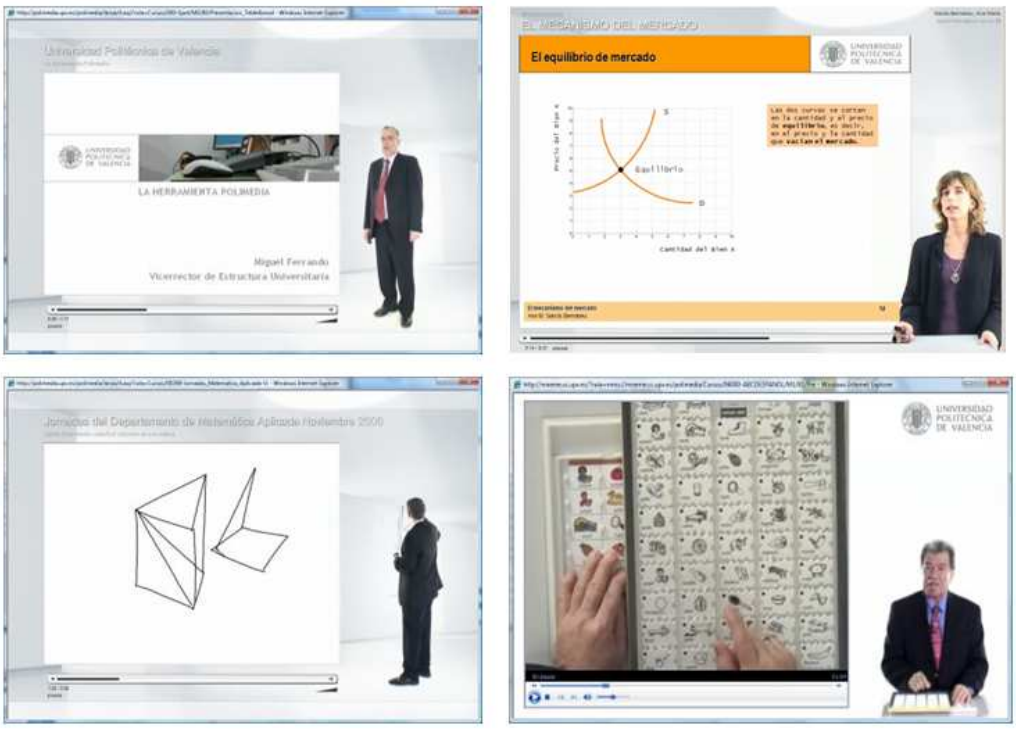

Figure 5.2: Setup choices for poliMedia

The slide area is a raw capture from the screen of the speaker's computer, so, in addition to slides, it can show websites, computer applications, or any other graphical resource that can be executed in a computer. Furthermore, speakers are capable to write over the image displayed on the slides area using a tablet PC connected to the computer.

The production process for a poliMedia repository has been carefully designed to achieve both a high rate of production and a high quality output, comparable to that of a TV production but at a lower cost. A poliMedia production studio consists of a 4x4 metre room with a white background in which we find a video camera; a capture station; a pocket microphone; lighting; and A/V equipment including a video mixer and an audio noise gate. It is worth noting that the use of lighting in such a small set allows us to get a sharper image much more easily than in a lecture recording hall.

The recording process is quite simple: lecturers (speakers) are invited to come to the studio with their presentation and slides. They deliver their lecture, while they and their computer screen are recorded in two different streams. The two streams are put side-by-side to generate a raw preview of the poliMedia content, which can be reviewed by the speaker at any time.

Figure 5.3 shows a picture taken in the UPV poliMedia recording studio during a recording session.

If the speaker is satisfied with the end result, a post-process is applied to the raw recordings, which includes cropping, joining (with a little overlap) and h.264 encoding, in order to

Figure 5.3: A poliMedia recording session at the UPV.

generate a mp4 file suitable for distribution. This process is fully automatic, meaning that the speaker can review the post-processed video file in just a few minutes. Next, the mp4 file is distributed online via a streaming server.

5.3 transLectures

transLectures [37, 39], acronym of *Transcription and Translation of Video Lectures*, was an EU (FP7-ICT-2011-7) STREP project in which advanced automatic speech recognition and machine translation techniques were tested on large video lecture repositories. The project started in November 2011 and finished in October 2014. The transLectures consortium included video lecture providers (users), experts in ASR and MT, and professional transcription and translation providers:

- Universitat Politècnica de València (UPV), Valencia, Spain (Coordinator).
- Xerox S.A.S. (XEROX), Grenoble, France.
- Jozef Stefan Institute (JSI), Ljubljana, Slovenia.
- Rheinisch-Westfaelische Technische Hochschule (RWTH), Aachen, Germany.
- European Media Laboratory GmbH (EML), Heidelberg, Germany.

- Deluxe Digital Studios Ltd (DDS), London, UK.

Although the transLectures project involved all these institutions and companies, it is important to note that the work and results reported in this thesis were exclusively generated by the UPV, being the author of this thesis the designer and developer of TLP.

transLectures was grounded on the three following scientific and technological objectives:

1. Improvement of transcription and translation quality by massive adaptation.

   ASR had not yet revealed its full potential in the generation of acceptable transcriptions for large-scale collections of audiovisual objects. However, relatively little further research into ASR technology was required: we had to learn how to better exploit the wealth of knowledge we have at hand. More precisely, the aim was to demonstrate that acceptable transcriptions can be obtained through the massive adaptation of general-purpose models from lecture-specific knowledge such as speaker, topic and, more importantly, lecture slides. It is only by having acceptable transcriptions that adaptation of translation models can also provide acceptable results.

2. Improvement of transcription and translation quality by intelligent interaction.

   Massive adaptation can deliver substantial contributions to the improvement of overall quality, but sufficiently accurate results are unlikely to be obtained through fully-automated approaches alone. Instead, in order to reach the desired levels of accuracy, user interaction needs to be taken into consideration. The typical user models for the transcription and translation of audiovisual objects are batch-oriented. Under this model, an initial transcription/translation is first computed by the system off-line and then sent to the user to be post-edited manually without system participation. However, these models only yield satisfying results when highly collaborative users are working on near-perfect system output. Otherwise, a more intelligent interaction model that saves on user supervision and allows the system to learn from user supervision actions is proposed.

3. Integration into Opencast to enable real-life evaluation.

   In contrast to many past research efforts in which system prototypes are evaluated in the lab alone and are largely inapplicable to real-life settings, transLectures aimed to develop tools and models for use with Opencast, in order to evaluate their usefulness using real-life data in a real-life context.

All transLectures ideas were tested on the VideoLectures.NET and poliMedia media repositories. For automatic transcription the English and Slovenian languages were considered in VideoLectures.NET, whilst Spanish was chosen for poliMedia. Meanwhile, automatic translation was carried out from {Spanish, Slovenian} into English, and English into {French, German, Slovenian, Spanish}.

5.4 System Architecture

This section describes the architecture of the *transLectures-UPV Platform* (TLP) that was developed and integrated for the first time with the poliMedia repository on June 2013. Fig-

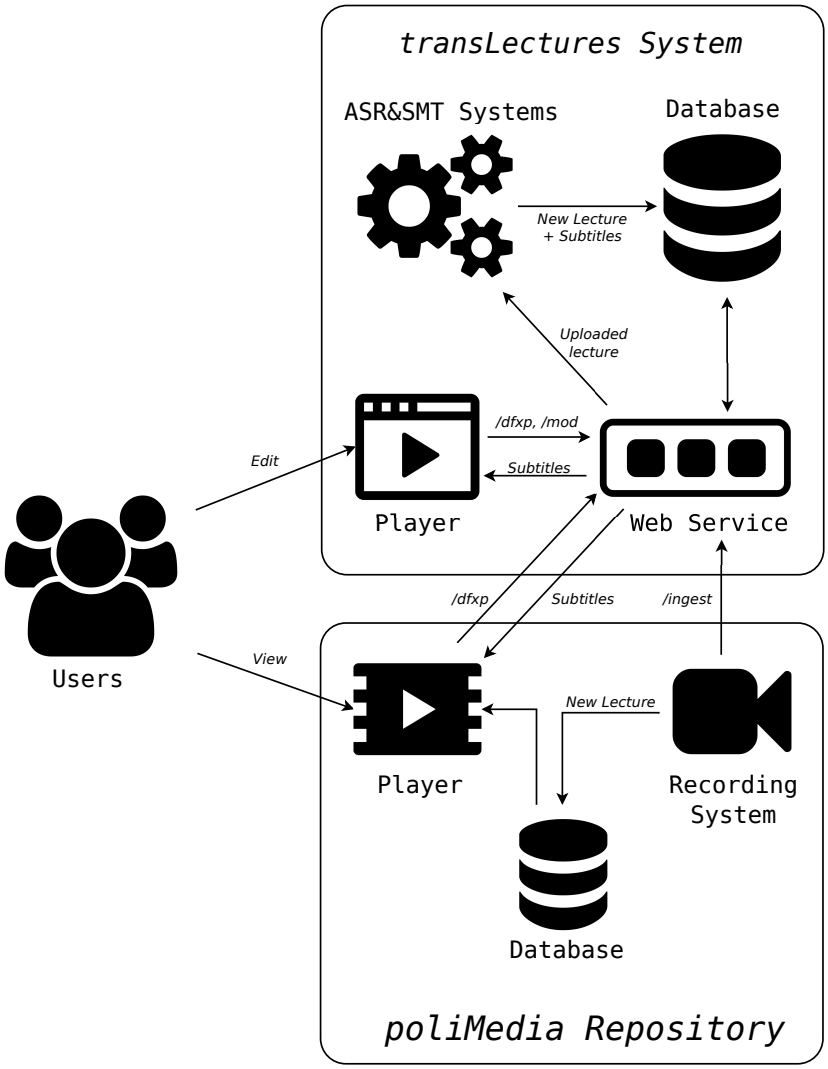

Figure 5.4: Overview of TLP when deployed over the poliMedia repository.

ure 5.4 shows a global overview of the components and processes involved in TLP when deployed in a media repository such as poliMedia.

First of all, we have identified two use cases to cover the different ways in which the user is able to interact with the system. Specifically, the user may want simply to view a video lecture and its corresponding subtitles, which would be the first use case, or he might choose to also edit/supervise the automatic subtitles, which would be the second.

In the first use case, the user browses the poliMedia catalogue and selects the lecture he wants to watch. The user then is redirected to the poliMedia player (Figure 5.1), where he can watch the requested lecture. This player allows the user to select subtitles for the video in different languages, where available. The list of languages available is obtained by sending a request to an external service: the TLP Web Service (see Section 5.4.1). This Web Service checks to see if there are subtitles available for a given lecture and, if so, in which languages. Once the user has selected a language, the player sends a request to the web service asking for

the corresponding subtitles file. This file is then sent in DFXP format [44] and immediately displayed on the player. All subtitles are generated automatically by the ASR and SMT systems (see Section 5.4.4). The generated subtitles along with the corresponding metadata, are stored in the TLP Database (see Section 5.4.3).

In the second use case, the user, while watching a lecture with the corresponding subtitles as described in the first use case, notices that the subtitles displayed contain transcription errors and decides to correct them. The poliMedia player offers the possibility to redirect the user to the TLP Player, a tool with subtitle editing capabilities, among other features (see Section 5.4.2). Corrections made by the user are sent back to the Web Service, appending them into the original DFXP file. The DFXP format is extended in this use case in order to be able to track the history of modifications made by users and the automatic systems, allowing the Player to show the best subtitles available for every segment. In other words, a DFXP file can be understood as a mini repository of caption modifications.

TLP needs to be permanently synchronised with the poliMedia repository in order to provide transcriptions and translations for any newly recorded videos. For this purpose, the TLP Web Service provides a lecture upload service, which is used by the poliMedia recording system. Then, once a new lecture has been uploaded to the Web Service, the transcription of this lecture and its subsequent translation into different languages is carried out by the ASR and SMT systems.

It is worth noting the distributed nature of TLP architecture (i.e. that each component could be deployed on a different machine), although in this case study all components were hosted by a single machine mounting a Intel i7 CPU with 64GB of RAM.

In the following sections, we give more detailed descriptions of the key components of TLP.

5.4.1 Web Service

The TLP Web Service is the interface for exchanging information and data between the poliMedia repository and TLP. It also enables the subtitle visualisation and editing capabilities of the TLP Player. This Web Service is implemented as a python Web Server Gateway Interface (WSGI), and defines a set of HTTP interfaces related to caption delivery and media upload:

- */ingest*: POST request which allows the client to upload audio/video files and other related material, such as slides and textual resources, which could be useful for adapting the ASR system. This POST request also allows additional metadata to be submitted, such as the language of the recording and speaker ID. This metadata can be used to enhance the accuracy of the ASR system by applying speaker adaptation techniques. The automatic translation is then generated from the automatic transcription, and both are stored in DFXP format.

- */status*: GET request to check the status of a video lecture uploaded through the ingest interface.

- */langs*: GET request that provides the client with a list of the subtitles and languages available for a given lecture.

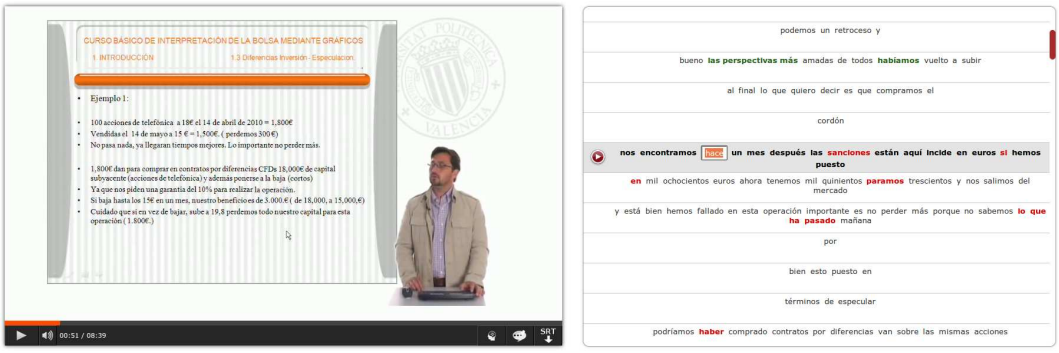

Figure 5.5: TLP Player showing the default side-by-side editing layout. The video playback is shown on the left-hand side of the screen, whilst the transcription editor appears on the right-hand side.

- */dfxp*: GET request that returns the subtitles in DFXP format for a specific lecture and language.
- */mod*: POST request that sends and commits changes made by a user when supervising a transcription or translation.

All these interfaces operate with the Database, which stores all information needed by the Web Service.

5.4.2 Player

Although the quality of automatic transcriptions and translations that TLP provide are accurate enough to be considered useful (see Section 5.5.1), they may contain errors that need to be corrected by the user. A PHP/HTML5 video player and caption editor was carefully designed to expedite the error supervision task and obtain subtitles of an acceptable quality in exchange for a minimum amount of user effort. The TLP Player was developed in accordance with Nielsen's usability principles [27, 28], and it was iteratively improved during the user evaluations described in Section 5.5.2.

Three alternative editing layouts were available for users to choose from according to their personal preferences. Figure 5.5 shows the default side-by-side editing interface with the video playback on the left and transcription editor on the right.

At that time, the TLP Player offered two interaction modes: batch/post-editing interaction, in which the user can freely supervise any segment of the video; and intelligent interaction, in which the player asks the user to correct only those words considered most likely to be incorrect by the system. Figure 5.6 shows a zoomed screenshot of the intelligent interaction mode. Probably incorrect words that need revision from the user appear highlighted in red. Both post-editing and intelligent interaction strategies are evaluated under an experimental setup with real users in Section 5.5.2.

Additionally, a complete set of key shortcuts were implemented to enhance expert user capabilities.

Figure 5.6: A zoomed screenshot of the transcription interface of the TLP Player in intelligent interaction mode.

Figure 5.7 shows an alternative editing interface where the transcription is located below the video, which is intended to bring a clear look of what is being shown on the video in case it contains relevant information.

Captions are retrieved from the Web Service through the */dfxp* interface, which provides the complete video transcription or translation in DFXP format. As already indicated, the availability of language-specific subtitles can be consulted by querying the Web Service */langs* interface. Once retrieved, users are able to correct transcription/translation errors and save modifications back to the Web Service through its */mod* interface.

Figure 5.8 shows the third editing interface, which is suitable for users who do not pretend to make a full supervision of the video but to perform specific modifications.

5.4.3 Database

The TLP Database stores all the data required by the Web Service, and the ASR and SMT systems. Specifically, it stores the following entities:

- Lectures: All the information related to a specific lecture is stored in the Database, such as language, duration, title, keywords and category. In addition, an external ID, recognised by the poliMedia database, is stored and used in all transactions performed between the poliMedia player and the Web Service for lecture identification purposes.

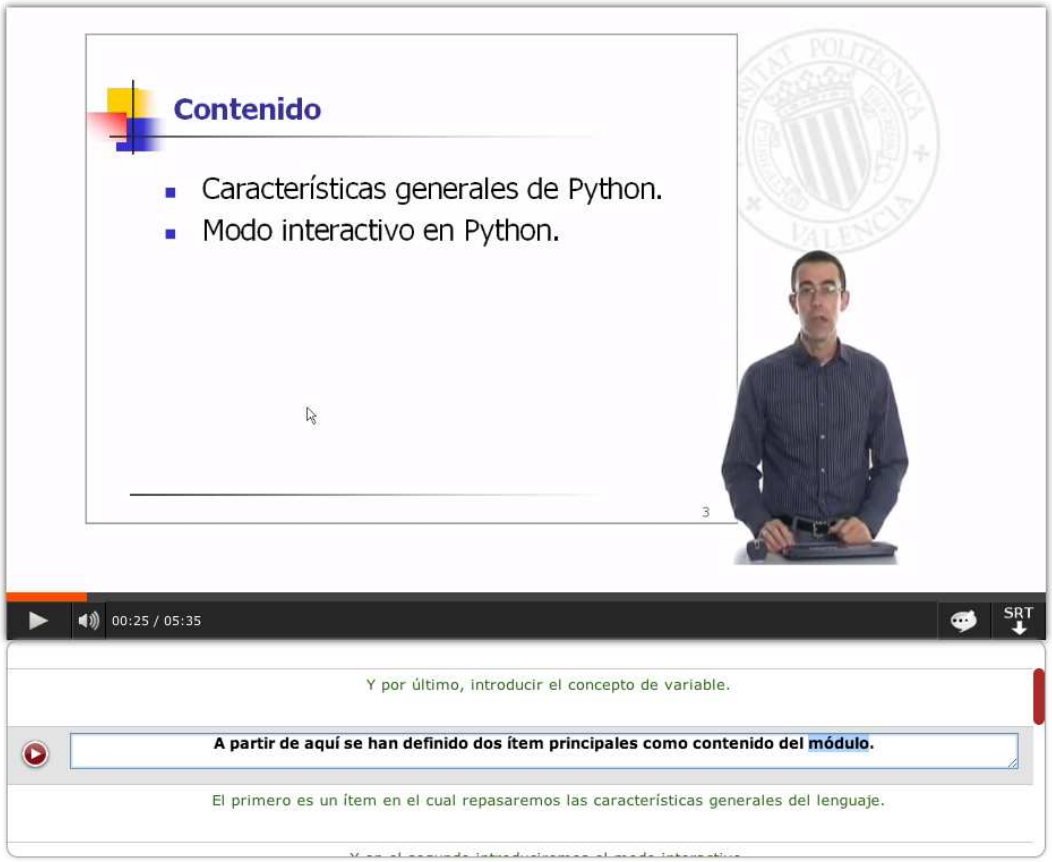

Figure 5.7: TLP Player, second layout.

- Speakers: The information about the speaker of a given lecture can be exploited by the ASR system, adapting the underlying models to the speech peculiarities of a specific speaker. The result is a better transcription and, consequently, a better translation.
- Captions: All subtitles generated by the ASR and SMT systems are kept in the Database and retrieved by the Web Service.
- Uploads: Every time a Web Service */ingest* operation is requested by the poliMedia recording system, a new entry is stored in the Database. Once the computation of the automatic transcription and translation of the lecture is performed, a new entry is added to the lectures, speakers and subtitles entities.

5.4.4 ASR & SMT Systems

ASR and SMT systems are key components of TLP. These systems generate the automatic transcriptions and translations of every lecture available in the poliMedia repository. They are designed to exploit all available information regarding the repository in order to enhance transcription and translation quality.

Figure 5.8: TLP Player, third layout.

On the one hand, regarding Automatic Speech Recognition, information about the speaker is used to inform acoustic model adaptation techniques, while text extracted from slides are used to adapt language models to the specific topic of the lecture.

On the other hand, translations of a given lecture into different languages are generated from its automatic transcription. This means that translation accuracy is highly correlated with transcription quality; that is, the better the transcription, the better the translation. SMT systems are adapted to the topic of the lecture by updating the translation and language models with related training data extracted from out-of-domain parallel corpora.

It must be noted that TLP was designed to easily allow an automatic regeneration of subtitles following major upgrades to the ASR or SMT systems. These upgrades might be the result of better acoustic, translation or language models, or of new ASR and SMT techniques. This means that the repository's overall transcription and translation quality can be constantly improved.

In the initial deployment of TLP over the poliMedia repository, only Spanish video lectures were considered, and thus, an Spanish ASR system was built to generate a complete set of transcriptions for all those videos. Additionally, an Spanish to English SMT system was trained to generate automatic English translations from the Spanish transcriptions. Below it follows a detailed description about the training of these systems.

Spanish ASR System

The first Spanish ASR system was based on the conventional GMM-HMM approach, using the TLK toolkit [4] to train acoustic models and the SRILM toolkit [36] to deploy n-gram language models.

In order to train the baseline Spanish acoustic models, training speech segments were first transformed into 16KHz, and then parametrised into sequences of 39-dimensional real feature vectors. In particular, every 10 milliseconds 12 Mel Frequency Cepstral Coefficients (MFCC) [45], the log-energy, and their first and second order time derivatives were calculated. Regarding the transcriptions, 23 basic phonemes were considered plus a silence model. Words were transformed into phonemes according to the Spanish pronunciations rules. The resulting phoneme transcriptions also include speech disfluencies as hesitations, incorrect pronunciations, etc.

The previously extracted features were used to train tied triphoneme HMMs with Gaussian mixture models for emission probabilities. The training scheme was similar to the one described in [45]. Firstly, each Spanish phoneme was modelled as a three state HMM with a conventional left to right topology. Secondly, these HMMs were used to initialise triphonemes which were extracted by modelling each phoneme with its context. Thirdly, an automatic tree-based clustering based on manually crafted rules was used to tie the original states [46]. As a product of the previous process, tied triphoneme HMMs with only one Gaussian component per state were obtained similarly to [46]. Finally, mixture components in each state were repeatedly split using an iterative training process. During all training steps the Baum-Welch algorithm was used to estimate the model parameters that maximises the log-likelihood for the training data [30, 45].

Two speaker adaptation techniques were applied in order to improve the baseline acoustic models: fast vocal tract length normalisation (VTLN) [42], and Constrained Maximum Likelihood Linear Regression (CMLLR) [8, 10, 35].

On the one hand, the motivation of VTLN is that different speakers usually have vocal tracts of different lengths. These different lengths result in linear shifts of the formant frequencies (frequency regions where the acoustic energy of the voice is high). To alleviate this variability, frequencies from the acoustic data are linearly transformed (warped) to maximise the log-likelihood of the training data.

On the other hand, the CMLLR technique aims to linearly transform the mean and variance of all Gaussian mixture components from all HMM states in order to better fit to the acoustic features of the speaker. Let μ_{si} and Σ_{si} be the mean and variance of the i-th Gaussian mixture component of an HMM state s. The CMLLR technique computes a new $\hat{\mu}_{si}$ and $\hat{\Sigma}_{si}$ such that

$$\hat{\mu}_{si} = A\mu_{si} + b \qquad \text{and} \qquad \hat{\Sigma}_{si} = A\Sigma_{si}A^T$$

for an adaptation matrice A and an adaptation vector b that maximise the log-likelihood of the input data. However, it is more convenient to apply this transformation to the input features instead to the acoustic models, as it has been demonstrated to be equivalent [8]. Thus, CMLLR adaptation introduces a second recognition step, in which the output of the first recognition step is used to estimate the adaptation matrices A and b, used afterwards to transform the standard input features to CMLLR features used in the second recognition step [35].

Regarding language modelling, the baseline language model was a linearly interpolated 4-gram language model composed by several n-gram models which were trained with different in-domain and out-of-domain corpora, one per corpus. All n-gram models were smoothed with modified Kneser-Ney absolute interpolation method [15]. Interpolation weights were optimised to minimise the perplexity on a given in-domain development set [13].

To improve our baseline language model, a 3-gram language model trained with the text from the slides of the video to be recognised is added to the interpolation scheme [23]. Unluckily, the poliMedia repository does not keep a separated digital text version (i.e. PDF, PPT) of the slides for each video lecture, and therefore text information has to be extracted directly from the video. To this purpose, we used the free Optical Character Recognition (OCR) software Tesseract [34] accompanied with proper pre-processing and post-processing steps as done in [23], although this automatic process can introduce text recognition errors that might ruin the potential improvement that slides may introduce.

Finally, the well-known Viterbi algorithm, implemented in the TLK toolkit, is used to automatically recognise the videos lectures.

Spanish→English SMT System

The Spanish into English translation system initially deployed was based on the state-of-the-art phrase-based SMT toolkit Moses [17] for training translation models, and on the SRILM toolkit [36] to train n-gram language models.

Obtaining accurate automatic translations of educational videos is one of the most challenging tasks due to the lack of in-domain parallel data to train SMT systems. However, there exist large amounts of out-of-domain parallel corpora which are publicly available. Since our objective was to use lecture-specific knowledge in order to adapt the translation models to the lecture domain, we applied a sentence selection technique called Infrequent n-gram Recovery or Infrequent n-gram Sentence Selection (ISS) [9]. This technique selects the minimal set of sentences from an out-of-domain parallel corpora that yield reliable estimations for all the n-grams that occur in the text to be translated, which, in our case, is the transcription of the video lecture.

When selecting sentences, it is important to choose sentences that contain n-grams that have rarely occurred in the training corpus but that are going to be used in the translation. Such n-grams are denoted as *infrequent n-grams*, i.e. n-grams that occur less than a given threshold t, the so-called infrequent threshold.

Sentences from the out-of-domain pool are sorted by their infrequency score in order to select the most informative ones. The infrequency score depends on the infrequent score that the source n-grams to be translated receive. Let $\mathbf{w}$ be a n-gram from the set of all n-grams $\mathcal{X}$ that appear in the source text; $C(\mathbf{w})$ the occurrences of $\mathbf{w}$ in the source part of the training corpus; and $N(\mathbf{w})$ the counts of $\mathbf{w}$ in the source sentence f from the out-of-domain pool to be scored. The infrequency score $i(\cdot)$ of the source sentence f is defined as

$$i(f) = \sum_{\mathbf{w} \in \mathcal{X}} \min(1, N(\mathbf{w})) \max(0, t - C(\mathbf{w})) \tag{5.1}$$

In order to avoid giving a high score to noisy sentences with a lot of occurrences of the same infrequent n-gram, only one occurrence of each n-gram is taken into account to compute the score. Additionally, the infrequency score tends to give more importance to the n-grams with the lowest counts in the training corpora. Consequently, the score is updated every time a sentence is added to the training set. The process is stopped whenever a maximum number of sentences is obtained, or no infrequent n-ngram is left in the training corpora [9]. Finally, the selected sentences from the out-of-domain corpora are added to the in-domain training data

(if available) in order to create the final training data set. This final training set is used then to train the phrase-based and word-based translation models with the Moses SMT toolkit.

Regarding language modelling, we trained a large linearly interpolated English n-gram language similarly to the Spanish ASR system. Each individual language model was trained in a different corpus. All models used order 4 and were smoothed with modified Kneser-Ney absolute interpolation method [15]. Interpolation weights were optimised to minimise the perplexity on a given in-domain development set [13]. In addition, we considered to use a second n-gram language model trained with the in-domain training data obtained after applying the ISS technique. Both language models were incorporated into the log-linear translation model as new feature functions [18].

Finally, the built-in decoder from the Moses toolkit was used to translate the input sentences given by the Spanish ASR system into English.

5.5 System Evaluation

TLP was evaluated at the first stage from two different viewpoints. On the one hand, Section 5.5 describes how transcription and translation quality was automatically assessed to gauge the performance of the underlying ASR and SMT systems. On the other hand, Section 5.5.2 shows how user satisfaction and productivity were analysed through real user evaluations.

5.5.1 Automatic Evaluations

In this section we empirically assess the performance of the Spanish ASR and the Spanish into English SMT systems described in Section 5.4.4. After an exhaustive analysis of several massive adaption techniques that exploit lecture-related information available a priori in the repository, best systems on both tasks were put into production to automatically generate Spanish and English subtitles for all Spanish poliMedia lectures.

Spanish ASR System

To train and evaluate the Spanish ASR system, 114 hours of Spanish poliMedia video lectures were manually transcribed as part of the transLectures project. This manually transcribed data was partitioned into training, development and test sets that were allocated to train, tune and evaluate our ASR system. Statistics on these three sets are shown in Table 5.2.

On the one hand, the acoustic models were trained using the training set of the poliMedia corpus. In summary, the ASR system accounted 1745 HMM triphonemes with 3342 tied-states having up to 64 Gaussians per state. The training configuration was tuned on preliminary experiments using the development set.

On the other hand, a baseline 4-gram language model was trained by interpolating several individual 4-gram models built from different in-domain and out-of-domain corpora. Table 5.3 summarises the main statistics of these out-of-domain corpora. Speech transcriptions from the training set of the Spanish poliMedia corpus were used to train the in-domain

Table 5.2: Statistics on the training, development and test sets of the Spanish poliMedia speech corpus.

	Training	Development	Test
Videos	655	26	23
Speakers	73	5	5
Hours	107h	3.8h	3.4h
Sentences	39.2K	1.3K	1.1K
Running Words	936K	35K	31K
Vocabulary Size	26.9K	4.7K	4.3K

Table 5.3: Basic statistics of the external corpora involved in the generation of the Spanish language model for the Spanish ASR system.

Corpus	Sentences	Running Words	Vocabulary
EPPS [1]	132K	0.9M	27K
News-Commentary-v8 [3]	183M	4.6M	174K
TED	316M	2.3M	133K
Europarl v7 [16]	2.1M	55M	439K
El Periódico	2.7M	45M	916K
News (07-11)	8.6M	217M	2.9M
United Nations [6]	9.5M	253M	1.6M
UnDoc	10M	318M	1.9M
Google-Counts-v1 [24]	-	44.8G	2.2M

language model. Interpolation weights were optimised on the Spanish poliMedia development set. The vocabulary of the resulting interpolated language model was restricted to the 60K most frequent words.

Finally, recognition parameters such as the grammar scale factor or the word insertion penalty [45] were tuned on the development set of poliMedia.

Table 5.4 reports WER results on the test set of the Spanish poliMedia corpus. The baseline system achieved 30.3 WER points. After applying the VTLN and CMLLR speaker adaptation techniques, the WER was reduced to 24.8% WER points, that is, a 18% relative improvement compared with the baseline.

Once determined the best speaker-adapted acoustic model, we assessed the impact of the adaptation of language models to the text from the slides extracted via OCR. In order to measure the full potential of this adaptation technique, we also considered an optimistic scenario in which a text version of the slides is available or the OCR system provides error-free transcriptions. To do this, text transcriptions from the slides of all development and test videos of the Spanish poliMedia corpus were manually annotated. With these references we were able to measure the quality of the OCR system, obtaining 64% WER points in the slides of the test set. Although this WER is high, the experiments revealed that the ASR performance was significantly increased by using this automatic slides.

Table 5.4: Evolution of the WER of the Spanish ASR system computed on the test set of the Spanish poliMedia corpus, as speaker adaptation techniques are applied to the baseline system.

	WER
Baseline	30.3
+ VTLN	28.8
+ CMLLR	24.8

Table 5.5: WER of the Spanish ASR system computed on the test set of the Spanish poliMedia corpus using topic-adapted language models (using text extracted from slides).

	WER
Baseline (BL)	24.8
BL + Correct slides	21.4
BL + OCR slides	23.8

Results in terms of WER are summarised in table 5.5. The first row depicts the results obtained with the baseline language model without any information from the slides. The second row adds a 3-gram language model trained with the human annotated slides for each video. It is observed that error free slide text heavily improves performance, with a relative WER improvement of 14%. Similarly to the second row, the third row adds a slide-dependent 3-gram to the linear interpolation but using the slide transcriptions obtained automatically via OCR instead of the error-free transcriptions. When comparing automatic versus error-free transcriptions, it is observed that roughly 70% of the improvement is lost. However, even a slide-dependent language model obtained from noisy text outperforms the baseline.

Spanish→English SMT System

To evaluate the Spanish into English SMT system, both development and test sets from the Spanish poliMedia corpus (see Table 5.2) were manually translated into English. In addition, 8 hours from training set were also manually translated to be used as a small in-domain training data. Table 5.6 summarises the main statistics of this Spanish-English parallel corpus.

On the one hand, our baseline translation model was trained using the in-domain poliMedia training set plus the Europarl [16] corpus, while topic-adapted translation models were trained using an appropriate set of in-domain training sentences. This set was composed by the in-domain poliMedia training set plus a set of relevant sentences extracted from an out-of-domain pool of parallel data populated by the Europarl and the United Nations [6] corpora. Table 5.7 summarises the most important statistics of these corpora.

On the other hand, we trained a large linearly interpolated English 4-gram language model using the English part of the Spanish-English poliMedia training set as in-domain corpora

Table 5.6: Main statistics of the parallel Spanish-English poliMedia corpus.

	Sentences	Running Words		Vocabulary	
		en	es	en	es
Training	1.5K	40.2K	40.3K	3.9K	4.7K
Development	1.4K	38.7K	37.8K	3.7K	3.5K
Test	1.1K	32.1K	32.1K	3.3K	4.1K

Table 5.7: Main figures of the out-of-domain corpora from which sentences were selected using the ISS technique in order to train the first topic-adapted Spanish into English SMT system.

	Sentences	Running Words		Vocabulary	
		en	es	en	es
Europarl-v7 [16]	2M	54.5M	57M	132K	195K
United Nations [6]	11.2M	320M	366M	132K	195K

(see Table 5.6), as well as several bilingual and monolingual corpora as out-of-domain corpora. Table 5.8 summarises the most important statistics of these out-of-domain corpora. All individual 4-gram models were interpolated to minimise the perplexity of the Spanish-English poliMedia development set.

Table 5.9 shows the results obtained when evaluating the Spanish into English translation quality on the test set of poliMedia. These results are reported in terms of BLEU. The baseline system scored 24.0 BLEU points. This result was significantly improved to 25.3 BLEU points by re-training the translation models using a proper in-domain training set obtained by applying the ISS technique. Finally, the inclusion into the log-linear translation model of the small in-domain language model, trained with the selected in-domain data, led to the best result of 26.0 BLEU points.

Table 5.8: Basic statistics of the external corpora involved in the generation of the large English language model for the Spanish into English SMT system.

Corpus	Sentences	Running Words	Vocabulary
TED	142K	2.3M	100K
News Commentary	208K	4.5M	150K
Europarl v7 [16]	2.2M	54.1M	326K
United Nations [6]	10.6M	286M	1.8M
GIGA [2]	19.8M	504.8M	5.8M
News (07-11)	48.8M	986M	6.1M
Google-Counts-v1 [24]	-	356.3G	7.3M

Table 5.9: Evolution of BLEU computed on the test set of the Spanish-English poliMedia corpus as topic adaptation techniques are incorporated to the baseline Spanish into English SMT system.

	BLEU
Baseline	24.0
Topic Adaptation (ISS)	25.3
+ in-domain LM	26.0

5.5.2 User Evaluations

In this section we describe the evaluations carried out with real users to assess TLP under a real-life scenario. These evaluations were done under the Docència en Xarxa (Online Teaching) programme of the Universitat Politècnica de València (UPV) in collaboration with the transLectures project (see Section 5.3). Docència en Xarxa is an on-going incentive-based programme to encourage university lecturers at the UPV to develop digital learning resources.

Methodology

A total of 27 lecturers signed up for this study, reviewing a sample of 86 video lectures organised into three phases. Most participants had degrees in different branches of engineering (17), while the rest mastered business management (6), social science (2) and biology (2).

The participants were asked to review the automatic transcriptions of five of their own poliMedia videos using the TLP Player described in Section 5.4.2. Participants were free to choose where and when to review those transcriptions, without our supervision. These videos were transcribed with the Spanish ASR system described in Section 5.4.4. Lectures to be reviewed were allocated across the following three evaluation phases:

1. Post-editing: The automatic transcription for the first video of each participant is manually and freely reviewed until obtaining a perfect transcription.

2. Intelligent interaction: Given the automatic transcription of the second and third videos, only a subset of probably incorrectly-recognised words are reviewed.

3. Two-step review: This phase was organised in two consecutive rounds of evaluation for the fourth and fifth videos. The first round mimics phase two above, where the lecturer reviewed only the least confidence words. Once this first round is completed, the video is then automatically re-transcribed on the basis of the lecturer's review actions preserving their corrections. In a second round, the updated transcriptions are completely reviewed as in the phase one.

During these trials, the TLP Player logged precise user interaction statistics, from which we computed two of the main variables of this study: Word Error Rate and Real Time Factor. Also, having both variables in hand, we also assessed the WER reduction per RTF unit for the three aforementioned evaluation phases. In addition, we collected feedback from the

participants as subjective statistics after each phase, in the form of a brief satisfaction survey based on [21]. Lecturers were asked to rate various aspects related to intuitiveness, likeability and usability on a Likert scale from 1-10.

Below we describe the main experimental results attained over the three evaluation phases.

First phase: Post-editing

In the first phase, 20 lecturers freely reviewed the automatic transcription of their first video lecture using the default post-editing interaction mode of the TLP Player, accounting for a total of 2.6 hours in 20 video lectures. WER and RTF values from the first phase are shown at the left-most plot of Figure 5.9, where each data point represents a supervision made by a lecturer on a certain video.

In order to evaluate the impact of automatic transcriptions on the total time required to generate usable subtitles, we compared the effort made by the participants in this first phase with the time needed to generate manual subtitles from scratch from a previous study [40]. We found a statistically significant difference between the mean RTFs for subtitles generated automatically (Mean (M)=5.4, Std (S)=2.9) and the mean RTFs for those generated manually from scratch (M=10.1, S=1.8). This suggests that the automatic transcriptions, at their reported accuracy in terms of WER, allow lecturers to generate subtitles more efficiently than manually from scratch. In this first phase, the mean WER reduction per RTF unit was 3.2 (S=1.3). Also, we found that RTF can be reliably explained as a function of WER by means of a linear regression:

$$\text{RTF} = 2.025 \cdot \log_e(\text{WER}) \tag{5.2}$$

This logarithmic adjustment can be explained by the fact that users essentially ignore automatic transcriptions above a certain WER threshold, preferring to transcribe from scratch.

Finally, the satisfaction survey revealed that users rated very positively (Overall Mean = 9.1) this user interaction strategy, designed in accordance with intuitiveness (9.3), likeability (8.8) and usability (8.9) principles.

Second phase: Intelligent Interaction

This second phase incorporated a new interaction strategy called intelligent interaction [33]. This strategy is based on the application of active learning (AL) techniques to ASR [5]. More concretely, we apply batch AL based on uncertainty sampling [20] using confidence measures [11, 31, 43], which provide the probability of correctness of each word appearing in the automatic transcription. The idea is to focus user's review actions on incorrectly-transcribed words saving time and effort. However, the lecturer may need to review (confirm) some correctly-recognised words incorrectly identified as errors (false positives), although many of the incorrectly-recognised words are spotted correctly (true positives).

In this phase, lecturers are requested to review a subset of least confidence words in increasing order of probable correctness. This subset typically constituted between 10-20% of all words, but lecturers could modify this range from 5% to 40% depending on the perceived accuracy of the transcription.

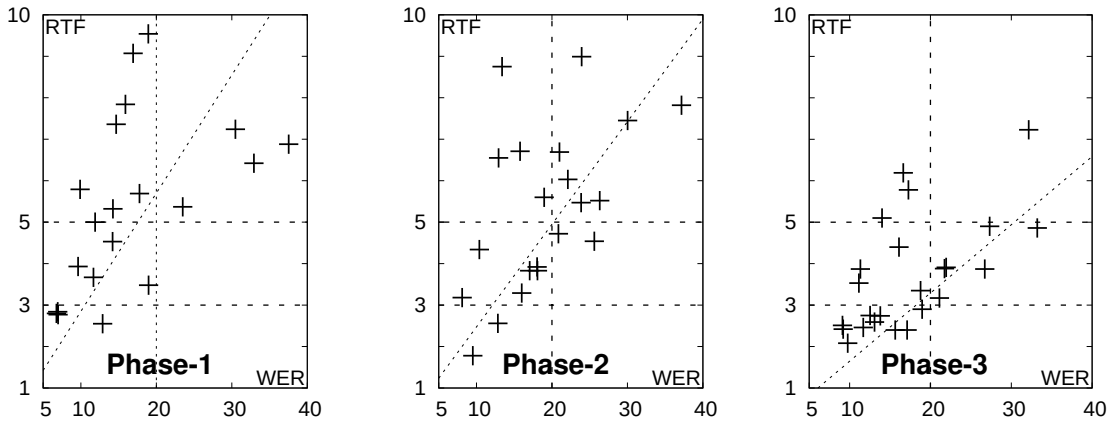

Figure 5.9: Evolution of RTF as a function of WER in the post-editing mode across the three phases. Data points of the second phase correspond to those lecturers that declined to use intelligent interaction and switch back to the conventional post-editing strategy. Data points of the third phase are those obtained in the second step of that phase.

Figure 5.6 shows a screenshot of the transcription interface of the TLP Player in this phase. Low-confidence words are shown in red and corrected low-confidence words in green. The text box including the low-confidence word can be expanded in both directions to correct the context, if needed. For this phase, the intelligent interaction mode was activated in the TLP Player by default, although lecturers could switch back to the post-editing mode as in the first phase.

Interaction logs shown that only 12 of the 23 participants from this second phase stayed in the intelligent interaction mode for the full review of their poliMedia videos (2.8 hours over 18 video lectures). In the other cases (3 hours over 22 video lectures), lecturers switched back to the post-editing mode. Participants wanted to make sure that perfect transcriptions were obtained no matter how much time could be saved by the intelligent interaction mode. WER and RTF values corresponding to those lecturers that declined to use intelligent interaction and switched back to the conventional post-editing strategy are shown at the plot in the middle of Figure 5.9.

For those lecturers that stayed in the intelligent interaction mode, mean review time was reduced to a RTF of 2.2, although the resulting transcriptions were not error-free, unlike in phase one. The mean residual WER of the transcriptions after being reviewed was 8.0, which is not so far from that achieved by non-expert transcriptionists [12]. This indicates that confidence measures successfully identify most of all incorrectly-recognised words.

The mean WER reduction per RTF unit in the intelligent interaction strategy was M=4.6 (S=3.9). In comparison with the conventional post-editing mode (M=3.9, S=1.3), we did not find statistically significant differences between both methods. This means that intelligent interaction is in fact just as efficient in terms of WER decrease per RTF unit as conventional post-editing.

Satisfaction surveys for this second phase were not as excellent as in the first phase (Overall Mean = 7.2). In comparison with the first phase, surveys statistically reflected that post-

editing (M=9.1, S=1.3) was preferred over intelligent interaction (M=7.2, S=1.7) by our lecturers. The figures collected on intuitiveness (8.1), likeability (6.8) and usability (6.3), dropping from the conventional post-editing phase, reflect this assessment. However, lecturers did seem to embrace confidence measures, suggesting that low confidence words denoted in red could be incorporated into the conventional post-editing strategy.

Third phase: Two-step Supervision

The third phase was organised into two sub-phases or rounds and is essentially a combination of the previous two phases. In this phase, lecturers first reviewed a subset of the least confidence words, as in the second phase. The videos were then re-transcribed on the basis of all previous review actions, preserving those corrections made by users. These updated transcriptions were expected to be of high quality than the original transcriptions [32], reducing overall review times. In the second round of this third phase, lecturers completely reviewed the regenerated transcription as in phase one. The fourth and fifth videos of each lecturer were reviewed in this phase.

This two-step review process was successfully completed by 15 lecturers on a total of 26 video lectures with 3.7 hours of video. More precisely, a total of 1 and 2.7 hours were reviewed in the first and second steps, respectively.

In the first step of this phase, average review time was 1.4 RTF. The reviewed transcriptions in this step, but also in phases one and two, were used to adapt the ASR system via a process of massive adaptation. Specifically, on the one hand, we adapted the acoustic features to the speaker with the CMLLR technique (see Section 5.5.1) using the reviewed transcriptions, and on the other hand, we adapted the language model by incorporating a small language model, trained with the reviewed transcriptions, into the language model interpolation scheme (see Section 5.5.1). Then, the automatic transcriptions were regenerated, preserving those segments already reviewed by lecturers, and using them to improve the recognition of the context words using a constrained search [19, 33].

As reported in Table 5.10, WER dropped significantly from the initial 28.4% to the regenerated transcriptions 18.7%. That is, almost 10 WER points over 1.4 RTF, meaning that intelligent interaction plus adaptation (M=8.6, S=5.8) achieved a higher statistically significant WER reduction per RTF unit than intelligent interaction alone (M=4.6, S=3.9). This suggests that intelligent interaction plus adaptation is, in fact, more effective in terms of WER decrease per RTF unit than intelligent interaction alone.

In the second step, lecturers completely reviewed the regenerated transcriptions to obtain perfect final subtitles, as in the first phase. Right-most plot of Figure 5.9 depicts RTF and WER values for each lecture supervised by the participants. Average RTF for this task stood at 3.9. As expected, when comparing WER reduction per RTF unit in the first phase (M=3.2, S=1.3) and the second step of this phase three (M=5.3, S=2.0), we can observe a statistically significant learning curve in lecturers' performance. As a result, we proved that there is a learning curve involved in getting to grips with the TLP Player.

In order to fairly compare the first (M=3.2, S=1.3) and third (M=5.3, S=2.0) phases in terms of WER reduction per RTF unit, we subtract the effect of the learning curve for each lecturer leading to a corrected WER reduction per RTF unit (M=4.7, S=2.8). Even so, we found a lower yet statistically significant difference in favour of the third phase explained

Table 5.10: Summary of results obtained in the two-step review phase

	WER	RTF	ΔRTF
Initial transcriptions	28.4	0.0	-
First step: Intelligent interaction	25.0	1.4	1.4
Massively adapted transcriptions	18.7	1.4	-
Second step: Complete review	0.0	5.3	3.9

by the application of massive adaptation. This result suggests that the two-step strategy is a bit more efficient than the conventional post-editing strategy. However, this statistically significant difference only holds when enough reviewed data is available for adaptation. That is, the reviewed data generated in the first step of this phase (1.0h) is not sufficient to improve the ASR performance so that the overall WER reduction per RTF unit of the third phase compared to that of the first phase is statistically higher.

The two-step supervision implied that lecturers have to put time aside on two separate occasions to review the same video. However, lecturers preferred to carry out the review process in a single step rather than in two steps. This fact was reflected on the average score of the user satisfaction surveys (M=7.8, S=2.0). For this reason, the two-step strategy was less preferred by lecturers than the post-editing strategy.

Discussion

Provided that the review of automatic transcriptions was more efficient than generate them from scratch, alternative user interaction strategies were explored to generate subtitles from automatic transcriptions as efficiently and comfortably as possible for our lecturers [26]. First of all, we determined that WER was the main factor involved in explaining the values of RTF.

In line with [22], more sophisticated user interfaces alone, like our intelligent interaction strategy, have not proved to be more efficient in terms of WER decrease per RTF unit than conventional post-editing, nor were they preferred by lecturers over the simple (though more time-costly) interaction model. We find it particularly noteworthy how important it was for lecturers to be able to produce high quality (perfect) end transcriptions, prioritising this over any time-savings afforded by the more intelligent strategies [7, 25, 29]: a full half of our lecturers reverted to the conventional post-editing model to complete the review of their video transcriptions.

Nevertheless, the combination of intelligent interaction with massive adaptation techniques led to statistically significant savings in user effort in comparison to intelligent interaction alone and, consequently, to the conventional post-editing strategy. This conclusion differs from that of [22] mainly because a greater amount of adaptation data has been used in our study. All in all, lecturers preferred the simple “one-step” post-editing strategy over the sophisticated two-step strategy.

5.6 Conclusions

In this chapter we have presented a system architecture that allows on-line video lecture repositories to provide users acceptable transcriptions and translations in exchange for relatively little user effort. The implementation of this architecture is called *The transLectures-UPV Platform*. We have also described TLP's main components: the Web Service, Player, Database, and the ASR and SMT systems; as well as how these components interact with each other under two use cases: viewing and editing video lectures. Preliminary automatic and human evaluations suggested, first, that the delivered transcriptions and translations were of an acceptable quality; second, that these automatic transcriptions were a very good start point for users to generate perfect transcriptions, saving about a half of the human effort needed to generate transcriptions from scratch; and third, that the TLP Player, the tool devoted to correct lecture subtitles, was very comfortable and easy to use.

It must be noted that TLP is still serving transcriptions and translations for the poliMedia repository at the time of writing. In the meantime, TLP has been significantly enhanced in terms of design and implementation. Furthermore, under the scope of the transLectures project, new ASR and MT systems have been integrated into TLP, and the existing ones have been progressively improved over time. An updated description of TLP, along with an extension of the preliminary results presented here, is presented in Chapter 8.

Bibliography

[1] The EPPS Speech corpus at ELRA. `http://catalog.elra.info/product_info.php?products_id=1035`.

[2] The GIGA corpus for the WMT'10. `http://www.statmt.org/wmt10/training-giga-fren.tar`.

[3] The News Commentary corpus for the WMT'13 (Monolingual). `http://www.statmt.org/wmt13/training-monolingual-nc-v8.tgz`.

[4] M. A. del Agua, A. Giménez, N. Serrano, J. Andrés-Ferrer, J. Civera, A. Sanchís, and A. Juan. The translectures-upv toolkit. In *Advances in Speech and Language Technologies for Iberian Languages - Second International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain, November 19-21, 2014. Proceedings*, pages 269–278, 2014.

[5] L. Deng and X. Li. Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):1060–1089, May 2013.

[6] A. Eisele and Y. Chen. Multiun: A multilingual corpus from united nation documents. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, 2010.

[7] B. Favre, K. Cheung, S. Kazemian, A. Lee, Y. Liu, C. Munteanu, A. Nenkova, D. Ochei, G. Penn, S. Tratz, C. R. Voss, and F. Zeller. Automatic human utility evaluation of ASR systems: does WER really predict performance? In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 3463–3467, 2013.

[8] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, 12(2):75–98, 1998.

[9] G. Gascó, M. Rocha, G. Sanchis-Trilles, J. Andrés-Ferrer, and F. Casacuberta. Does more data always yield better translations? In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*, pages 152–161, 2012.

[10] D. Giuliani, M. Gerosa, and F. Brugnara. Speaker normalization through constrained MLLR based transforms. In *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*, 2004.

[11] D. Z. Hakkani-Tür, G. Riccardi, and A. L. Gorin. Active learning for automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2002, May 13-17 2002, Orlando, Florida, USA*, pages 3904–3907, 2002.

[12] T. J. Hazen. Automatic alignment and error correction of human generated transcripts for long speech recordings. In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 2006.

[13] F. Jelinek and R. L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proc. of the Workshop on Pattern Recognition in Practice*, pages 381–397, Amsterdam, The Netherlands: North-Holland, 1980.

[14] M. Ketterl, O. A. Schulte, and A. Hochman. Opencast matterhorn: A community-driven open source solution for creation, management and distribution of audio and video in academia. In *ISM 2009, 11th IEEE International Symposium on Multimedia, San Diego, California, USA, December 14-16, 2009*, pages 687–692, 2009.

[15] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing, ICASSP '95, Detroit, Michigan, USA, May 08-12, 1995*, pages 181–184, 1995.

[16] P. Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT.

[17] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*, 2007.

[18] P. Koehn and J. Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 224–227, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[19] T. T. Kristjansson, A. Culotta, P. A. Viola, and A. McCallum. Interactive information extraction with constrained conditional random fields. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, July 25-29, 2004, San Jose, California, USA*, pages 412–418, 2004.

[20] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *In Proc. of ICML*, pages 148–156, 1994.

[21] J. R. Lewis. IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. *International Journal of Human-Computer Interaction*, 7(1):57–78, Jan. 1995.

[22] S. Luz, M. Masoodian, B. Rogers, and C. Deering. Interface design strategies for computer-assisted speech transcription. In *Proceedings of the 20th Australasian Computer-Human Interaction Conference, OZCHI 2008: Designing for Habitus and Habitat, Cairns, Australia, December 8-12, 2008*, pages 203–210, 2008.

[23] A. A. Martinez-Villaronga, M. A. del Agua, J. Andrés-Ferrer, and A. Juan. Language model adaptation for video lectures transcription. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 8450–8454, 2013.

[24] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[25] C. Munteanu, R. Baecker, G. Penn, E. Toms, and D. James. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pages 493–502, 2006.

[26] H. Nanjo and T. Kawahara. Towards an efficient archive of spontaneous speech: Design of computer-assisted speech transcription system. *The Journal of the Acoustical Society of America*, 120(5):3042–3042, 2006.

[27] J. Nielsen. User interface directions for the web. *Communications of the ACM*, 42(1):65–72, Jan. 1999.

[28] J. Nielsen and J. Levy. Measuring usability preference vs. performance. *Communications of the ACM*, 37(4):66–75, 1994.

[29] Y. Pan, D. Jiang, L. Yao, M. Picheny, and Y. Qin. Effects of automated transcription quality on non-native speakers' comprehension in real-time computer-mediated communication. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1725–1734, 2010.

[30] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice-Hall, 1993.

[31] G. Riccardi and D. Hakkani-Tur. Active learning: theory and applications to automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(4):504–511, 2005.

[32] I. Sanchez-Cortina, N. Serrano, A. Sanchís, and A. Juan. A prototype for interactive speech transcription balancing error and supervision effort. In *17th International Conference on Intelligent User Interfaces, IUI '12, Lisbon, Portugal, February 14-17, 2012*, pages 325–326, 2012.

[33] N. Serrano, A. Giménez, J. Civera, A. Sanchis, and A. Juan. Interactive handwriting recognition with limited user effort. *International Journal on Document Analysis and Recognition*, pages 1–13, 2013.

[34] R. Smith. An overview of the tesseract OCR engine. In *9th International Conference on Document Analysis and Recognition (ICDAR 2007), 23-26 September, Curitiba, Paraná, Brazil*, pages 629–633, 2007.

[35] G. Stemmer, F. Brugnara, and D. Giuliani. Adaptive training using simple target models. In *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005*, pages 997–1000, 2005.

[36] A. Stolcke. SRILM - an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*, 2002.

[37] The transLectures Consortium. transLectures Project Web Page. `http://translectures.eu`.

[38] Universidad Politècnica de València. The polimedia repository. http://media.upv.es, May 2012.

[39] UPVLC, XEROX, JSI-K4A, RWTH, EML, and DDS. transLectures: Transcription and Translation of Video Lectures. In *Proc. of EAMT*, page 204, 2012.

[40] J. D. Valor Miró, A. Pérez González de Martos, J. Civera, and A. Juan. Integrating a state-of-the-art ASR system into the opencast matterhorn platform. In *Advances in Speech and Language Technologies for Iberian Languages - IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings*, pages 237–246, 2012.

[41] VideoLectures.NET: Exchange ideas and share knowledge. `http://www.videolectures.net`.

[42] L. Welling, S. Kanthak, and H. Ney. Improved methods for vocal tract normalization. In *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '99, Phoenix, Arizona, USA, March 15-19, 1999*, pages 761–764, 1999.

[43] F. Wessel, R. Schluter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 9(3):288–298, 2001.

[44] World Wide Web Consortium (W3C). Distribution Format Exchange Profile (DFXP). `http://www.w3.org/tr/2006/cr-ttaf1-dfxp-20061116`.

[45] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK, 2006.

[46] S. J. Young, J. J. Odell, and P. C. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the workshop on Human Language Technology*, HLT '94, pages 307–312, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

CHAPTER 6

RECOMMENDER SYSTEMS FOR ONLINE LEARNING PLATFORMS

Contents

6.1 Introduction

In the previous chapter we have presented *The transLectures-UPV Platform*, an implementation of a system architecture that makes cost-effective transcription and translation of large video lecture repositories possible. The main motivation of this particular application of ASR and MT technologies was to make audiovisual resources accessible to speakers of different languages and to people with disabilities [4, 18] at large scale. However, the availability of transcriptions and translations for the whole repository enable numerous digital content management applications such as lecture categorisation, summarisation, automated topic finding, plagiarism detection or lecture recommendation, among others. This latter application has become essential for media repositories due to their fast growth and their increasing popularity. Users are often overwhelmed by the amount of lectures available and may not have the time or knowledge to find the most suitable videos for their learning requirements.

Up until recently, recommender systems have mainly been applied in areas such as music [8, 11], movies [2, 19], books [12] and e-commerce [3], leaving video lectures largely to one side. Only a few contributions to this particular area can be found in the literature, most of them focused on VideoLectures.NET [1]. However, none of them has explored the possibility of using lecture transcriptions to better represent lecture contents at a semantic level.

In this chapter we describe a content-based lecture recommender system that uses automatic speech transcriptions, alongside lecture slides and other relevant external documents, to generate semantic lecture and user models. The proposed system was designed and mainly developed by the author of this thesis under the PASCAL Harvest Project entitled *La Vie* during a three-month research stay in the Jozef Stefan Institute (JSI) at Ljubljana, Slovenia, from July to October 2012. The recommender system was finally deployed over the VideoLectures.NET site, and, at the time of writing, it is still serving recommendations to its users.

The chapter is structured as follows. First, Section 6.2 gives an overview of the recommendation system, focusing on the text extraction and information retrieval process, topic and user modelling and the recommendation process. Next, in Section 6.3 we address the dynamic update of the recommender system and the required optimisations needed to maximise the scalability of the system. Then, the integration of the proposed system into the VideoLectures.NET repository is described in detail in Section 6.4. Finally, Section 6.5 draws some conclusions and future work.

6.2 Recommendation system overview

Figure 6.1 gives an overview of the recommender system. The left-hand side of the figure shows the topic and user modelling procedure, which can be seen as the training process of the recommender system. The right-hand side describes the recommendation process. The aim of topic and user modelling is to obtain a simplified representation of each video lecture and user. The resulting representations are stored in a recommender database. This database will be exploited later in the recommendation process in order to recommend lectures to users.

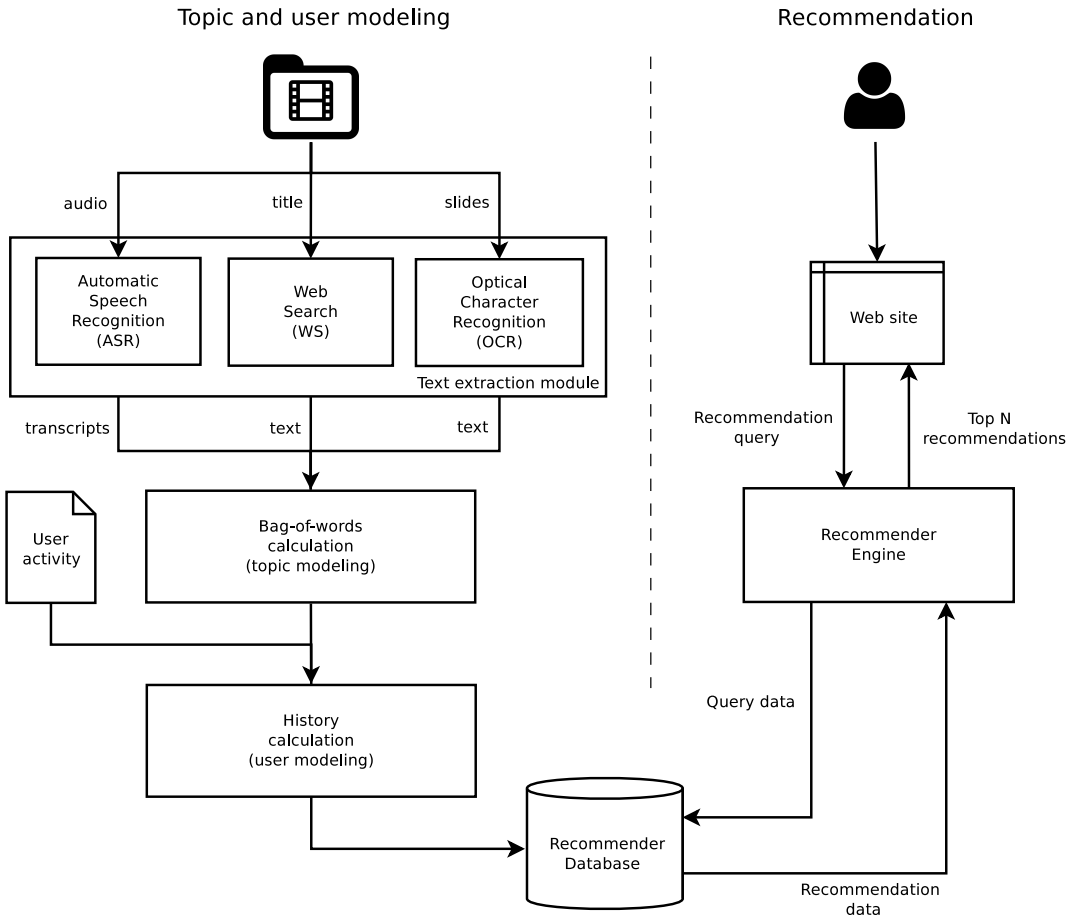

Figure 6.1: Recommender system overview.

As shown in Figure 6.1, every lecture in the repository goes through the topic and user modelling process, which involves three steps. The first step is carried out by the text extraction module. This module comprises three sub-modules: Automatic Speech Recognition (ASR), Web Search (WS) and Optical Character Recognition (OCR).

- Automatic Speech Recognition sub-module: generates an automatic speech transcription of the video lecture. However, in case the transcriptions are already available, they are automatically retrieved, for instance using the TLP Web Service (see Section 5.4.1).

- Web Search sub-module: uses the title of the lecture to retrieve related documents and publications from the web, also extracting text information from these documents.

- Optical Character Recognition sub-module: where available, it extracts text from the lecture slides when these are encoded into video or image formats. In case slides are available in a digital text format such as PDF or PPT, text information is directly extracted from them.

The second step takes the text retrieved by the text extraction module and computes a bag-of-words representation. This bag-of-words representation consists of a simplified text description commonly used in natural language processing and information retrieval. More precisely, the bag-of-words representation of a given text is a vector of its word counts over a fixed vocabulary. In the third step, lecture bags-of-words are used to represent the users of the system. That is, each user is represented as the bag-of-words computed over all the lectures the user has ever seen. Finally, user and topic models are stored in the recommender database, that is accessed afterwards by the recommender engine in the recommendation process.

The right-hand side of Figure 6.1 shows the recommendation process, that is conducted by the recommender engine. This engine uses the pre-trained topic and user models to calculate a suitability or utility function $s(u, v, r)$, being u the user, v the current lecture, and r a hypothetical lecture recommendation. Specifically, this function indicates how likely it is that a user u would want to watch lecture r after viewing lecture v. In our case, the utility function is computed as a linear combination of several recommendation features:

$$s(u, v, r) = \vec{w} \cdot \vec{x} = \sum_{i=1}^{I} w_i \cdot x_i \tag{6.1}$$

where x_i is the i-th feature vector computed for the triplet (u, v, r), w_i is the i-th feature weight and I is the number of recommendation features. In this work we considered the following recommendation features:

1. *Lecture popularity:* number of visits to lecture r.
2. *Content similarity:* weighted dot product between the lecture bags-of-words v and r [6].
3. *Category similarity:* number of categories (from a predefined set) that v and r have in common.
4. *User content similarity:* weighted dot product between the bags-of-words u and r.
5. *User category similarity:* number of categories in common between lecture r and all the categories of lectures the user u has watched in the past.
6. *Co-visits:* number of times lectures v and r have been seen in the same browsing session.
7. *User similarity:* number of different users that have seen both v and r.

Feature weights w_i can be learned by training different statistical classification models, such as support vector machines (SVMs), using positive and negative (u, v, r) recommendation samples as training data.

The most suitable recommendation $\hat{r}$ for a given u and v is computed as follows:

$$\hat{r} = \underset{r}{\operatorname{argmax}}\, s(u, v, r) \tag{6.2}$$

However, in recommender systems the most common practice is to provide the user the N recommendations r that achieve the highest utility values s, for instance, the first 10 lectures.

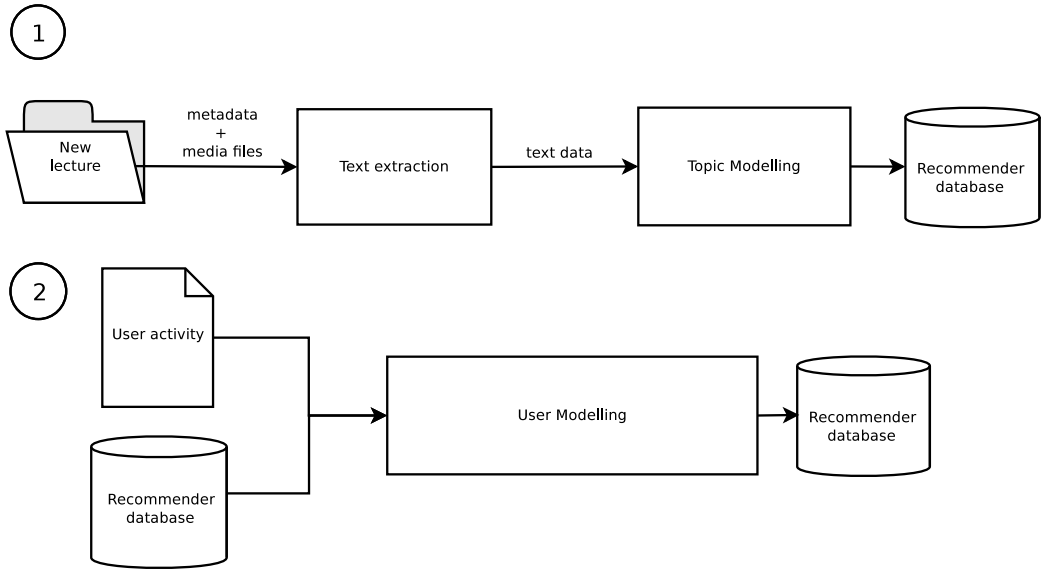

Figure 6.2: Regular update process overview.

6.3 System Updates and Optimisation

Lecture repositories are rarely static. They may grow on a daily basis to include new lectures, or have outdated videos removed. Also, users' learning progress or interactions with the repository influence the user models. The recommender database must therefore be constantly updated in order to include the new lectures added to the repository and update the user models. Furthermore, the addition of new lectures to the system might lead to changes to the bag-of-words (fixed) vocabulary. Any variation to this vocabulary involves a complete regeneration of the recommender database. That said, changes to the vocabulary may not be significant until a substantial percentage of new lectures has been added to the repository.

Two different update scenarios can be defined: on the one hand, the incorporation of new lectures and updating the user models, and on the other hand, the redefinition of the global bag-of-words vocabulary, including the regeneration of both the lecture and user bags-of-words. We will refer to these scenarios as regular update and occasional update, respectively, after the different periodicities with which they are meant to be run.

- *Regular update*: The regular update is responsible for including the new lectures added to the repository and updating the user models with the last user activity, both in the recommender database. As its name suggests, this process is meant to be run on a daily basis, depending on the frequency with which new lectures are added to the repository, since new lectures cannot be recommended until they have been processed and included into the recommender database. Figure 6.2 illustrates the *regular update* process.

- *Occasional update*: As mentioned in Section 6.2, lecture bags-of-words are calculated under a fixed vocabulary. Since there is no vocabulary restriction on the text extraction process, we need to modify the bag-of-words vocabulary as new lectures are added to

the system. The occasional update carries out the process of updating this vocabulary, which involves recalculating both the lecture and user bags-of-words.

In order to maximise the scalability of the system, while also reducing the response time of the recommender, the features *Content similarity*, *Category similarity*, *Co-visits* and *User similarity* described in Section 6.2 are pre-computed for every possible lecture pair and stored in the recommender database. Then, during the recommendation process, the recommender engine loads the values of these features, leaving the computation of features *User content similarity* and *User category similarity* until runtime. The decision to calculate the features *User content similarity* and *User category similarity* on-the-fly was driven by the highly dynamic nature of the user models, in contrast to the lecture models, which remain constant until the bag-of-words vocabulary is changed.

6.4 Integration into VideoLectures.NET

The proposed recommendation system was implemented and integrated into the VideoLectures.NET repository during the PASCAL2 Harvest Project *La Vie* (*Learning Adapted Video Information Enhancer*) [13]. Said integration is discussed here across five subsections. First, we describe the LaVie project and the VideoLectures.NET repository in Sections 6.4.1 and 6.4.2 respectively. Next, we address topic and user modelling from video lecture transcriptions and other text resources in Section 6.4.4. Then, Section 6.4.5 we describe how recommender feature weights were learned from data collected from the existing VideoLectures.NET recommender system. Finally, the recommendation system is evaluated in Section 6.4.6.

6.4.1 The LaVie project

One problem created by the success of *VideoLectures.NET* was the difficulty that individual users had in identifying the best video for their needs among the vast range of possibilities afforded by the site. Each video has a particular mix of content and style of presentation with implicit assumptions about background knowledge and level of expertise of its intended audience. On the other hand the video consumer has an approximate understanding of his/her abilities and material that he/she would like to learn about. For example, they may have a background in basic classification methods (e.g. SVMs) applied to text, some knowledge of probability theory, but not know about Bayesian reasoning. He/she would like to learn about Topic Models. The question of which sequence of videos would be most appropriate to help him/her to attain the desired knowledge would also depend on the style of presentation he/she prefers and so on. Before LaVie, *VideoLectures.NET* provided a recommender system that was only based on keywords extracted from the lecture titles. The relation would typically be based on the topic of the video or the lecturer. Furthermore, the system was not able to adapt its responses to the interests or background of the user.

The aim of La Vie project was to develop a proof-of-concept system that would provide users with advice on suitable videos for their needs. It was hosted by the Jozef Stefan Institute (JSI), at Ljubljana, Slovenia, from July to October 2012. Although the LaVie project was expected to involve a large team, finally only three people participated on it: Matjaz

Table 6.1: Basic statistics on the VideoLectures.NET repository (June 2014)

Number of videos	18,824
Total number of authors	12,252
Total duration (in hours)	11,608
Average lecture duration (in minutes)	37

Rihtar, a local JSI researcher that provided the required infrastructure and access to VideoLectures.NET data; Alejandro Pérez, researcher from the UPV; and the author of this thesis.

6.4.2 The VideoLectures.NET Repository

VideoLectures.NET [17] is a free and open access repository of video lectures mostly filmed by people from the Jozef Stefan Institute (JSI) at major conferences, summer schools, workshops and other events from many fields of science. It collects high quality educational content, recorded to high quality, homogeneous standards. The portal is aimed at promoting science, the exchange ideas and knowledge sharing by providing high quality didactic contents not only for the scientific community, but also the general public. VideoLectures.NET has so far published more than 18,000 educational videos. Relevant details regarding the repository can be found in Table 6.1.

The generation of accurate speech transcriptions for the VideoLectures.NET repository was carried out as part of the transLectures project (see Section 5.3). The recommender system was able to access the transcriptions via the TLP Web Service (see Section 5.4.1), and therefore, the ASR sub-module from the text extraction module described in Section 6.2 was replaced by the proper Web Service API calls.

6.4.3 VideoLectures.NET user-lecture interaction analysis

Before the development of the recommendation system for VideoLectures.NET, we performed an exhaustive analysis of the VideoLectures.NET usage, in order to identify user behaviour patterns, product of a user-lecture interaction, that could be exploited or taken into account in its development. In particular, we carried out two different studies. On the one hand, we performed a lecture popularity analysis, that is, how evolves the number of visits of an specific lecture over time. On the other hand, we analysed the user activity to determine how an specific user interacts with the website. Both analysis provided us answers to questions like: Does the video popularity depend on its antiquity? Does this apply to all kind of videos? Do registered users come back to VideoLectures.NET?

Both analysis were performed extracting the data available in the VideoLectures.NET website logs for a 46 months period (from September 2008 to July 2012).

Please note that all plots shown below, which exhibits the evolution of a concrete type of event by the time, presents noisy data due to the “sparse activity” of the users, and, as the time dimension increases, this effect becomes more drastic. To avoid this and to provide “eye-friendly” results, we smoothed these curves with Bezier curves. These smoothed curves are shown together with the original noisy data. Also note that the y-axis is plotted in log-scale.

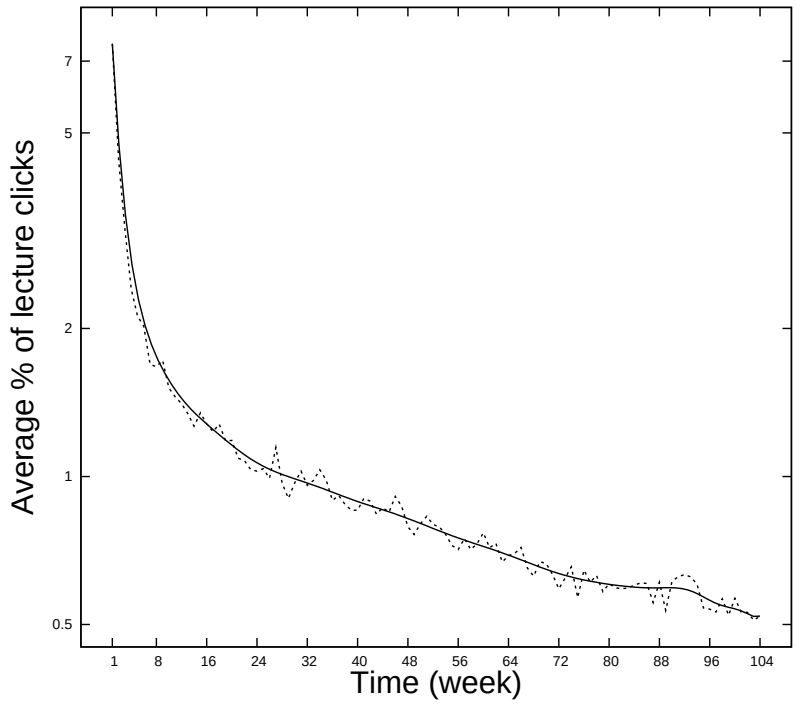

Figure 6.3: Evolution of the average percentage of total lecture clicks, computed over all videos, as a function of time (in weeks).

Lecture popularity analysis

The lecture popularity analysis was devoted to check the evolution of the number of visits over time. Figure 6.3 shows this evolution, for a period of two years, of the average percentage of total lecture clicks for all videos. This figure shows that on average, about 7% of the total lecture clicks are made on the first week after the date of publication of the lecture. Then, lecture popularity suffers an exponential decay during the first month, and afterwards, it decreases practically linearly with time.

However, depending on the lecture type, the evolution of popularity over time shows some singularities. For example, Figure 6.4 shows this evolution for two different videos extracted from two different conferences. These kind of videos seems to follow the global trend, but, surprisingly, we noticed that some light peaks of popularity come up approximately after one and two years of their publication dates. This fact might be attributed to the celebration of the same conference in the following years, which could lead some users to review the lectures from past editions of that conference.

Another example are tutorials. Figure 6.5 shows the evolution of the popularity for two different tutorials. In both cases, popularity remains roughly constant over the time. This seems reasonable, since tutorials are very requested by users with independence of their antiquity.

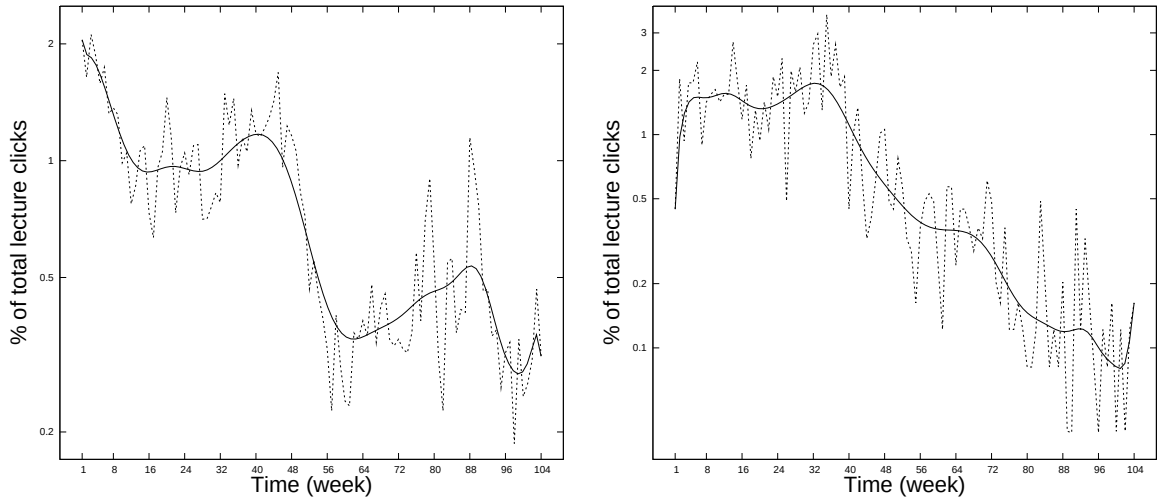

Figure 6.4: Evolution of the percentage of total lecture clicks as a function of time (in weeks) for two lectures that belong to different conferences: on the left hand, *wsdm09_dean_cblirs* (WSDM'09), and on the right hand, *www09_auer_tlwldp* (WWW'09).

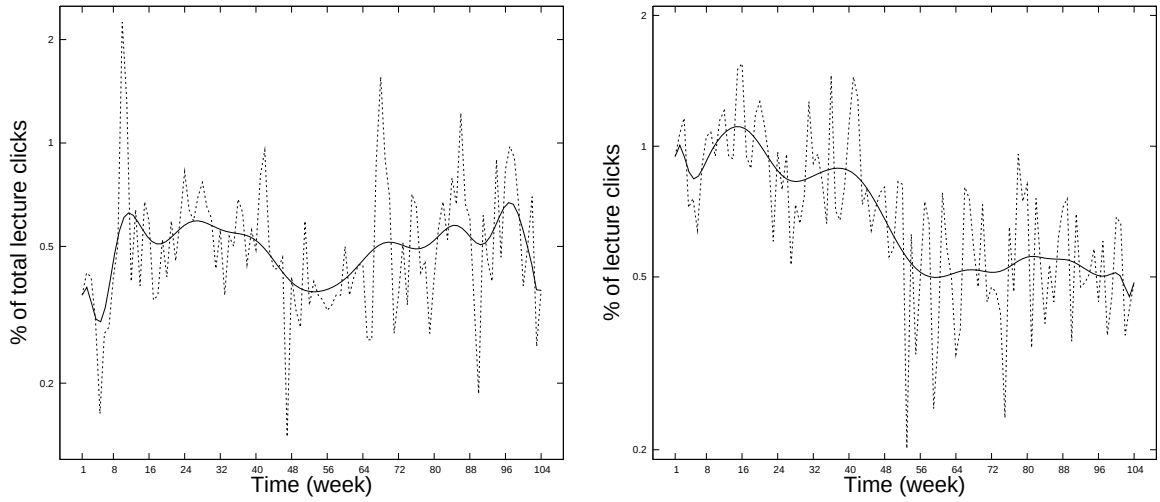

Figure 6.5: Evolution of the percentage of total lecture clicks as a function of time (in weeks), for two lectures from different tutorials: on the left hand, *ssll09_gore_iml*, and on the right hand, *ssll09_tiu_intlo*.

User activity analysis

In this section we perform an analysis of the activity of both types of VideoLectures.Net users: registered and anonymous users. We considered two types of interactions:

- Click event: A user reaches one lecture's page either from the VideoLectures.Net site or from the outside, regardless the user plays the video or not.

Table 6.2: Global statistics computed for all VideoLectures.Net registered users (3545 users).

	Mean
Connection days	2.4 ($\pm$ 4.6)
Total lecture clicks	11.4 ($\pm$ 35.1)
Total lecture views	4.6 ($\pm$ 12.7)
Lecture clicks per day	4.0 ($\pm$ 5.6)
Lecture views per day	1.7 ($\pm$ 2.6)
Ratio views/clicks	0.4 ($\pm$ 0.4)

- View event: A user plays a video lectures for at least 15 seconds.

We gathered the counts of both events for each user and computed the following statistics:

- Connection days: Number of distinct days in which a user logged activity in the site.
- Total lecture clicks: Total number of click events performed by a user.
- Total lecture views: Total number of view events performed by a user.
- Lecture clicks per day: Average number of click events performed by a user per connection day.
- Lecture views per day: Average number of view events performed by a user per connection day.
- Ratio views/clicks: Ratio for a specific user of total lecture views divided by total lecture clicks. It reveals if users actually play or not the lectures they visit.

In addition, we studied the user activity over time, in order to assess whether the average user visits regularly the website or not.

All these figures are shown separately for each user type: registered and anonymous. Identifying anonymous users was problematic, since it is not possible to match all anonymous interactions to its real, physical user. In our case, we assumed that each distinct cookie session belonged to a distinct user.

Registered users

Table 6.2 shows some statistics computed over the activity logged by registered users during the time scope of the analysis (3545 users). These figures reveals that registered users do not usually come back to VideoLectures.NET (2.4 connection days on average). Also, only in the 40% of cases the user actually plays and watches the video after visiting the lecture page. This suggests that most lectures that the site recommends to visit do not fit with the interests of the user.

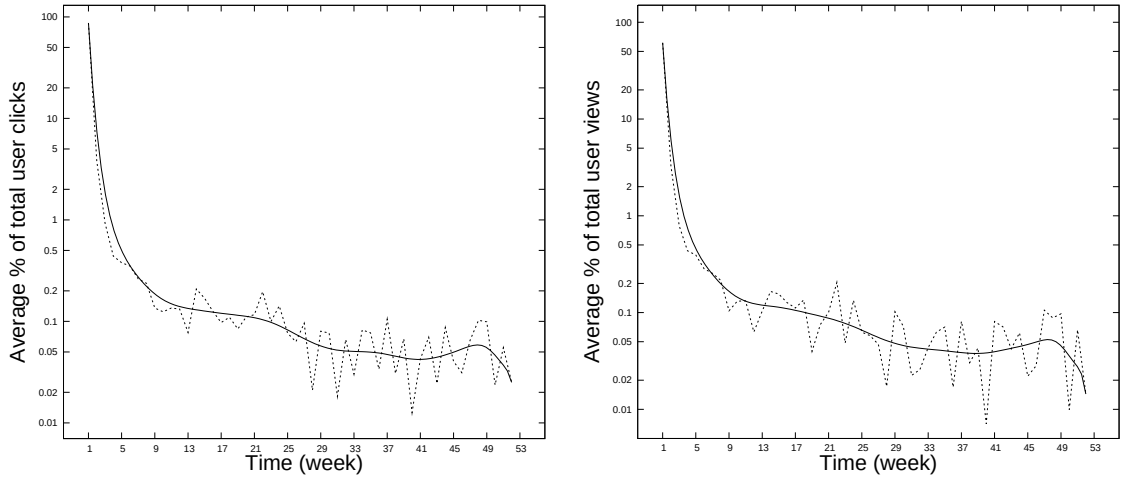

Figure 6.6: Evolution of the average percentage of total clicks, computed over all registered users, as a function of time (in weeks) is shown in the left side, whilst the same data but for the average percentage of total views is shown on the right side.

Table 6.3: Global statistics computed for all VideoLectures.Net anonymous users (3.7 millions of distinct users based on cookies).

	Mean
Connection days	1.3 ($\pm$ 1.0)
Total lecture clicks	4.4 ($\pm$ 334.9)
Total lecture views	1.7 ($\pm$ 4.1)
Lecture clicks per day	2.4 ($\pm$ 64.1)
Lecture views per day	1.2 ($\pm$ 1.6)
Ratio views/clicks	0.3 ($\pm$ 0.4)

Figure 6.6 shows, on the left hand, the average percentage of total clicks made by registered users as a function of time (in weeks), for a period of two years, and on the right hand, it shows the same figures but for the average percentage of total views. User activity decays exponentially on the first three weeks, afterwards only a marginal activity is logged. It is surprising that, on average, the 87% of the total clicks made by a user on VideoLectures.NET are performed within his/her first week after the registration. One possible explanation to this phenomena is that VideoLectures.NET doesn't offer any benefits to registered users in comparison to anonymous users, so that they finally decide to not log in again anymore.

Anonymous users

Table 6.3 shows user activity statistics computed over the activity logged by anonymous users during the time scope of the analysis (3.7 millions of users). From these figures we can draw similar conclusions as to the registered users. However, in this case we found strange values, such as anonymous users that performed hundreds of thousands of clicks. We believe

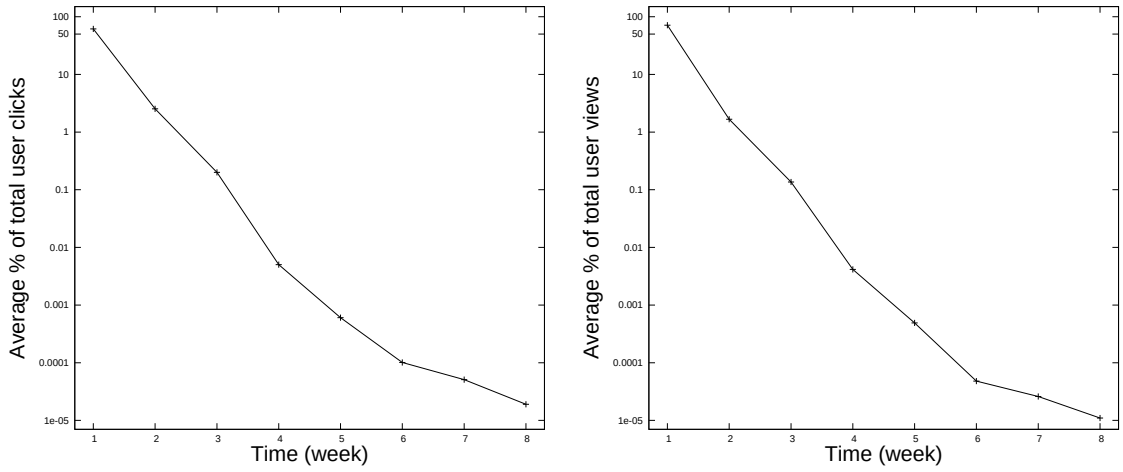

Figure 6.7: Evolution of the average percentage of total clicks, computed over all anonymous users, as a function of time (in weeks) is shown in the left side, whilst the same data but for the average percentage of total views is shown on the right side.

that these users were actually bots that were indexing VideoLectures.NET contents into their databases.

Figure 6.7 shows, on the right-most hand, the average percentage of total clicks made by anonymous users as a function of time (in weeks), for a period of two months, and on the left-most hand, the same figures but for the average percentage of total views. Anonymous users behave in a similar way like registered users, so the same conclusions can be applied here: anonymous users also do not usually come back to VideoLectures.NET after their first visit.

Conclusions

The analysis of the user-lecture interaction on VideoLectures.NET revealed some interesting facts. First of all, as we expected, lecture popularity rapidly decreases as the time goes by. However, it depends on the type of the lecture. For instance, popularity of tutorials remains more or less constant over time. Secondly, it seems that VideoLectures.NET did not provide real benefits to registered users, since most of them did not log in again after their first week on the platform. Even worse: most anonymous users (90%) did not come back to VideoLectures.NET after their first week in the site, being most of them one-day users. This means that VideoLectures.NET did not offer any mechanisms to catch the attention of casual users. Also, the existing recommender system seemed to provide bad recommendation links, given that most users actually did not play lecture videos after visiting the lecture page.

6.4.4 Topic and User Modelling

The first step in generating lecture and user models for VideoLectures.NET involved collecting textual information from different sources. In particular, the text extraction module gathered textual information from the following sources:

- Speech transcriptions from the TLP Web Service (see Section 5.4.1).
- Web search-based textual information from Wikipedia, DBLP and Google.
- Text extracted from lecture presentation slides (PPT, PDF or PNG using OCR).
- VideoLectures.NET internal database metadata.

Next, the text extraction module output was used to generate lecture bags-of-words for every lecture in the repository. These bags-of-words, as mentioned in Section 6.2, were calculated under a fixed vocabulary that was obtained by applying a threshold to the number of different lectures in which a word must appear in order to be included. By means of this threshold, vocabulary size is significantly reduced, since uncommon and/or very specific words are disregarded. Once defined, term weights were calculated using term frequency-inverse document frequency (td-idf), a statistical weighting scheme commonly used in information retrieval and text mining [9]. Specifically, tf-idf weights are used to calculate the features *Content similarity* and *User content similarity*.

Finally, the VideoLectures.NET user activity log was parsed in order to obtain values for the *Co-visits* feature for all possible lecture pairs, as well as a list of lectures viewed per user. This list was used together with the lecture bags-of-words to generate the user bags-of-words and categories. These, in turn, were used to calculate *User content similarity* and *User category similarity*, respectively, as well as *User similarity* for all possible lecture pairs.

In a final step, all this data was stored in the recommender database in order to be exploited by the recommender engine in the recommendation process.

6.4.5 Learning Recommendation Feature Weights

Once the data needed to compute recommendation feature values for every possible (u, v, r) triplet in the repository was made available, the next step was to learn the optimum feature weights for the calculation of the utility function shown in Equation 6.1. To this end, an SVM classifier was trained using data collected from the existing VideoLectures.NET naive recommender system. Specifically, every time a user clicked on any of the 10 recommendation links provided by this recommender system, 1 positive and 9 negative samples were registered. SVM training was performed using the SVM^{light} open-source software [7]. The optimum feature weights were those that obtained the minimum classification error over the recommendation data.

The feature weights that granted the minimum classification error for this data are shown in Table 6.4. As feature values are not normalised, feature weights cannot be directly compared. Nevertheless, the negative (almost zero) contribution of the lecture popularity to the recommendation score is remarkable.

6.4.6 Evaluation

Although there are many different approaches to the evaluation of recommender systems [5, 14], it is difficult to state any firm conclusions regarding the quality of the recommendations made until they are deployed in a real-life setting. The La Vie project therefore provided an

Table 6.4: Optimum recommender feature weights values.

Feature	Weight
Lecture popularity	-2.66089e-05
Content similarity	0.00452
Category similarity	0.00148
User content similarity	0.02724
User category similarity	0.04167
Co-visits	0.00187
User similarity	0.01519

ideal evaluation framework, being deployed across the official VideoLectures.NET site. The strategy followed for the objective evaluation of the La Vie recommender was to compare it against the existing VideoLectures.NET recommender by means of a coin-flipping approach. Specifically, this approach consisted of logging user clicks on recommendation links provided by both systems on a 50/50 basis and comparing the total number of clicks recorded for each system.

Table 6.5 shows the results computed using this evaluation technique. The number of clicks made on each recommender system were gathered after processing the VideoLectures.NET access logs for a time span of 171 days. Results show that the old, naive recommender system was slightly more used (clicked) than our proposed system. However, we believe that a simple comparison of user-click counts is not a legitimate point of comparison for recommendation quality. For instance, nuisance variables not taken into account might influence how users respond to the recommendation links provided. As an alternative, we can compare the rank of the recommendations clicked by users within each system. Specifically, for each recommendation clicked by a user in either system, we can compare how the same recommendation link ranked in the other system. This might be a more appropriate measure for comparing both recommender systems. However, additional data need to be logged in order to carry out this alternative evaluation. This data is currently being collected and future evaluation results will be obtained following this rank comparison approach.

Apart from this, we believe that the low quality of the English transcriptions used to train the recommender for the first time could have severely compromised the quality of the recommendations. The English ASR system that generated automatic transcriptions for the first time in VideoLectures.NET as part of the transLectures project scored 44.0 WER points in the test set of the English VideoLectures.NET corpus [15], very far from the 20% WER threshold under which ASR output becomes useful for users [10].

Despite the lack of a robust objective evidence for assessing the comparative performance of the La Vie system, informal subjective evaluations indicated that the proposed recommender system seems to provide better recommendations than the existing VideoLectures.NET recommender. Fig. 6.8 shows recommendation examples from both systems for a new user viewing a random VideoLectures.NET lecture.

Table 6.5: Coin-flipping technique evaluation results

	Clicks
Naive	33.2K (56.4%)
LaVie	25.6K (43.6%)
TOTAL	58.8K (100%)

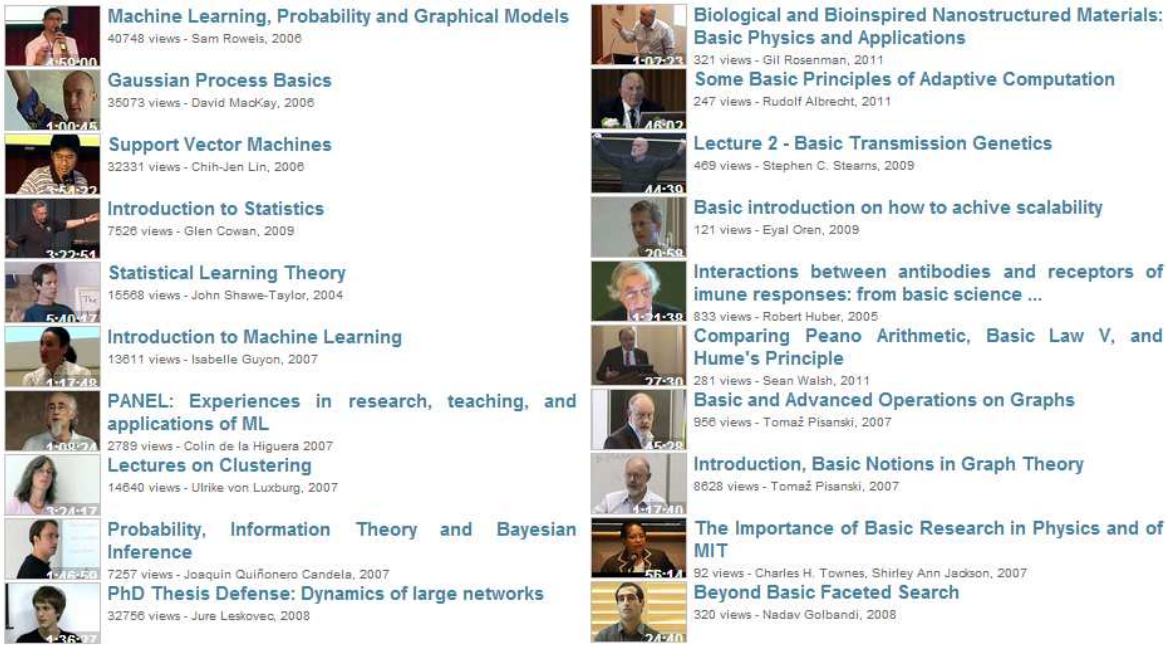

Figure 6.8: On the left, La Vie system recommendations for a new user after viewing "Basics of probability and statistics" VideoLectures.NET lecture. On the right, recommendations offered by VideoLectures.NET's existing system.

6.5 Conclusions

In this chapter we have demonstrated how automatic speech transcriptions of video lectures can be exploited to develop a lecture recommender system that can zoom in on user interests at a semantic level. In addition, we have described how the proposed recommender system has been particularly implemented for the VideoLectures.NET repository. This implementation was later deployed in the official VideoLectures.NET site and, at the time of writing, it is still generating recommendation links for the production version of VideoLectures.NET. The proposed system could also be extended for deployment across more general video repositories, provided that video contents are well represented in the data obtained by the text extraction module.

By way of future work, first, we intend to retrain the recommender system using lecture transcripts of significant better quality. The last English ASR System used in transLectures to generate automatic transcriptions in VideoLectures.NET achieved 23.4 WER points in the test set of the English VideoLectures.NET corpus [16], that is, almost a 50% WER reduction

in comparison with the system that generated the initial transcriptions used to train the current RS. This will give us an idea of the importance of the speech transcription with respect to other variables regarding recommendations quality. Our hypothesis regarding this is that the better overall transcription quality, the better recommendations can be delivered. Second, we plan to evaluate the recommender system using other evaluation approaches that measure the suitability of the recommendations more accurately, such as the aforementioned recommendation rank comparison. Finally, since lectures can be subtitled in several languages, we would like to extend this RS in order to also provide recommendations of related lectures in other languages.

Bibliography

[1] N. Antulov-Fantulin, M. Bošnjak, M. Znidaršic, and M. Grcar. ECML-PKDD 2011 discovery challenge overview. *Discovery Challenge*, 2011.

[2] W. Carrer-Neto, M. L. Hernández-Alcaraz, R. Valencia-García, and F. García-Sánchez. Social knowledge-based recommender system. application to the movies domain. *Expert Systems with Applications*, 39(12):10990–11000, 2012.

[3] J. J. Castro-Schez, R. Miguel, D. Vallejo, and L. M. López-López. A highly adaptive recommender system based on fuzzy logic for b2c e-commerce portals. *Expert Systems with Applications*, 38(3):2441–2454, 2011.

[4] A. Fujii, K. Itou, and T. Ishikawa. Lodem: A system for on-demand video lectures. *Speech Communication*, 48(5):516 – 531, 2006.

[5] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[6] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, DTIC Document, 1996.

[7] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA, 1999.

[8] S. K. Lee, Y. H. Cho, and S. H. Kim. Collaborative filtering with ordinal scale-based implicit ratings for mobile music recommendations. *Information Sciences*, 180(11):2142–2155, 2010.

[9] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[10] C. Munteanu, R. Baecker, G. Penn, E. Toms, and D. James. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pages 493–502, 2006.

[11] A. Nanopoulos, D. Rafailidis, P. Symeonidis, and Y. Manolopoulos. Musicbox: Personalized music recommendation based on cubic analysis of social tags. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(2):407–412, 2010.

[12] E. R. Núñez-Valdéz, J. M. Cueva Lovelle, O. Sanjuán Martínez, V. García-Díaz, P. Ordoñez de Pablos, and C. E. Montenegro Marín. Implicit feedback techniques on recommender systems applied to electronic books. *Computers in Human Behavior*, 28(4):1186–1193, 2012.

[13] PASCAL Harvest Programme. `http://www.pascal-network.org/?q=node/19`.

[14] G. Shani and A. Gunawardana. Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer, 2011.

[15] UPVLC, XEROX, JSI, RWTH, EML, and DDS. D2.2: Status report on complete transcriptions and translations. Technical report, The transLectures project.

[16] UPVLC, XEROX, JSI, RWTH, EML, and DDS. D2.3: Final report on complete transcriptions and translations. Technical report, The transLectures project.

[17] VideoLectures.NET: Exchange ideas and share knowledge. `http://www.videolectures.net`.

[18] M. Wald. Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. *Interactive Technology and Smart Education*, 3(2):131–141, 2006.

[19] P. Winoto and T. Y. Tang. The role of user mood in movie recommendations. *Expert Systems with Applications*, 37(8):6086–6092, 2010.

CHAPTER 7

LANGUAGE MODEL ADAPTATION USING EXTERNAL RESOURCES FOR SPEECH RECOGNITION

Contents

7.1 Introduction

Although state-of-the-art ASR systems have been proved to yield accurate speech transcriptions in most cases, their outputs can be greatly improved through the use of in-domain data. This opportunity can be exploited when transcribing large video lecture repositories, because media files are typically accompanied by descriptive metadata such as title, keywords, abstracts, or related text documents. This is the case of our first Spanish ASR system, described in Section 5.4.4. The system is adapted via language model interpolation from different out-domain and in-domain text resources including the text of the slides attached to the video lecture, significantly improving the transcription quality.

However, it is not always possible to obtain slides from video lectures: in some cases the author does not grant access to that document, in others the repository simply does not keep track of such files. Also, when slides are not available in a separated document, they can be extracted directly from the video recording applying Optical Character Recognition (OCR), although recognition errors introduced by this technology ruin almost three fourths of the potential benefits of this adaptation technique, as we saw in Section 5.5.1.

Besides slide adaptation, related works have explored language model adaptation by using text documents downloaded from the Internet. To gather relevant in-domain data, document retrieval techniques based on building search queries to locate documents through common search engines are typically applied. In the particular case of ASR, some authors tried to build these queries from keyword detection [10] or from a first pass recognition [2, 7, 11]. Other works tried to use the training set itself [14] or the text extracted from the slides [9] to build the query.

The aforementioned studies clearly focused they effort on how to build the optimal queries in order to have a competitive recall and precision trade-off. In contrast, our proposal is to use the title of each video lecture as the search query, since in almost all cases they describe with high precision the topic and the contents of the media. Our proposed document retrieval technique was inspired in the one briefly described in Section 6.2.

To sum up, in this work we propose a language model adaptation technique by document retrieval for video lecture transcription. This approach is compared with a baseline computed from a large collection of out-domain and in-domain resources. Also, we compare our results with those obtained by sole slide adaptation [8] using as slides the text extracted via OCR directly from the video file. Finally, both approaches are combined to check whether further improvements can be obtained with respect to both the baseline model and the slide-adapted model. All comparisons are made with two different acoustic modelling approaches: a classical GMM-HMM and a state-of-the-art DNN-HMM, in order to probe that transcription quality improvements are consistent with increasingly better acoustic models. All these techniques were developed within the scope of the transLectures project (see Section 5.3).

The rest of the chapter is structured as follows. First, Sections 7.2 and 7.3 describe the document retrieval and language model adaptation techniques. Then, these techniques are assessed in Section 7.4. Finally, conclusions and future work are drawn in Section 7.5.

7.2 Document Retrieval

In this work we focus on document retrieval from the web by building queries using the title of the video lecture. This is in contrast to other works where more complex techniques are proposed, such as rendering the lines of each slide as queries [9], or extracting keywords from a first pass recognition to build queries [7]. Our method is built on the hypothesis that the title is very informative and tends to contain the most important keywords and they appear in the proper order. The proposed method has several advantages among other works, such as it can be used without the need of slides or a first pass recognition on the video. We should remark that sometimes the paper, lecture notes, or even the transcript, on which the lecture is based, is downloaded. This is very useful for the adaptation, although finding this exact document is not the primary goal of the search.

To ensure that the documents retrieved are of a minimum-required quality, we constrain the search to PDF documents only, and not web pages. Note that typically PDF files are of a higher standard since they are usually papers, books or notes related to the lecture topic. Also, in case of some of the retrieved documents might be in a language different from that of the video, we perform a conventional language identification/classification process based on n-gram characters [1] over the extracted text.

We propose the following two search methods for retrieving N documents per video:

- Exact search: documents that exactly match the title of the video lecture are downloaded, i.e. the title is contained within the text of the document. Sometimes the search produces less than N results. For instance, the lecture "*Applications. V-mWater: an e-Government Application for Water Rights Agreements*" produced 0 results.

- Extended search: first, an exact search is performed, and, if less than N documents are found, the search is extended with documents that partially match the title. In other words, the extended search will retrieve all the documents from the exact search plus other documents that contain some of the words of the lecture title, up to N documents.

The impact of both search methods on the transcription quality is assessed in Section 7.4.4.

7.3 Language Model Adaptation

In nearly all language modelling applications, it is common to have corpora of different size that also stem from different sources. The basic approach to language model estimation consists of merging all the different corpora into a single one and estimating a language model on all of the data. On the other hand, this incorporates several problems. For instance, when the corpora are very different in size, the resulting n-gram counts will be dominated by the large corpus only, so including (few) in-domain data will have almost no effect on the resulting language model.

Instead, individual language models are estimated on each data source separately. In a second step, the linear interpolation of the individual probability estimates is used. Let I be the number of corpora, and $p_i(w|h)$ be the estimate for a word w given the preceding history words h for the language model trained on the i-th corpus [5]. Then we have

$$p(w|h) = \sum_i \lambda_i p_i(w|h) \tag{7.1}$$

where

$$\sum_{i=1}^{I} \lambda_i = 1$$

for the probability estimate $p(w|h)$ of the combined language model. The interpolation weights are denoted by λ_i.

In this way we can also adapt a language model to a given domain: Given a small corpus of in-domain development data, the weight parameters λ_i can be tuned such that the perplexity on the development data is minimised. The weight parameters then weigh the individual corpora, be it in-domain or out-of-domain, according to their adequacy related to the in-domain development data.

Equation 7.1 is extended to consider language models trained with text extracted from documents retrieved from the internet:

$$p(w|h, V) = \sum_i \lambda_i p_i(w|h) + \lambda_D p_D(w|h) \tag{7.2}$$

where

$$\sum_{i=1}^{I} \lambda_i + \lambda_D = 1$$

being V the current video and $p_D(w|h)$ the language model trained on the documents downloaded for V.

Furthermore, we consider the scenario where lecture slides can be extracted applying OCR to the video file [8], or the scenario where both the text from the slides and the retrieved documents are combined:

$$p(w|h, V) = \sum_i \lambda_i p_i(w|h) + \lambda_D p_D(w|h) + \lambda_S p_S(w|h) \tag{7.3}$$

where

$$\sum_{i=1}^{I} \lambda_i + \lambda_D + \lambda_S = 1$$

In case the document retrieval technique does not download any resource for a given video, λ_D is constrained to 0.

7.4 Experiments

In this section our approach is compared with a baseline language model computed with both out-of-domain corpora and in-domain data. The baseline system is compared against three systems: a system adapted with documents, a system adapted with slides, and a system adapted with both slides and documents. In all experiments two acoustic models are used: a DNN-HMM and a classical GMM-HMM. Empirical results are reported on the test set of the Spanish poliMedia corpus (see Section 5.2).

Table 7.1: Main statistics of the out-of-domain corpora used to train the baseline language model.

	Sentences	Words	Vocabulary
Google Ngram	–	44.8G	2.2M
UnDoc	10M	318.0M	1.9M
news (07-11)	8.6M	217.2M	2.9M
El Periódico	2.7M	45.4M	916K
Europarl-v7	2.1M	54.9M	439K
UnitedNations	448K	10.8M	234K
TED	316K	2.3M	133K
news-commentary	183K	4.6M	174K
EPPS	132K	0.9M	27K

7.4.1 Corpora

The baseline model was trained using a large set of out-of-domain and in-domain corpora. On the one side, the out-of-domain corpora is summarised in Table 7.1 with basic statistics. On the other side, as in-domain corpus we used the poliMedia corpus (see Section 5.2). Details of this corpus are given in Table 5.2.

In order to carry out experiments with language models adapted to slides and documents, we first needed to extract the text from the slides. However, as stated in Section 5.4.4, poliMedia does not maintain a separated text version of the slides, and therefore text was extracted automatically applying OCR. To do this we used the Tesseract OCR tool [12] as in our previous experiments (see Section 5.5.1), but applying an improved preprocessing step which performs several filters such as despeckling, enhancing or pixel negation to reduce the noise generated by different factors, including size variability; irregular structure due to different objects (charts, images, tables); or background and foreground colour variation. This improved OCR technique offered a Word Error Rate (WER) of 40% on the recognition of the text from the slides of the test set, which clearly outperforms the previous OCR system (64% WER points, see Section 5.5.1). The text extracted from the slides of the poliMedia test set accounted for 16.4K words and a vocabulary size of 3.1K words.

Regarding the document retrieval technique, we carried out experiments with language models trained with the text extracted from up to a maximum of 5, 10 and 20 documents downloaded per video, for both exact and extended searches. As we will see in Section 7.4.4, the extended search reports significantly better results than the exact search for 5 documents. So for 10 and 20 documents we only considered the extended search. Details of the retrieved documents for the videos of the poliMedia test are depicted in Table 7.2.

7.4.2 Acoustic Models

The proposed language model adaptation techniques were tested with two different acoustic models: standard GMMs and DNNs. We used the TLK [4] to train both acoustic models using the training set of the poliMedia corpus (see Table 5.2).

On the one side, GMM-HMM acoustic models were based on triphonemes, modelled

Table 7.2: Global statistics of all downloaded documents for every video from the test set of the poliMedia corpus.

		Documents	Words	Vocabulary
Exact Search	(5 docs)	102	1.2M	41K
Extended Search	(5 docs)	115	1.4M	42K
	(10 docs)	230	2.7M	65K
	(20 docs)	459	6.4M	104K

with a 3-state left-to-right topology. A decision tree based state-tying was applied, resulting in a total of 5039 triphoneme states. Each triphoneme was trained for up to 128 mixture components per Gaussian and 4 iterations per mixture. Also, fCMLLR was applied in order to reduce speaker variability.

On the other side, to train the hybrid DNN-HMM acoustic model an accurate forced alignment of the input features at triphoneme state (senone) level is required. We computed this alignment using the aforementioned GMM-based system. With respect to the topology of the DNN, the size of the output layer was determined by the number of phonetic targets derived during the previous forced alignment, in this case, 5039 classes. The development set was used to determine the optimum network configuration: 4 hidden layers of 3000 neurons. Finally, Gaussian mixtures of the previous HMM acoustic model were replaced with the neural network state posterior probabilities.

7.4.3 Language Models

Concerning the language models, the baseline model was trained interpolating several individual language models, as discussed in Section 7.3, and trained on the corpora described in Section 7.4.1.

It must be noted that the language model interpolation technique is quite expensive in computational terms. Since we are looking for a cost-effective solution, instead of interpolating the in-domain corpora with all the out-of-domain corpora, a prior interpolation of out-of-domain corpora is performed and this model is interpolated with the in-domain corpora. This change had no impact on error rates. However, it led to a considerable improvement of performance in terms of both time and memory.

Therefore, we trained a 4-gram language model for each out-of-domain corpus using the SRILM [13] toolkit. In addition, the training set of the poliMedia corpus was used to train the in-domain language model. Every language model was smoothed with the modified Kneser-Ney absolute interpolation method [3, 6]. The vocabulary of the interpolated language model was computed using 200K words over all the out-of-domain corpora plus the in-domain vocabulary, resulting in a 205K words vocabulary. Regarding the adapted models, the vocabulary was built extending the base vocabulary with the words in the slides and/or the documents.

Table 7.3: WER (%) computed over the test set of the poliMedia corpus for the baseline language models and for a set of adapted language models augmented with different number of documents retrieved with either exact and extended searches.

Language Model	Acoustic Model	
	GMM	DNN
Baseline (BL)	21.8	15.7
BL + Exact (5 docs)	20.7	14.6
BL + Extended (5 docs)	20.6	14.4
BL + Extended (10 docs)	20.6	14.4
BL + Extended (20 docs)	20.0	14.2

Table 7.4: Evolution of the WER (%) when adding OCR slides and retrieved documents to the baseline language models, computed over the test set of the poliMedia corpus.

Language Model	Acoustic Model	
	GMM	DNN
Baseline	21.8	15.7
+ OCR Slides	19.4	13.8
+ Documents	18.9	13.5

7.4.4 Evaluation

First we run experiments to assess whether the exact or the extended search is better for querying documents. For these experiments the number of documents per video is set to 5. Table 7.3 depicts these results, in which it is observed that document adaptation significantly improves the baseline results independently of the AM used. The extended search obtains better results than the exact search where the smaller amount of documents retrieved leads to slightly higher WER values.

After setting the retrieval technique to extended search, we assessed the impact of the number of documents retrieved when using up to 10 and 20 documents, instead of 5. The same table shows a significant improvement for all acoustic models when using 20 documents, up to 11% relative WER.

In cases where slides are available, not only it is possible to perform adaptation by using either the documents or the slides [8], but also a combination of these two resources. These combined results are summarised in Table 7.4. It is observed that the inclusion of documents significantly improves the results of all the previous systems (adapted or not) where documents were not used. It is also interesting to note that the combination of slides and documents outperforms both the system without slides and the system without documents.

7.5 Conclusions

We have proposed a new simple yet effective method to retrieve documents from the web and use them to build adapted language models for video lecture transcription. These doc-

uments have proven to be a very valuable resource for adapting language models, obtaining relative WER improvements of up to 9% using GMM, and up to 11% when using DNN, with respect to a strong baseline. Furthermore, if we combine the document adaptation with slide adaptation the system yields relative WER improvements of 13-14% with respect to a strong baseline. It is worth noting that, in general, the improvements are consistent for all proposed acoustic models, which makes us think that this kind of adaptation will provide significant improvements as the acoustic models get even better.

The documents obtained have led to significant improvements, proving that this method is a good way of retrieving documents for the purpose of adapting language models to the topic of the lecture. However, in the future, we plan to compare this document retrieval method with the alternative methods proposed by other authors [9–11]. Furthermore, we would like to explore the possibility of also adapting language models to the vocabulary of the speaker, using supervised or unsupervised transcriptions of previous lectures from the same speaker, in order to disambiguate certain words and expressions frequently used by the lecturer.

Bibliography

[1] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

[2] P.-C. Chang and L.-S. Lee. Improved language model adaptation using existing and derived external resources. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 531–536. IEEE, 2003.

[3] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393, 1999.

[4] M. A. del Agua, A. Giménez, N. Serrano, J. Andrés-Ferrer, J. Civera, A. Sanchís, and A. Juan. The translectures-upv toolkit. In *Advances in Speech and Language Technologies for Iberian Languages - Second International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain, November 19-21, 2014. Proceedings*, pages 269–278, 2014.

[5] F. Jelinek and R. L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proc. of the Workshop on Pattern Recognition in Practice*, pages 381–397, Amsterdam, The Netherlands: North-Holland, 1980.

[6] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing, ICASSP '95, Detroit, Michigan, USA, May 08-12, 1995*, pages 181–184, 1995.

[7] G. Lecorvé, G. Gravier, and P. Sébillot. An unsupervised web-based topic language model adaptation method. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, March 30 - April 4, 2008, Caesars Palace, Las Vegas, Nevada, USA*, pages 5081–5084, 2008.

[8] A. A. Martinez-Villaronga, M. A. del Agua, J. Andrés-Ferrer, and A. Juan. Language model adaptation for video lectures transcription. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 8450–8454, 2013.

[9] C. Munteanu, G. Penn, and R. Baecker. Web-based language modelling for automatic lecture transcription. In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, pages 2353–2356, 2007.

[10] I. Rogina and T. Schaaf. Lecture and presentation tracking in an intelligent meeting room. In *4th IEEE International Conference on Multimodal Interfaces (ICMI 2002), 14-16 October 2002, Pittsburgh, PA, USA*, pages 47–52, 2002.

[11] T. Schlippe, L. Gren, N. T. Vu, and T. Schultz. Unsupervised language model adaptation for automatic speech recognition of broadcast news using web 2.0. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 2698–2702, 2013.

[12] R. Smith. An overview of the tesseract OCR engine. In *9th International Conference on Document Analysis and Recognition (ICDAR 2007), 23-26 September, Curitiba, Paraná, Brazil*, pages 629–633, 2007.

[13] A. Stolcke. SRILM - an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*, 2002.

[14] A. Tsiartas, P. G. Georgiou, and S. Narayanan. Language model adaptation using WWW documents obtained by utterance-based queries. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA*, pages 5406–5409, 2010.

CHAPTER 8

TRANSCRIPTION AND TRANSLATION PLATFORM

Contents

8.1 Introduction

In Chapter 5 we presented the *transLectures-UPV Platform* (TLP), which consists of a set of software tools that enable multilingual automatic subtitling in large video repositories. Indeed, we described the integration of this software with the poliMedia repository (see Section 5.2) under the scope of the transLectures EU project (see Section 5.3). This integration made possible the generation for the first time of automatic Spanish and English subtitles for several thousands of Spanish video lectures available in the poliMedia repository. These subtitles were produced by the first Spanish ASR and Spanish into English SMT systems described in Section 5.4.4. Also, thanks to the TLP Player (see Section 5.4.2), poliMedia users were able to edit subtitles in order to amend transcription and translation errors.

During the transLectures project, new ASR and MT systems were built and incorporated into TLP to support transcription and translation of poliMedia lectures in other languages than Spanish. Furthermore, all those ASR and MT systems were progressively improved throughout the project by introducing several technological upgrades. In this chapter we will describe chronologically all changes and improvements made in this period of time. For the sake of brevity and clarity, instead of calendar dates, we will use relative months since the start of the transLectures project to state the time instant in which an improvement was introduced into TLP. For instance, M12 will refer to the 12th month after the beginning of transLectures. As a reminder, the transLectures project started in November 2011 and finished in October 2014.

Also, during this period of time and beyond, the TLP software have been significantly improved by correcting bugs and adding new features and enhancements. These improvements resulted in successive public releases of TLP under an open-source license. Indeed, it has been adopted recently by different institutions, either physically and remotely as a cloud service via the *Transcription and Translation Platform* [a] (TTP) of the MLLP [b] research group, a cloud-based transcription and translation service fully grounded on TLP.

The rest of the chapter is structured as follows: Section 8.2 describes the latest TLP release existing at the time of writing. Next, Section 8.3 reviews the integration of TLP with the poliMedia repository within the scope of the transLectures project. Then, Section 8.4 describes the aforementioned MLLP's Transcription and Translation Platform, and finally, Section 8.5 gives some concluding remarks and future work.

8.2 The transLectures-UPV Platform

In this section we will give a complete overview of TLP, describing the workflows involved when integrating ASR, MT and Text-To-Speech Synthesis (TTS) technologies into large media repositories with the aid of this platform.

The main features of TLP are the following:

- **Easy integration of ASR, MT and TTS technologies:** TLP provides a framework in which custom ASR, MT and TTS systems can be easily integrated into the platform to

[a] `http://ttp.mllp.upv.es`
[b] `http://mllp.upv.es`

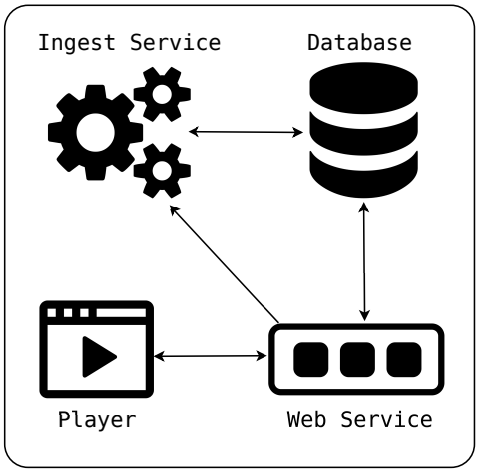

Figure 8.1: Main Components of *The transLectures UPV Platform*.

provide automatic subtitles and synthesised text-to-speech audio tracks. This feature is widely reviewed in Section 8.3.1.

- **Exploitation of media-related resources:** Language model adaptation techniques for ASR such as the one described in Chapter 7 are supported by TLP.
- **Support to grid computing**: TLP has been developed expressly to take advantage of computer clusters in order to parallelize at maximum transcription, translation, and text-to-speech processes.
- **Ergonomic solution for subtitle editing:** TLP includes a web-based software tool, the TLP Player (See Section 8.2.4), that can be used to review subtitles with ease and comfort.
- **Support to collaborative subtitling:** TLP enables collaborative subtitling by managing edit sessions and keeping track of all user editions made on subtitle files. Changes can be postponed to be reviewed by authors, who can later review these changes to accept or reject them.
- **Automated actions to enhance subtitle quality:** Whenever a transcription or translation file is edited, translations and/or synthesised audio tracks are automatically regenerated.
- **Extensive and powerful public API:** TLP offers a complete API (see Section 8.2.3) that makes very easy the integration process between TLP and remote media repositories.
- **Based on open-source tools and libraries:** All software tools included in TLP are entirely based on open-source languages and libraries.
- **Publicly released as open-source software:** Successive versions of TLP have been publicly released under the open-source Apache License 2.0.

Figure 8.1 shows the main components of TLP and a simplification of all existing interactions among them. These components are the Database, the Web Service, the Ingest Service and the Player, each of which are described in their corresponding sections.

As stated before, all incremental versions of TLP have been released under the open-source Apache License 2.0:

- **v1.0.0**: July 2014
- **v1.0.1**: October 2014 (End of the transLectures project)
- **v1.1.0**: December 2014
- **v1.2.0**: April 2015
- **v2.0.0**: June 2015
- **v2.1.0**: July 2015

The latest version of TLP can be freely downloaded from the MLLP research group's official web page[c]. The following subsections describe TLP at its version 2.1.0.

8.2.1 Use Cases

We have defined the following three use cases to illustrate the main ways a remote media repository and its users can interact with TLP:

1. A new recording from the media repository needs to be automatically transcribed and translated.
2. A user plays a video along with subtitles in the media repository's website.
3. A user decides to review video subtitles.

Use Case 1: A new recording from the media repository needs to be automatically transcribed and translated

Figure 8.2 describes graphically all human-machine interactions taken place in this first use case. A lecturer records a new video lecture, for instance, in a recording studio, classroom or conference. To get this video transcribed and translated into several languages, a Media Package File (MPF) made up with the recorded media file plus other metadata (see Section 8.2.5) is created and uploaded to TLP via the */ingest* interface of the Web Service API (see Section 8.2.3). The Ingest Service (see Section 8.2.5) unpacks the MPF and launches the required transcription, translation and/or speech synthesis processes. During this stage, the client (media repository) can check at any time the progress of the upload using the */status* endpoint of the Web Service. Finally, the Ingest Service creates a new media record in the Database (see Section 8.2.2) and stores all media, subtitles, and synthesised audio track files.

[c] `http://mllp.upv.es/tlp`

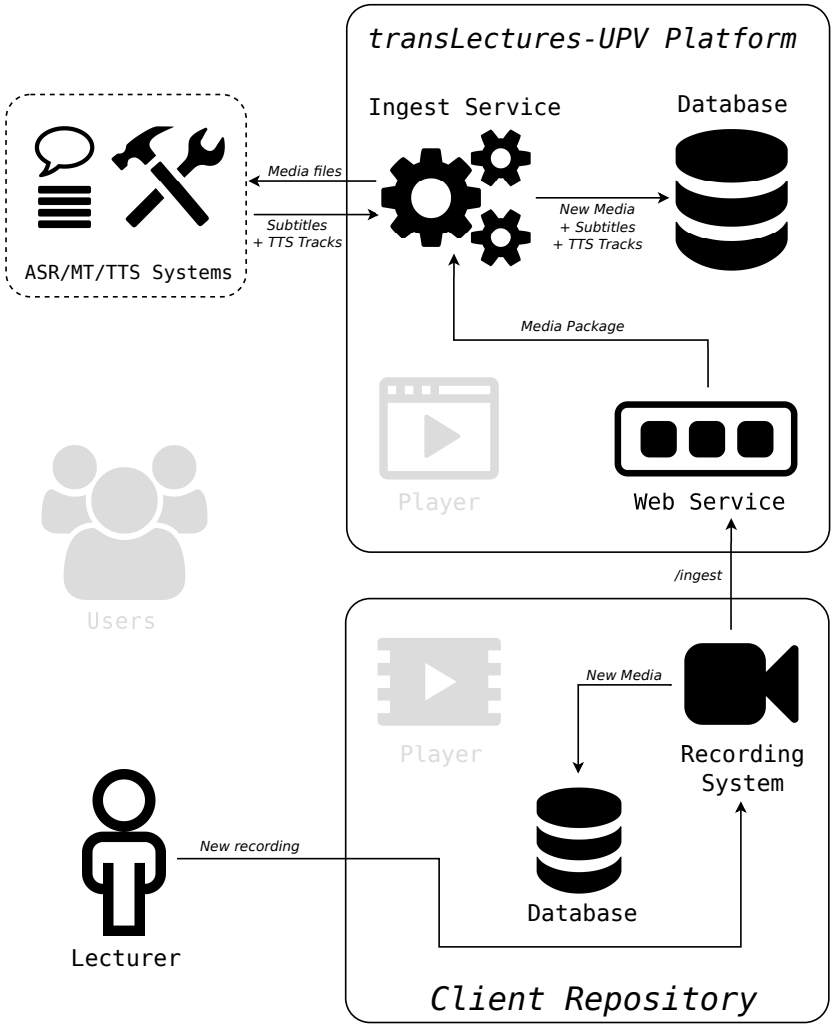

Figure 8.2: TLP Use Case 1: A new recording from the media repository needs to be automatically transcribed and translated.

It must be noted that this process is fully transparent for the lecturer, in the sense that it does not alter the educational content production life cycle: subtitles are automatically generated without human intervention after the recording session ends. Afterwards, these subtitles are delivered on-demand to the users, as described in the next use case.

Use Case 2: A user plays a video along with subtitles in the remote repository's website

Figure 8.3 depicts the second use case. Here, a user browses the remote media repository's catalogue and selects one video file. The user can watch the selected media with subtitles in different languages, or even listen to it in another language by selecting an automatically synthesised audio track. To get the list of all available subtitle languages, the repository's media player sends a request to the */langs* interface of the Web Service. As the user selects the desired subtitle language, the remote repository's media player calls the */subs* endpoint to

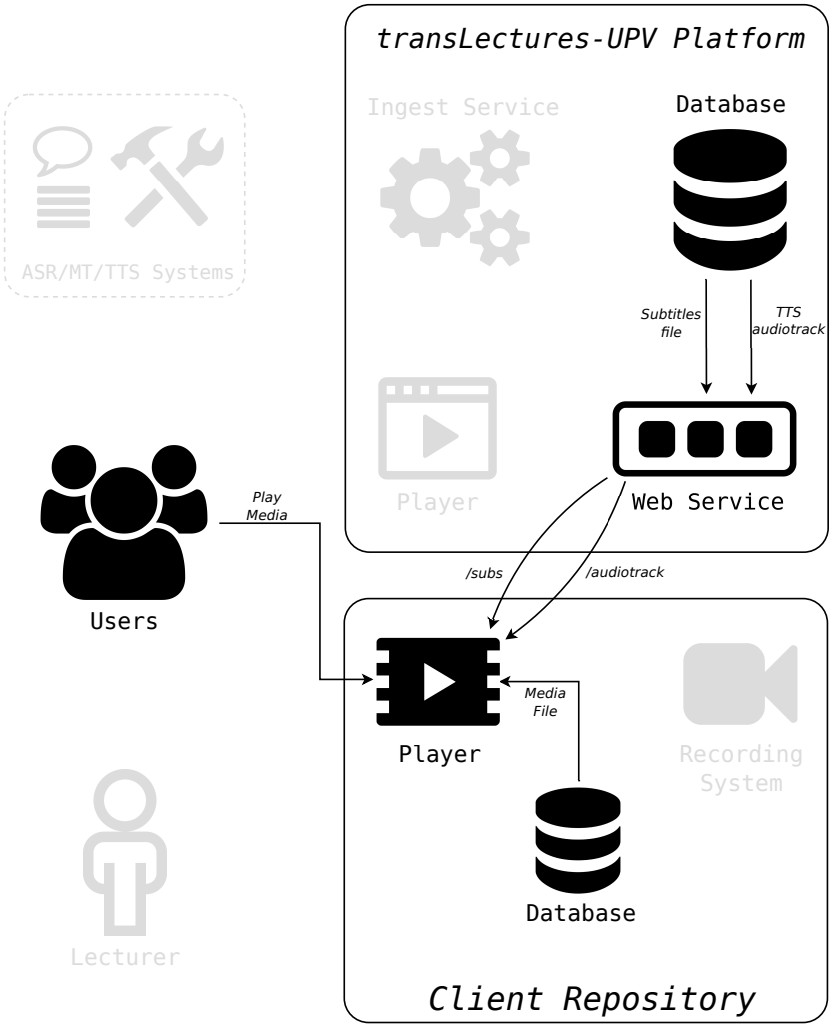

Figure 8.3: TLP Use Case 2: A user plays a video along with subtitles in the remote repository's website.

download the latest version of the corresponding subtitle file, which is immediately displayed. Similarly, the media player calls the */audiotrack* interface to retrieve a synthesised audio track in another language.

Use Case 3: A user decides to review video subtitles

An illustration of the third use case is shown in Figure 8.3. A user, while playing a media file with subtitles (use case 2), notices that the subtitles contain errors and he/she decides to correct them. To do this, the repository's media player redirects the user to the TLP Player[d] (see Section 8.2.4). The TLP Player offers an ergonomic and efficient interface for subtitle

[d]It must be noted that the remote repository's media player can allow subtitle editing, whenever it implements the needed API calls to the Web Service in order to submit all changes to TLP. In this case, the redirection to the TLP Player is not needed.

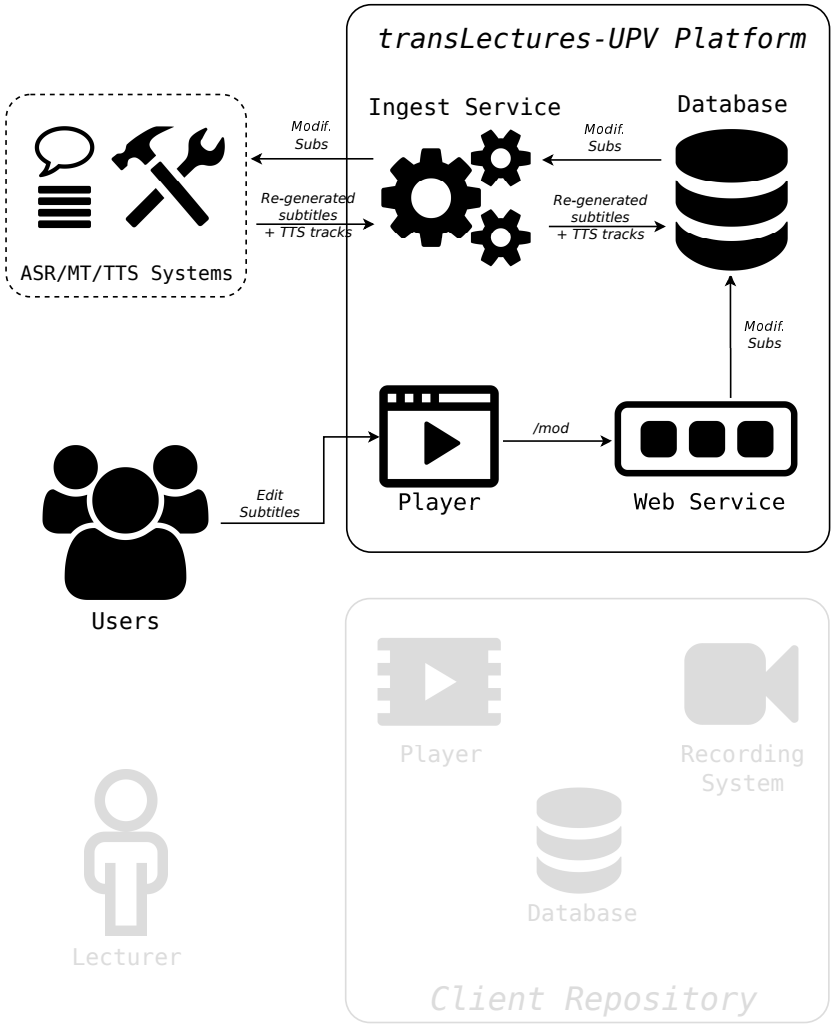

Figure 8.4: TLP Use Case 3: A user decides to review video subtitles.

editing.

The TLP Player loads the main media file and the selected subtitles file by calling the */metadata* and */subs* interfaces of the Web Service, respectively. Any corrections made by the user are then sent back to the Web Service via the */mod* interface and appended into the internal DFXP [35] representation of the subtitles file. The DFXP format was extended to store extra information such as timestamps, user information and confidence measures. The updated DFXP file plus other metadata are committed to the Database.

Finally, automatic translations and synthesised audio tracks are automatically re-generated taking into account user amendments. In parallel, WER or BLEU metrics, depending on the type of subtitles (transcription or translation, respectively), are automatically computed by taking as hypothesis the fully automatic version of the subtitles, and as reference the reviewed version from the user(s). These metrics provide useful information about the quality of the ASR or MT system that generated the subtitle file.

Alternatively, TLP can be configured in order to put lecturers in the middle of the process.

Under this setting, subtitle changes made by regular users are not committed, and therefore not shown to other users, until the lecturer reviews these changes and decides to approve or discard them. Lecturers are properly notified by the remote media repository when there exist subtitle modifications that require their approval. The Web Service is conveniently equipped with a set of interfaces to support this use case variant (i.e. */revisions*, */accept* or */reject*, among others).

8.2.2 Database

TLP features an object-relational database which stores all the data required by the Web Service and the Ingest Service. The Database manages the following entities:

- **Uploads**: Every time a new media file is uploaded using the */ingest* operation, a new upload entry is stored in the database to track its progress.
- **Systems**: Metadata from all ASR, MT, and TTS systems featured by the Ingest Service (see Section 8.2.5) are also stored into the database.
- **Media/Lectures**: All the information related to a specific media/lecture is stored in the database, including language, duration, title, keywords and category. An external ID, provided by the client repository, is used to identify the media object in all transactions performed between the client repository and the Web Service API.
- **Speakers**: Information about the speaker/lecturer can be used by the ASR system to adapt the underlying models to the unique characteristics of the given speaker and, therefore, improve the quality of the resulting subtitles.
- **Subtitles**: All subtitles automatically generated by the Ingest Service are stored in DFXP format into the database and retrieved by the client via the Web Service.
- **Audiotracks**: As in the case of subtitles, automatically synthesised audio tracks from translated subtitles are also stored in the database.
- **Edit sessions**: Information about user edit sessions, either alive or finished, is stored in the database. The Web Service uses this information to allow or disallow a user to start editing a subtitle file (see Section 8.2.3).
- **Edit history**: All user editions within an edit session are stored individually in the database, so that it is possible to recover previous versions of a subtitle file at any time. In fact, the Web Service offer several API interfaces devoted to this purpose (see Section 8.2.3).
- **Users**: Conceptually, TLP is presented as a single-user system (the client repository to which it is integrated). However, it is possible to add more users to give support to multiple repositories at the same time. The MLLP's Transcription and Translation Platform described in Section 8.4 is a good example of a multiple-user TLP installation.

8.2.3 Web Service

The Web Service is the key component for integration: it is the information exchange point between the remote media repository and TLP (see Section 8.2.4). Specifically, the Web Service defines an API composed of a wide set of HTTP interfaces that allow a full integration of all TLP services with the remote media repository. These interfaces can be divided in four groups:

- **Interfaces for media upload and management:**
  - */ingest*: allows to upload media (audio/video) files and any attachments and metadata to TLP, in order to generate automatic subtitles and/or translated audiotracks.
  - */uploadslist*: returns a list of all uploads being processed by the Ingest Service (see Section 8.2.4).
  - */status*: allows to check the current status of a specific upload ID.
  - */systems*: returns a list of all available ASR, MT, and TTS systems featured by the Ingest Service.
- **Interfaces for downloading media and subtitle files:**
  - */metadata*: returns metadata and media file locations for a given media ID.
  - */langs*: returns a list of all subtitles and audio track languages available for a given media ID.
  - */subs*: allows to download a subtitle file for a given media ID and language.
  - */audiotrack*: allows to download an audio track file for a given media ID and language.
- **Interfaces for the edit of subtitles:**
  - */start_session*: starts an edit session to send and commit modifications of a subtitles file.
  - */session_status*: returns the current status of the given session ID.
  - */mod*: allows to send subtitle corrections under an open edit session.
  - */end_session*: ends an open edit session, and depending on the confidence of the user, changes are directly stored into the corresponding subtitles files, or left for revision by the author.
- **Interfaces for the management of user edits:**
  - */lock_subs*: (for authors) allow/disallow regular users (non-authors) to send subtitles modifications for an specific Media ID.
  - */edit_history*: returns a list of all edit sessions that involved an specific media ID.
  - */revisions*: returns a list of all edit sessions of all media files of the remote repository that are pending to be revised.

- */mark_revised*: (for authors) mark/unmark as revised an specific edit session ID.
- */accept*: (for authors) accept modifications of one or more pending edit sessions without having to revise them. Modifications are committed into the corresponding subtitles files.
- */reject*: (for authors) reject modifications of one or more pending edit sessions without having to revise them.

A detailed description of the Web Service's API can be found in the on-line documentation of TLP[e]. It is worth noting that public releases of TLP come along with several tools and libraries ready to interact with this API.

8.2.4 Player

The TLP Player is an PHP/HTML5 media player that allows users to review and edit subtitles with ease. It provides a highly ergonomic editing interface optimised to reduce user effort. Several refinements and enhancements have been implemented on it based on objective data and subjective feedback gathered from the different user evaluations we carried out (see Sections 5.5.2 and 8.3.3). Figure 8.5 shows an screenshot of the current Player. Its main features are the following:

- **Four different editing layouts:** depending on where the subtitle text edit area is located: *on-screen*, *side-by-side*, *top-bottom* and *subtitles only*.
- **A single yet effective interaction method**: conventional post-editing with a complete set of keyboard shortcuts to boost expert users' capabilities. The Intelligent Interaction mode, implemented in the first version of TLP as described in Section 5.4.2, was deprecated as a consequence of the results obtained in the first user evaluations described in Section 5.5.2.
- **Two edit modes**: *basic mode*, in which only the text of the subtitles can be edited, and *advanced mode*, where also the segmentation of the audio signal can be modified, this is, subtitle segments can be added, deleted or re-sized.
- **Support to external audio tracks**: it is capable to play the translated and synthesised text-to-speech audio tracks generated by the Ingest Service (see Section 8.2.4).
- **Adaptable to user preferences**: it includes a user settings section in which some options can be adjusted according to user preferences.
- **User sessions management**: it notifies to users when subtitles are being edited by another user at the same time.
- **Edit history management**: it allows authors and non-authors to load modifications made by other users to any of the subtitles.

[e] `http.//mllp.upv.es/tlp`

Figure 8.5: Screenshot of the advanced mode of the TLP Player.

8.2.5 Ingest Service

The Ingest Service is the component devoted to coordinate the automatic transcription, translation, and audio track synthesization of either new and existing media files. Specifically, it takes as input Media Package Files uploaded via the */ingest* interface of the Web Service. A MPF is a ZIP file that contains media files and attachments, plus a JSON file that states the uploaded media files and attachments included in the MPF, in addition to other metadata. The Ingest Service checks periodically whether new MPFs have been uploaded in order to start their processing, and if the ongoing uploads are progressing correctly or have failed.

Internal Structure

Figure 8.6 shows the internal structure of the Ingest Service, which is split into two layers. On the one hand, the upper layer implements the main logic of the Ingest Service using a modular design. It has a central node, the Core, that implements the logic of all possible paths a MPF can follow, while data processing tasks are handled by external modules. This means that the functionalities of the Ingest Service can be easily modified, replaced or extended by swapping these external modules with others.

These external modules can be divided in two categories:

- **Base Modules:** Modules that implement APIs for the basic operations used by the Core.
  - **URL Downloader:** Module that allows to download media files from a given URL address, whenever URL links are sent in the MPF instead of physical media files. It also offers the possibility of downloading media files from encoded URLs such as YouTube or Vimeo using external plug-ins (URL decoders).
  - **Media Module:** Module that offers several methods to perform media format conversions.

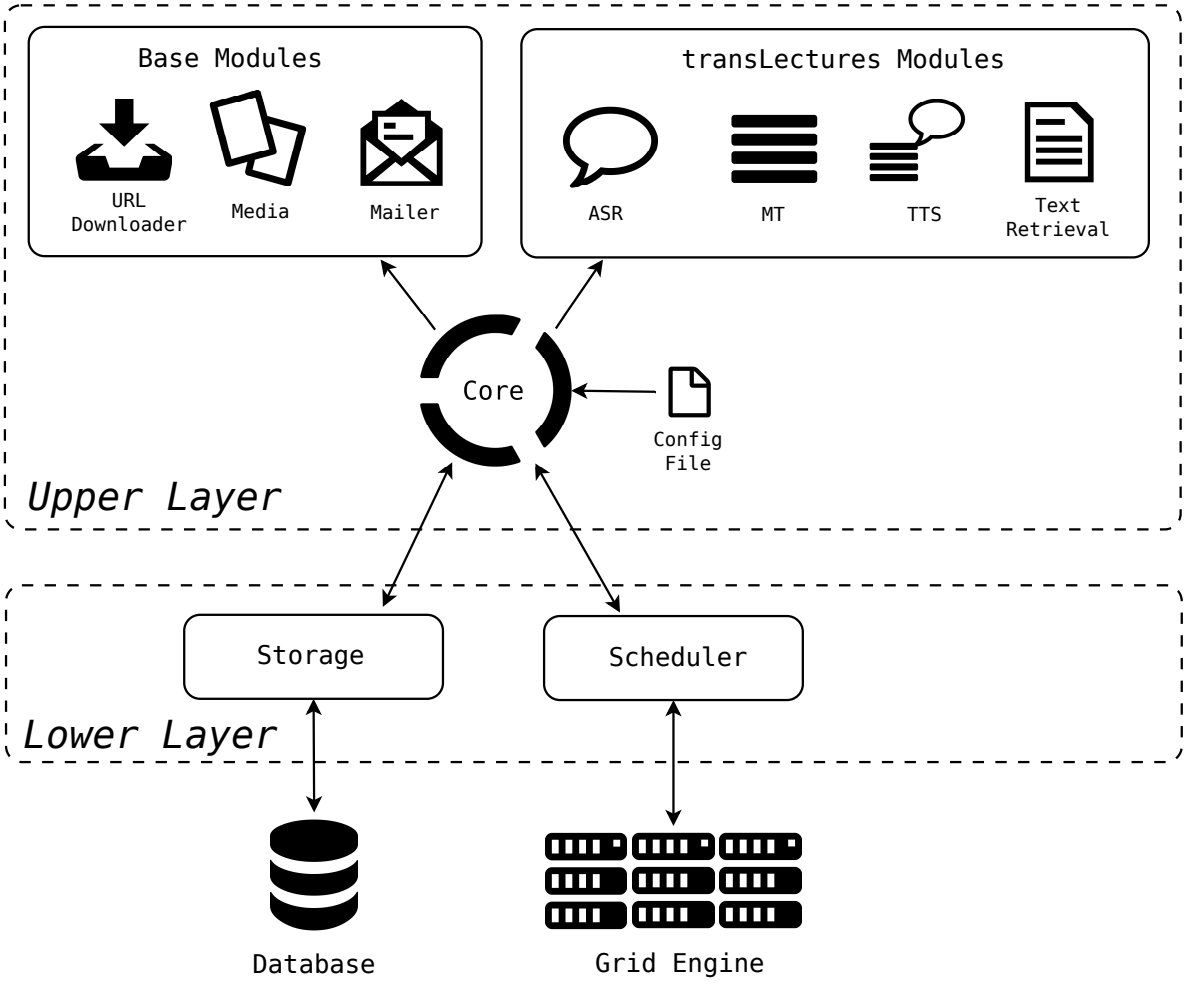

Figure 8.6: Internal structure of the Ingest Service.

  - **Mailer Module:** Module that implements routines to send e-mail notifications regarding upload status updates.

- **transLectures Modules:** Modules that allow to integrate ASR, MT, and TTS technologies into the Ingest Service.

  - **ASR Modules:** Automatic Speech Recognition Modules, used to generate transcription subtitle files.
  - **MT Modules:** Machine Translation Modules, used to generate translated subtitle files.
  - **TTS Modules:** Text-To-Speech Modules, used to generate synthesised audio tracks in a specific language.
  - **Text Retrieval Module**: Extracts plain text information from the different file resources included in the MPF. It also downloads related text documents from the

web, following the approach described in Section 7.2. The extracted text information can be exploited by ASR Modules to enhance the transcription quality by adapting the underlying ASR system to the topic of the media file, as described in Chapter 7.

On the other hand, the lower layer satisfies all local installation dependencies related to data storage and job scheduling. In it we can distinguish:

- **Scheduler layer:** Implements an API to manage transcription and translation processes, typically in a grid engine or job management system.
- **Storage layer:** Implements an API to access to the data stored in the Database and in the TLP Server's hard drive.

Uploads Workflow

In this section we explain the different steps that can be followed by an upload. First, we must distinguish between four types of operations:

- **New Media:** This operation is requested when a newly-recorded, non-existing media is uploaded to TLP for the first time. In this operation, a new media object is created in the Database.
- **Update Media:** This operation is requested when updates are applied to an existing media. For instance, new text resources such as lecture slides might be added to the Media Package File (MPF) to improve the automatic transcription and translations of the existing media.
- **Delete Media:** This operation is requested when a media is deleted from the remote repository.
- **Cancel Upload:** This operation is requested to cancel an ongoing upload for whatever reason.

Figure 8.7 shows the standard Ingest Service workflow: MPFs are uploaded to the TLP via the Web Service's */ingest* interface and stored in the Database. The Ingest Service reads the uploads table of the Database and starts processing the uploaded MPF. With some exceptions, an upload typically follows the following sequential steps:

1. **Media Package Processing:** The MPF is processed for the first time, performing several security, data integrity and data format checks.
2. **Transcription Generation:** A transcription file in DFXP format is generated from the main media file (video, audio) using the proper ASR Module.
3. **Translation(s) Generation:** One or more translation files in DFXP format are generated from a transcription file using the appropriate MT Modules.

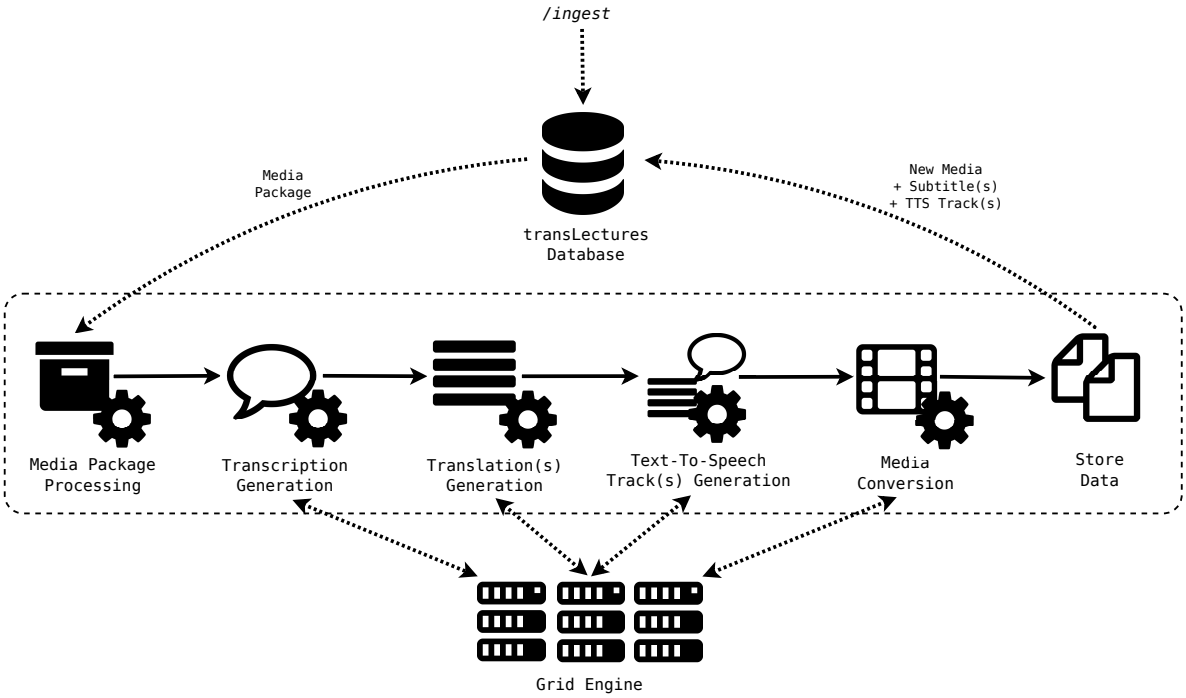

Figure 8.7: Ingest Service Workflow.

4. **Text-To-Speech Track Generation:** One or more synthesised audio track files are generated from a translation file using the appropriate TTS Modules.

5. **Media Conversion:** The main media file is converted into the media formats required by the TLP Player in order to maximise browser compatibility.

6. **Store Data:** For new and update operations, all the data attached in the MPF and automatically generated by the Ingest Service are stored in the Database. For delete operations, all previously stored media files and data are deleted.

In addition, in each awake of the Ingest Service, the Core checks, for every ongoing upload, the statuses of all the submitted processes it is waiting for. These statuses can be:

- **Queued:** Processes are on queue, waiting to be executed.

- **Running:** Processes are being executed.

- **Finished:** All submitted processes finished successfully. The Core moves the upload to the next processing step.

- **Failed:** Some submitted processes failed. The Core changes the upload status to a descriptive error state.

8.3 Integration with poliMedia

As described in Chapter 5, TLP was initially integrated with the poliMedia repository under the scope of the transLectures project. In this early stage, first Spanish ASR and Spanish into English SMT systems were built and integrated in M12 to provide Spanish and English subtitles to Spanish poliMedia video lectures. During the transLectures project, these systems were improved every six months by means of technological upgrades or the inclusion of more training data. Besides, in M24 we added support to English and Catalan poliMedia lectures, so that every lecture in the poliMedia repository was automatically transcribed and translated into Spanish, English and Catalan. To do this, English and Catalan ASR systems, as well as English into Spanish, English into Catalan, Catalan into Spanish, and Catalan into English MT systems were built and integrated into TLP. Similarly to the first M12 systems, these new M24 systems were also improved every six months.

It is worth noting that, following to major ASR and MT system upgrades, the whole repository was automatically re-transcribed and re-translated using improved systems. This led to a massive, overall improvement of the quality of all subtitles of the poliMedia repository.

The rest of this section is structured as follows. First, Section 8.3.1 describes the training procedures and the technological upgrades applied to all ASR and SMT systems. Second, Section 8.3.2 reviews the extensive and periodic automatic evaluations of the quality of these systems. Finally, Section 8.3.3 presents the results obtained in the second real user evaluations carried out in M24.

8.3.1 ASR and SMT Systems

In this Section we describe the ASR and MT systems that were integrated with TLP and poliMedia during the transLectures project, along with an explanation of the training and massive adaptation techniques that were applied to improve their quality. These techniques are referenced afterwards in Section 8.3.2.

ASR Systems

In Section 5.4.4 was described our first Spanish ASR system, based in the conventional GMM-HMM approach and built using the TLK toolkit [14] and the SRILM toolkit [33] to train acoustic and language models, respectively. This system was integrated into TLP in M12. Additionally, two new ASR systems to transcribe English and Catalan poliMedia lectures were built and integrated into TLP for the first time in M24. The initial versions of these systems were technologically grounded on the same acoustic and language modelling techniques used in the M12 Spanish ASR system.

During the whole project, several acoustic and language modelling techniques were applied consistently to all ASR systems to improve the overall transcription quality of the repository. These techniques are listed below:

- **Cluster-based Cepstral Mean and Variance Normalisation (CMVN) [34]**: CMVN aims to reduce the variability between feature representations of utterances from different speakers in order to increase the robustness of the acoustic models. To do this,

input acoustic features are first clustered using prior information about the speakers, and then normalised per cluster in mean and variance.

- **DNN-HMM approach [12]:** Under this approach, the previously trained GMM-HMM system is used to compute an alignment between the acoustic training data and their transcriptions to obtain a mapping from a time step of the acoustic signal (feature vector) to the HMM state associated with this time step. Then, a DNN is trained so that it predicts the HMM state given a feature vector. In this way, we can compute a probability distribution over the HMM states for each input feature vector. Our DNNs were made up of four hidden layers with 3000 units in each layer, and were trained using a discriminative pre-training scheme [12], using as input 15 MFCCs plus derivatives (48-dimensional features) with a sliding-window of 11 feature vectors.

- **Multilingual DNN [20]:** It is a technique that allows us to take advantage of all available training acoustic data, regardless of the language, which aims to train a DNN for a specific language. Specifically, knowing the language of the speech input features, the network is trained only to predict the corresponding language-specific HMM states, while the hidden layers are shared for all languages. This technique has proved that it can deliver significant improvements when dealing with languages in which the availability of training speech data is scarce, as it happens with the Catalan language. Our multilingual DNNs were made up of six hidden layers with 3000 units in each layer, and were trained similarly to monolingual DNNs.

- **Softmax DNN adaptation [36]:** the top layer of the DNN (the so-called softmax layer) is adapted to the speaker characteristics. This adaptation is based on a view of a DNN as non-linear feature extractor (lower layers) followed by a log-linear classifier (softmax layer). Therefore, the adaptation consists in adapting the input of the log-linear classifier (softmax layer) by introducing an affine transformation computed with the output of the system in a previous recognition step. The adapted DNN is therefore used in an additional third recognition step to re-compute the posterior HMM-state probabilities of the input features.

- **System combination with a CNN-HMM based system [8, 9]:** The speaker-adapted multilingual DNN is combined with a Convolutional Neural Network Hidden Markov Model (CNN-HMM) system. In the CNN-HMM approach, MFCC features are replaced by filter bank features. The basic idea here is to apply convolutional filters along the frequency instead of the time axis. More details about how these systems are trained can be seen in [8]. The DNN-CNN combination is performed before the third recognition step by means of a linear interpolation of the outputs of both networks.

- **n-gram based vocabulary selection:** It consists of training 1-gram models from all of the out-of-domain corpora. These 1-gram models are interpolated as in [21] to obtain the model with the least perplexity in a given development set. The final vocabulary is made up of the 50K most probable words according to the interpolated LM plus all the words from the in-domain corpora. Thus, vocabularies from out-of-domain corpora are combined assigning weights to each corpus based on their similarity to the in-domain development set, rather than just the size of the corpus. This technique allowed us to

increase the size of the vocabulary from 50K up to 200K words, since the number of noisy words introduced by this technique is significantly lower than in the previous method.

SMT Systems

In Section 5.4.4 is described our first Spanish into English MT system, which is based on the state-of-the-art phrase-based SMT toolkit Moses [24] to train translation models, and on the SRILM toolkit [33] to train n-gram language models. This system was integrated into TLP in M12 and afterwards improved over time, as we will see in Section 8.3.2. In addition, in M24 a new English into Spanish SMT system was introduced into TLP. This system was entirely based in the inverse SMT system. Enhancements made on both SMT systems were focused on improving the topic adaptation method, that is, the Infrequent n-gram Sentence Selection (ISS) technique [18] described in Section 5.4.4. Concretely, in M18, the ISS technique was slightly improved by better tuning of the infrequent threshold t, as well as by introducing a sentence length normalisation term to the infrequency scores, so that Eq. 5.1 became

$$i(f) = \sum_{\mathbf{w} \in \mathcal{X}} \min(1, N(\mathbf{w})) \max(0, t - C(\mathbf{w})) \cdot \frac{1}{Z(f)} \tag{8.1}$$

being $Z(f)$ the number of n-grams of f.

It is important to note that during the transLectures project we explored many other alternative topic adaptation techniques [10, 26, 28], however ISS was the one that delivered the best results, and therefore, the subsequent SMT systems were the chosen ones to be in the production system.

Finally, also starting from M24, translations from and to Catalan of poliMedia lectures were generated using the open-source rule-based translation system Apertium [16]. Nevertheless, the quality of the translations generated by these systems was not assessed due to the absence of proper test sets. However, Apertium yields high quality Spanish to Catalan (and vice versa) translations since Spanish and Catalan are very similar languages in lexical, syntactical and grammatical terms. On the other hand, acceptable translations from Catalan into English and vice versa are provided by Apertium, if we disregard the problem with out-of-vocabulary words.

8.3.2 Automatic Evaluations

In this section we analyse the progress of the transcription and translation quality in the poliMedia repository throughout the transLectures project. Figure 8.8 shows this progress for all ASR and SMT systems. On the one hand, left-most plot depicts the WER evolution for the Spanish (Es), English (En) and Catalan (Ca) ASR systems computed over their respective test sets during the 36-month period that comprised the transLectures project. Similarly, on the other hand, right-most plot shows the BLEU evolution of the Spanish into English (Es-En) and the English into Spanish (En-Es) SMT systems over the same time span. In both cases, it can be appreciated a positive evolution of the WER (the lower, the better) and BLEU (the higher, the better) curves in favour to a progressively improved transcription and translation

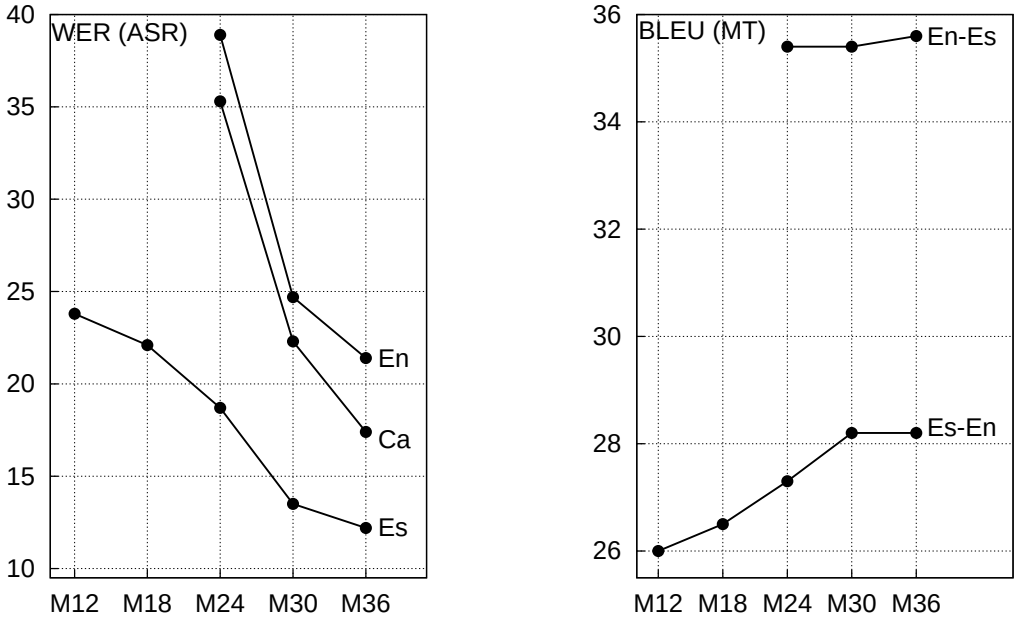

Figure 8.8: Transcription and translation quality progress for all languages considered in the poliMedia repository in ASR (in the left and in terms of WER) and in SMT (in the right and in terms of BLEU) during the 36-month period that comprised the transLectures project. WER and BLEU figures for each ASR and SMT system were computed on their respective test sets.

quality. Further training and evaluation details of every ASR and SMT systems are given all through the rest of this section.

Spanish ASR System

To gauge the progress of the Spanish transcription quality, WER figures were computed over the test set of the Spanish poliMedia corpus (see Table 5.2). Table 8.1 shows the evolution of the WER over time along with a description of the techniques incrementally incorporated to the system and their impact on the WER. The first Spanish ASR system (M12), described in Section 5.4.4, scored 23.8 WER points.

This system was first improved in M18 by applying the Cepstral Mean and Variance Normalisation (CMVN) technique, achieving 22.1 WER points.

A big improvement was obtained in M24, when the ASR migrated from the classical GMM-HMM approach to a hybrid NN-HMM system based on a shallow Neural Network [13], as a preliminary approach to Deep NN. The NN was made up of one hidden layer with 4000 neurons. This improvement was combined with a new language model adapted to slides and documents according to the technique described in Chapter 7. With both enhancements, the M24 system scored 18.7 WER points.

Then, in M30, another significant improvement was obtained by replacing the previous shallow NN with a Deep NN (DNN) made of four layers and 3000 neurons per layer, as well as increasing the number of MFCCs from 13 to 16, achieving 13.5 WER points. Please note

Table 8.1: Evolution of WER (%) for the Spanish ASR system from M12 to M36, computed on the test set of the Spanish poliMedia corpus. System tags labelled with (*) denote the same system but without language model adaptation.

Tag	Description	WER
M12	First system	**23.8**
M18	M12 + CMVN	**22.1**
	M18 + Shallow NN	19.5
M24	+ Improved LM adaptation	**18.7**
	M24* + DNN	17.4
M30	+ LM adaptation	**13.5**
	M30* + Multilingual DNN	16.8
	+ Softmax adaptation	16.5
	+ System combination CNN	16.3
M36	+ Improved LM adaptation	**12.2**

Table 8.2: Statistics of the English Videolectures.NET speech corpus.

	Training	Development	Test
Videos	20	4	4
Speakers	68	11	25
Hours	20	3.2	3.4
Sentences	5K	1K	1.3K
Running Words	130K	28K	34K
Vocabulary Size	7K	3K	3K

that this system is equivalent to the best one reported in Section 7.4.4.

The final M36 system included several improvements. First, we replaced the previous DNN by a 6-layer Spanish-Catalan multilingual DNN of 3000 neurons per layer trained with the Spanish poliMedia corpus plus the Catalan speech corpora used to train the Catalan ASR system (which is described later). Second, we introduced the softmax layer adaptation technique to the multilingual DNN. Third, we trained a CNN-HMM-based system to be combined with the previous speaker-adapted multilingual DNN-HMM system. Finally, we applied the new LM adaptation technique based on the new vocabulary selection algorithm. All these improvements together produced a boost in the system accuracy up to 12.2 WER points, which is almost the same quality that can be achieved by a human transcriber [19].

English ASR System

In M24, a new English ASR system was developed to transcribe English lectures from the poliMedia repository. To report results, we used the test set of the English VideoLectures.NET speech corpus created as part of the transLectures project. Table 8.2 shows basic statistics of this corpus.

Table 8.3: Statistics of the corpora used to train the English Language Model for the English ASR system.

Corpus	Sentences	Running Words	Vocabulary
Spanish-English poliMedia	1.5K	40.2K	3.8K
VideoLectures.NET	5K	130K	7K
COSMAT [25]	55.9K	1.4M	43.2K
TED-LIUM [30]	57K	2.6M	158K
VideoLectures.NET-subtitles	85K	2.1M	41.4K
WIT3 [11]	154K	3.1M	58.4K
Europarl-TV	180.4K	1.9M	31K
News-Commentary-v8 [6]	248K	6.2M	81.1K
EU-TT2	1M	2.1M	203K
Europarl-v7 [23]	2.2M	59.9M	139K
DGT-TM [2]	2.5M	49.2M	298K
Hal [1]	4.6M	104M	972.8K
United Nations [15]	11.2M	320M	132K
GIGA [4]	22.5M	668M	3.1M
News Crawl [7]	53.1M	1.3G	3.3M
Wikipedia	82.6M	1.7G	8.1M
Google-Counts-v1 [27]	-	356.3G	7.3M

This first system was based on the Spanish ASR system described in Section 5.4.4. It was a classical GMM-HMM composed by 3-state tied triphonemes and 64 components per mixture with CMVN and CMLLR adaptation.

On the one hand, the acoustic models were trained using the training set of the English VideoLectures.NET corpus (20h), as well as the EPPS [3] (102h) and TED-LIUM [30] (118h) speech corpora, accounting up to 240 hours of speech data.

On the other hand, the baseline language model was a linear interpolation of several 4-gram language models smoothed with interpolated Kneser-Ney absolute discounting [22] trained on several corpora. These corpora is listed in Table 8.3, accompanied by their main statistics. The linear interpolation weights were optimised in the development set of the English VideoLectures.NET corpus. The vocabulary of the resulting language model is reduced to the 50K most probable words of the out-of-domain corpora, in addition to all the training in-domain words. Finally, the adapted language models were trained by including information extracted from lecture slides and related documents to the interpolation scheme, as described in Chapter 7.

Table 8.4 shows the evolution of the WER over time along with a description of the techniques incrementally incorporated to the system and their impact on the WER. The first M24 English ASR system achieved 38.9 WER points. This system was heavily improved in M30 thanks to the replacement of GMMs by a DNN of 4 hidden layers and 3000 neurons per layer, scoring 24.7 WER points. Finally, in M36, the system was improved in several ways. First, the training data was augmented with 197 hours from the VideoLectures.NET subtitles (112h) and the VoxForge corpus (85h). Second, we applied the softmax layer adaptation technique

Table 8.4: Evolution of WER (%) for the English ASR system from M24 to M36, computed over the test set of the English VideoLectures.NET corpus. System tags labelled with (*) denote the same system but without language model adaptation.

Tag	Description	WER
M24	First system	**38.9**
	M24* + DNN	28.4
M30	+ LM adaptation	**24.7**
	M30* + Extra train data	26.8
	+ Softmax adaptation	25.1
	+ Multilingual DNN	24.6
	+ System combination CNN	24.1
M36	+ New LM adaptation	**21.4**

Table 8.5: Statistics of the Catalan poliMedia speech corpus.

	Training	Development	Test
Videos	177	17	16
Speakers	41	6	6
Hours	21.3	2.4	2.1
Sentences	11.1K	1.3K	1.3K
Running Words	160K	20K	18K
Vocabulary Size	17K	3.9K	3.5K

to the DNN. Third, the monolingual DNN was replaced by an English-Spanish-Catalan multilingual DNN of six hidden layers and 3000 neurons per layer trained using all the available training data, that is, the training data used on the Spanish and Catalan ASR systems plus all English training data, accounting for 620 hours of speech. Fourth, we combined the output of our speaker-adapted multilingual DNN with the output of a previously trained CNN-HMM based system. Finally, we applied the new vocabulary selection technique to train the adapted LM. With all enhancements in hand, the WER was finally reduced to 21.4 points.

Catalan ASR System

In M24 a new Catalan ASR system was developed to transcribe Catalan lectures from the poliMedia repository. To do this, 26 hours of Catalan lectures from poliMedia where manually transcribed and allocated into three different sets for training, tuning, and evaluation purposes. The main figures of this corpus are depicted in Table 8.5. The test set was used to report WER figures.

The first Catalan ASR system (M24) was based on the Spanish ASR system described in Section 5.4.4. It was a classical GMM-HMM composed of 3-state tied triphonemes and 128 components per mixture with CMVN and CMLLR adaptation.

On the one hand, acoustic models were trained using the training set of the Catalan poli-

Table 8.6: Statistics of the corpora used to train the Catalan Language Model.

Corpus	Sentences	Running Words	Vocabulary
Glissando [17]	2.5K	63.8K	3.5K
poliMedia-training	11.1K	160K	17K
Es→Ca poliMedia-training	43.1K	1.1M	29.2K
Àgora [32]	105K	422.7K	25.5K
El Periódico	1.9M	40.8M	342.9K
Wikipedia	4.1M	99.5M	1.4M

Table 8.7: Evolution of WER (%) for the Catalan ASR system from M24 to M36, computed over the test set of the Catalan poliMedia corpus. System tags labelled with (*) denote the same system but without language model adaptation.

Tag	Description	WER
M24	First system	**35.3**
	M24* + DNN	23.5
M30	+ LM adaptation	**22.3**
	M30* + Multilingual DNN	21.6
	+ Softmax adaptation	21.0
	+ System combination CNN	21.0
M36	+ New LM Adaptation	**17.4**

Media corpus (21h) plus two external corpora: Àgora [32] (45h) and Glissando [17] (6h), accounting for 72 hours of speech data.

On the other hand, we trained a linearly interpolated baseline language model computed from in-domain and out-of-domain corpora. In this case, individual 3-gram and 4-gram language models with interpolated Kneser-Ney absolute discounting [22] were trained for each of the following corpora: Catalan poliMedia, El Periódico, Catalan Wikipedia, Àgora [32] and Glissando [17]. Additionally, due to the lack of in-domain text corpora for the Catalan ASR system, the training set of the Spanish poliMedia corpus (see Table 5.2) was translated using the Spanish into Catalan MT system described in Section 8.3.1. This translated data was used to train another language model that was also added to the interpolated language model. Basic statistics of the aforementioned corpora can be found in Table 8.6. The linear interpolation weights were optimised in the Catalan poliMedia development set. The vocabulary of the resulting language model was reduced to the 50K most probable words of the out-of-domain corpora plus all the training in-domain words. Finally, the adapted language models were trained by including information extracted from lecture slides and related documents to the interpolation scheme, as described in Chapter 7.

Table 8.7 shows the evolution of the WER over time along with a description of the techniques incrementally incorporated to the system and their impact on the WER. The first version (M24) of the Catalan ASR system achieved 35.3 WER points on the test set. In M30, the system adopted the hybrid DNN-HMM approach with a DNN made of 4 layers and

Table 8.8: Main figures of the out-of-domain corpora from which sentences were selected using the ISS technique starting from M24 for both Es→En and En→Es SMT systems.

	Sentences	Running Words		Vocabulary	
		en	es	en	es
Europarl TV	180.4K	1.9M	1.8M	31K	42.3K
Europarl-v7 [23]	2M	54.5M	57M	132K	195K
DGT-TM [2]	2.5M	49.2M	54.8M	298K	315K
News-Commentary-v8 [5]	182K	4.7M	5.3M	75.3K	97.7K
United Nations [15]	11.2M	320M	366M	132K	195K

Table 8.9: BLEU figures of the Spanish-English MT system from M12 to M36 computed over the test set of the Spanish-English poliMedia corpus.

Tag	Description	BLEU
M12	First System	26.0
M18	M12 + Improved ISS	26.5
M24	M18 + More OD data	27.3
M30, M36	M24 + bugfix ISS	28.1

3000 neurons per layer, leading to a huge improvement, scoring 24.7 WER points. Finally, the combination of a speaker-adapted 6x3000 multilingual DNN-HMM system with a CNN-HMM system, plus the new vocabulary selection technique for language model adaptation, gave the best result of 17.4 WER points.

Spanish→English SMT System

To control the quality of Spanish into English translations, BLEU figures were computed over the test set of the Spanish-English poliMedia parallel corpus (see Table 5.6).

Table 8.9 shows the evolution of the BLEU over time along with a description of the improvements made to improve the system. The first Spanish into English SMT system (M12) described in Section 5.4.4 achieved 26.0 BLEU points. In M18, the aforementioned improvements made on the ISS technique (a better threshold tuning and adding a sentence-length normalisation term) increased the BLEU score up to 26.5 points. Then, in M24, the out-of-domain corpora pool was extended with additional parallel corpora, achieving 27.3 BLEU points. Table 8.8 shows basic statistics of the corpora that compounded the extended pool. Finally, in M30 it was fixed a bug found in the implementation of the ISS technique, meaning that the previously reported results were not as good as they could really be. Thus, our final system scored 28.1 BLEU points.

Table 8.10: Main figures of the parallel English-Spanish VideoLectures.NET corpus.

	Sentences	Running Words		Vocabulary	
		en	es	en	es
Training	2.1K	59.6K	54.3K	3.8K	5.4K
Development	1.0K	29.4K	27.5K	2.7K	3.5K
Test	1.4K	37.4K	33.6K	2.9K	3.9K

Table 8.11: BLEU results on the test set of the English-Spanish MT system from M24 to M36 computed on the test set of the VideoLectures.NET corpus.

Tag	Description	BLEU
M24, M30	First System	35.4
M36	M30 + bugfix ISS	35.5

English→Spanish SMT System

In M24, the newly incorporated English ASR system generated the first automatic transcriptions for English poliMedia lectures. In order to make these lectures more accessible to Spanish audiences, a new English into Spanish SMT system was built and incorporated to TLP.

To monitor the quality of English into Spanish translations, we report BLEU figures computed over the test set of the English-Spanish VideoLectures.NET parallel corpus, created as part of the transLectures project. Table 8.10 shows basic statistics of this corpus.

On the one hand, translation models for this system were trained using the same parallel data and techniques as in the M24 Spanish into English SMT system. On the other hand, we trained a linearly interpolated language model computed from in-domain and out-of-domain corpora. In particular, we trained individual 4-gram language models with interpolated Kneser-Ney absolute discounting [22] for each of the corpora used to train the baseline language model of the Spanish ASR system, which are listed in Table 5.3. The linear interpolation weights were optimised in the English-Spanish VideoLectures.NET development set.

Table 8.11 shows the evolution of the BLEU figures over time. The first M24 system scored 35.4 BLEU points, which is a very good result: BLEU figures above 30 points are correlated with good-quality translations according to human perception [29]. Finally, the fix of the bug discovered in the ISS technique implementation gave the actual and final quality of the system: 35.5 BLEU points.

8.3.3 User Evaluations

In this section we describe user evaluations carried out under UPV 2013-2014 DeX programme, as a continuation of the ones described in Section 5.5.2 which were carried out in

Table 8.12: Summary of the results obtained in the user evaluations carried out under the DeX 2013-2014 programme.

	Spanish	English	Catalan	Spanish→English
Lecturers	39	12	5	10
Videos	135	57	19	13
Hours	18.3	7.9	1.5	2.1
Error(%)	12.0	36.0	40.4	41.9
RTF	2.7	6.2	5.6	12.2

M12. However, in this case, only the conventional post-editing mode from the TLP Player (see Section 8.2.4) was considered, since it was the preferred interaction method of the users that participated in the previous evaluations. In this edition, participants were requested to review the automatic transcriptions of five poliMedia videos in Spanish, English or Catalan, or to review automatic English translations of three Spanish poliMedia lectures, using the TLP Player. Lecturers' background was diverse but mainly grounded on engineering studies, however a few lecturers were related to other areas, such as business management, social science, and biology. As in the previous edition, participants were free to choose where and when to review those transcriptions, without our supervision. WER and TER as error metrics, and RTF as effort metric, were computed from each revision.

Automatic Spanish, English and Catalan transcriptions to be reviewed were generated by the corresponding M24 ASR systems described and evaluated in Section 8.3.2. Conversely, automatic Spanish into English translations were generated by the corresponding M24 SMT system. Table 8.12 summarises the empirical results obtained in these evaluations.

Review of Spanish Transcriptions

Spanish automatic transcriptions were reviewed by 39 lecturers accounting for 18.3 hours (135 videos). On average, WER was as low as 12 WER points and RTF was 2.7.

As already stated in Section 5.5.2, non-expert users usually need 10 RTF to transcribe a lecture from scratch.In practical terms, that means that a user would require 100 minutes to fully transcribe a video 10 minutes. However, using our post-editing protocol to review our high-quality Spanish transcriptions the review time would be less than 30 minutes, that is, about two thirds reduction of the total user effort.

Review of English Transcriptions

In English, 57 video transcriptions accounting for 7.9 hours were reviewed by 12 non-native volunteers. The review process was the same as for the Spanish lectures. The average RTF was 6.2, still far below of 10 RTF achieved by non-expert users. The reason behind this higher RTF was the poorer transcription quality: the reviewed automatic transcriptions had on overall 36.0 WER points.

In terms of more qualitative feedback, volunteers valued the simplicity and efficiency of the player interface. Volunteers agreed that the quality of the English automatic transcriptions

must be increased to reduce review time. In summary, results were largely positive, and volunteers preferred to review these transcriptions instead of transcribing from scratch.

Review of Catalan Transcriptions

The review of Catalan transcriptions was carried out by 5 lecturers that reviewed 19 video transcriptions accounting for 1.5 hours. The review process was the same described in the Spanish and English reviews. It is important to note that, in this case, Catalan transcriptions had significantly lower quality than those in Spanish. The average RTF was 5.6, while the average WER of automatic transcriptions was 40.4. Lecturers exposed the idea that the transcription quality had to be improved.

Review of Spanish into English Translations

Ten lecturers took part in the review of 13 Spanish into English translations accounting for 2.1 hours. The average RTF was 12.2, while the average TER was 41.9. If we compare this RTF to that achieved by manual translations (about 30 RTF), we observe a significant decrease in user effort. As in transcription evaluations, generally speaking, lecturers were satisfied with the interface, but demanded higher translation quality.

Discussion

The results obtained on these user evaluations reflected significant reductions in user effort when reviewing automatic transcriptions and translations in comparison with doing it from scratch, as in the previous M12 user evaluations (see Section 5.5.2). More specifically, we observed relative user effort reductions of about 70%, 40% and 35% for Spanish, English and Catalan transcriptions, respectively. In the case of English translations, the effort savings were about 60%.

For example, generating from scratch a Spanish transcription and an English translation for a hypothetical Spanish poliMedia lecture of 10 minutes length would take approximately 400 minutes (10 RTF for transcription plus 30 RTF for translation). Using our subtitles as a starting point, a user would need about 30 minutes (2.7 RTF) to review the Spanish transcription plus about 120 minutes (RTF 12.2) to correct the English translation, that is, 150 minutes. This represents a two thirds reduction of user effort.

8.4 MLLP's Transcription and Translation Platform

As stated before, TLP was initially developed under the framework of the transLectures project to fully integrate ASR and SMT technologies into the poliMedia and VideoLectures.NET repositories. During transLectures and beyond, TLP has been maintained and greatly improved by the Machine Learning and Language Processing (MLLP) research group at the UPV. In order to promote the platform itself as well as the research activities of the group, the MLLP launched in February 2015 the Transcription and Translation Platform

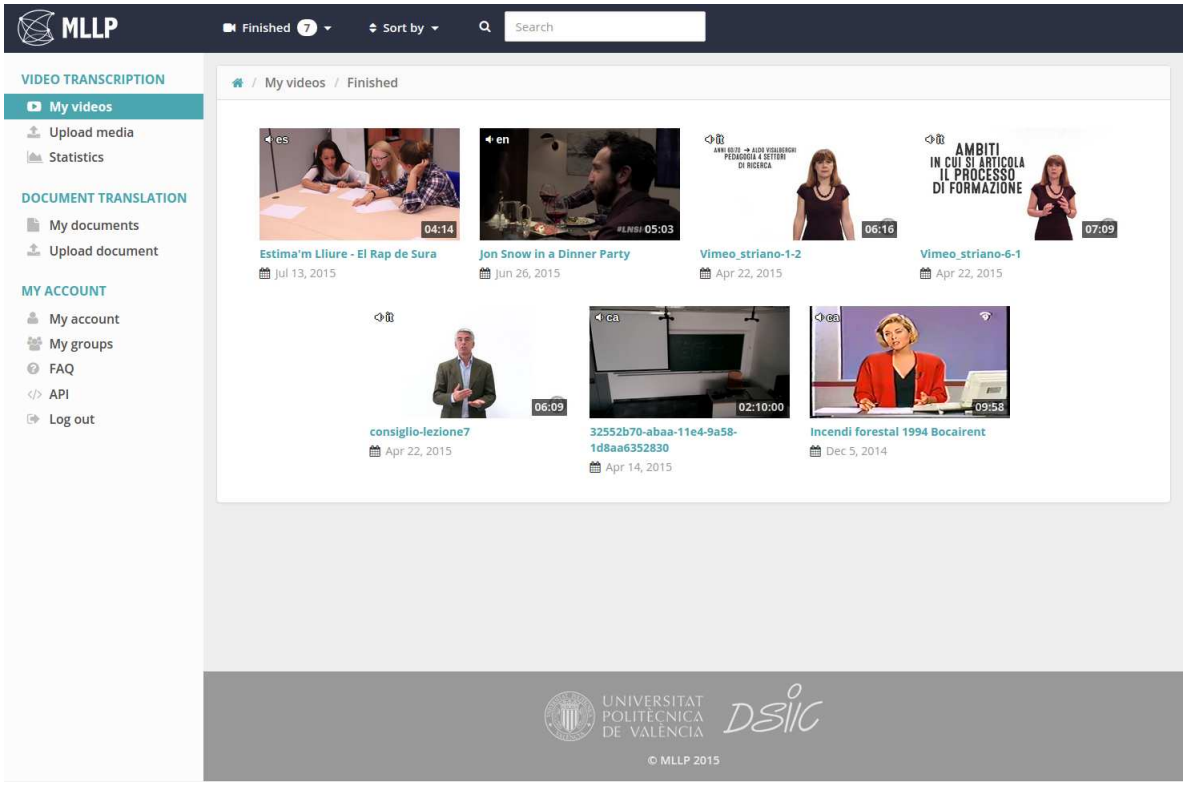

Figure 8.9: Screenshot of the *My Videos* page of the MLLP's Transcription and Translation Platform.

(TTP)[f], on-line service fully grounded on TLP which allows remote transcription and translation of video and audio files. Figure 8.9 shows an screenshot of the main page of TTP.

In TTP, anybody can create an user account to test the ASR and MT technologies offered by the MLLP group. By default, new accounts allow to upload up to five audio or video files accounting to a maximum of two hours. TTP features an *Upload Media* section to easily send video or audio files to the underlying TLP server to be transcribed and translated. Figure 8.10 shows an screenshot of the media upload form. In a first step, the user sets the media file/URL, title, and language. In a second step, the user can adjust some transcription settings, such as enabling or disabling LM adaptation of the ASR system (see Chapter 7), and attaching slides or related text document files for LM adaptation. Finally, in the third and last step, the user can select to which languages wants to translate the media file, in addition to which translated synthesised audio tracks wants to be generated. A Media Package File (MPF) is created and sent to TLP via the */ingest* interface of the Web Service API (see Section 8.2.3). The user can track the progress of the upload in the *My videos* section of his/her user account at any moment until the media file along with its subtitles and/or synthesised audio tracks are available to be played with the integrated TLP Player (see Section 8.2.4).

Advanced TTP users can exploit the possibilities of the TLP's public API (see Section 8.2.3), so that they can seamlessly integrate our ASR, MT and TTS technologies into

[f] `http://ttp.mllp.upv.es`

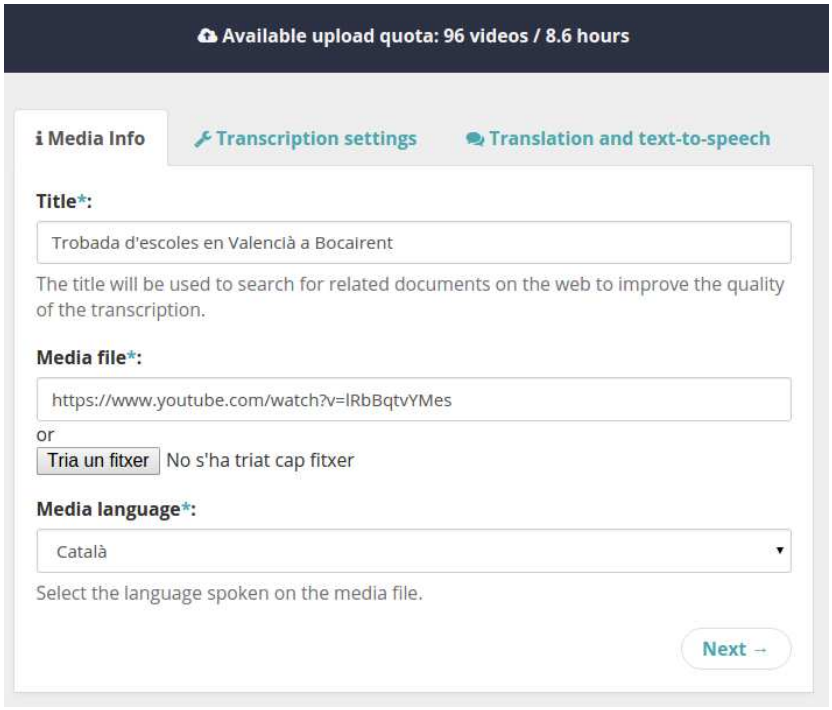

Figure 8.10: Screenshot of the *Upload Media* page of the MLLP's Transcription and Translation Platform.

their media repositories, as done with the poliMedia repository (see Chapter 5).

At the time of writing, 198 different users have created an account on TTP, uploading 879 videos that account for 192 hours. Among these users, we highlight the *Translation Centre for the Bodies of the European Union* (CDT), *Universidad Nacional de Educación a Distancia* (UNED), *Università degli Studi di Napoli Federico II* (UNINA), *Open Universiteit in the Netherlands* (OUNL), *University of Leicester* (ULEIC), *Universidad Carlos III de Madrid* (UC3M), *Universitat Oberta de Catalunya* (UOC), *Universidade Aberta* (UAb), *Tallinn University* (TU), *Université de Bourgogne* (UB), *edX*, *Axa Winterthur*, *Sonic Foundry*, or *Underthemilkyway*, among many others.

8.5 Conclusions

In this Chapter we have presented the final outcomes of this thesis. Firstly, we have described in depth the latest version of TLP, a free and open-source solution to enable cost-effective transcription and translation of video lectures. Secondly, we have shown a true-life example of how the overall transcription and translation quality of a media repository is enhanced by means of technological upgrades on the ASR and MT systems integrated into TLP. Also, we have proven that massive adaptation techniques provide significant improvements in transcription and translation quality. Furthermore, user evaluations reflected that using automatic transcriptions or translations as a start point to generate perfect subtitles saves about two thirds of the total time that would be needed to do that from scratch. Finally, we have pre-

sented and example of a transcription and translation cloud service based on TLP that is giving service to many institutions.

At the time of writing, there exist three physical installations of TLP running in the world:

- **Universitat Politècnica de València (UPV):** Since June 2013, an dedicated TLP server is serving automatic transcriptions and translations of Spanish poliMedia lectures, as described in Chapter 5. More recently, English and Catalan lectures are being transcribed too (see Section 8.3). In summary, TLP automatically managing Spanish, Catalan and English subtitles of roughly 15.000 video lectures (3.100 hours), and increasing.

- **Universidad Carlos III de Madrid (UC3M):** Starting from September 2014, TLP is generating Spanish and English subtitles for Spanish and English UC3M lectures and videos from their Massive Open Online Courses (MOOCs). Latest MLLP's M36 English and Spanish ASR systems as well as M36 English↔Spanish SMT systems were transferred and integrated into their local TLP Server to provide such subtitles. In overall, 418 UC3M videos accounting for 63 hours have been processed by their own TLP Server.

- **MLLP's Transcription and Translation Platform (TTP):** As described in Section 8.4, the MLLP research group launched on February 2015 this online subtitling platform grounded on TLP that is being currently used by several world-wide universities and companies. So far, TTP has processed 879 video and audio files in several languages accounting for 192 hours.

To conclude, we want to highlight that TLP is not a system prototype that has been tested in a lab under controlled conditions. It is working as a production system serving high-quality automatic subtitles in three different real-life scenarios. All in all, we believe that TLP has a true potential to become a widely used solution to enable automatic multilingual subtitling in large video lecture repositories.

As to future work, our plans are to extend TLP functionalities in order to give full support to the transcription and translation of Massive Open On-line Courses (MOOCs), either video and text contents. MOOCs are currently leading the open on-line learning framework, and we do not want to miss this opportunity. Also, it is important to note that TLP could be easily exported as a real solution to many other areas, such as television or cinema, i.e. using TLP as a professional tool to generate cost-effective multilingual subtitles for movies with crowd-sourcing capabilities; and therefore, we will also scrutinise this possibility. Finally, we plan to reactivate the intelligent interaction mode of the TLP Player using speaker-adapted word confidence measures [31], motivated by the fact that our current ASR systems are now yielding significant better transcriptions than those used in the past to evaluate this particular interaction mode (see Section 5.5.2).

Bibliography

[1] HyperArticles en ligne, HAL corpus. `http://hal.archives-ouvertes.fr/`.

[2] The DGT Translation Memory Corpus. `https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory`.

[3] The EPPS Speech corpus at ELRA. `http://catalog.elra.info/product_info.php?products_id=1035`.

[4] The GIGA corpus for the WMT'10. `http://www.statmt.org/wmt10/training-giga-fren.tar`.

[5] The News Commentary corpus for the WMT'13 (bilingual). `http://www.statmt.org/wmt13/training-parallel-nc-v8.tgz`.

[6] The News Commentary corpus for the WMT'13 (Monolingual). `http://www.statmt.org/wmt13/training-monolingual-nc-v8.tgz`.

[7] The News Crawl corpus for WMT'13. `http://www.statmt.org/wmt13/training-monolingual-news-2012.tgz`, 2012.

[8] O. Abdel-Hamid, L. Deng, and D. Yu. Exploring convolutional neural network structures and optimization techniques for speech recognition. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 3366–3370, 2013.

[9] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4277–4280, March 2012.

[10] A. Axelrod, X. He, and J. Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 355–362, 2011.

[11] M. Cettolo, C. Girardi, and M. Federico. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May 2012.

[12] G. 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 (receiving 2013 IEEE SPS Best Paper Award)*, 20(1):30–42, January 2012.

[13] G. E. Dahl, S. Member, D. Yu, S. Member, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. In *IEEE Transactions on Audio, Speech, and Language Processing*, 2012.

[14] M. A. del Agua, A. Giménez, N. Serrano, J. Andrés-Ferrer, J. Civera, A. Sanchís, and A. Juan. The translectures-upv toolkit. In *Advances in Speech and Language Technologies for Iberian Languages - Second International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain, November 19-21, 2014. Proceedings*, pages 269–278, 2014.

[15] A. Eisele and Y. Chen. Multiun: A multilingual corpus from united nation documents. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, 2010.

[16] M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144, 2011.

[17] J. M. Garrido, D. E. Mancebo, L. Aguilar, V. Cardeñoso-Payo, E. Rodero, C. de la Mota, C. G. Ferreras, C. Vivaracho-Pascual, S. Rustullet, O. Larrea, Y. Laplaza, F. Vizcaíno, E. Estebas-Vilaplana, M. Cabrera, and A. Bonafonte. Glissando: a corpus for multidisciplinary prosodic studies in spanish and catalan. *Language Resources and Evaluation*, 47(4):945–971, 2013.

[18] G. Gascó, M. Rocha, G. Sanchis-Trilles, J. Andrés-Ferrer, and F. Casacuberta. Does more data always yield better translations? In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*, pages 152–161, 2012.

[19] T. J. Hazen. Automatic alignment and error correction of human generated transcripts for long speech recordings. In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 2006.

[20] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7304–7308, May 2013.

[21] F. Jelinek and R. L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proc. of the Workshop on Pattern Recognition in Practice*, pages 381–397, Amsterdam, The Netherlands: North-Holland, 1980.

[22] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing, ICASSP '95, Detroit, Michigan, USA, May 08-12, 1995*, pages 181–184, 1995.

[23] P. Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT.

[24] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*, 2007.

[25] P. Lambert, J. Senellart, L. Romary, H. Schwenk, F. Zipser, P. Lopez, and F. Blain. Collaborative machine translation service for scientific texts. In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*, pages 11–15, 2012.

[26] S. Mansour, J. Wuebker, and H. Ney. Combining translation and language model scoring for domain-specific data filtering. In *2011 International Workshop on Spoken Language Translation, IWSLT 2011, San Francisco, CA, USA, December 8-9, 2011*, pages 222–229, 2011.

[27] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[28] R. C. Moore and W. D. Lewis. Intelligent selection of language model training data. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, Short Papers*, pages 220–224, 2010.

[29] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002*, pages 311–318, 2002.

[30] A. Rousseau, P. Deléglise, and Y. Estève. TED-LIUM: an automatic speech recognition dedicated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 125–129, 2012.

[31] I. Sánchez-Cortina, J. Andrés-Ferrer, A. Sanchís, and A. Juan. Speaker-adapted confidence measures for speech recognition of video lectures. *Computer Speech & Language (In Press)*.

[32] H. Schulz and J. A. Fonollosa. A catalan broadcast conversational speech database. In *I Joint SIGIL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages, Porto Salvo, Portugal, September*, 2009.

[33] A. Stolcke. SRILM - an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*, 2002.

[34] O. Viikki and K. Laurila. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1-3):133–147, 1998.

[35] World Wide Web Consortium (W3C). Distribution Format Exchange Profile (DFXP). `http://www.w3.org/tr/2006/cr-ttaf1-dfxp-20061116`.

[36] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong. Adaptation of context-dependent deep neural networks for automatic speech recognition. In *2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, December 2-5, 2012*, pages 366–369, 2012.

CHAPTER 9

CONCLUSIONS

Contents

9.1 Summary

In this section we summarise the work carried out for this thesis. Firstly, in Chapter 3 we have proposed two novel explicit conditional phrase length models for SMT: the standard length model and the specific length model. These phrase-length models have been integrated in a state-of-the-art log-linear SMT system as additional feature functions, providing in most cases a systematic boost of translation quality on unrelated language pairs.

Secondly, in Chapter 4 is described an efficient AS system clearly inspired in GMM-HMM based ASR that exhibits excellent performance detecting speech segments at near real-time speeds. This system was submitted to the Audio Segmentation competition of the Albayzin 2012 Evaluations, achieving the 2nd place, very close to the winner system.

Thirdly, Chapter 5 presented a system architecture that allows the integration of ASR and MT technologies into video lecture repositories. Its implementation, *The transLectures-UPV Platform*, was integrated into the UPV's poliMedia repository. Preliminary results on automatic and human evaluations suggested that the delivered transcriptions and translations were of an acceptable quality though had to be improved, and that the provided tools to edit subtitles were comfortable, productive, and very easy to use.

Then, Chapter 6 described a lecture RS that exploits automatic speech transcriptions of video lectures to zoom in on user interests at a semantic level. This RS was particularly implemented for the VideoLectures.NET repository, although preliminary, quantitative-based metrics computed in comparison with the previously existing RS were not encouraging, suggesting that qualitative-based metrics must be explored in order to fairly compare both systems.

Next, Chapter 7 proposed an effective method to retrieve documents from the web and use them to build adapted language models for video lecture transcription. The application of this technique under a solid experimental setting reported systematic and significant WER improvements of above 10%.

Finally, Chapter 8 presented the latest version of the *transLectures-UPV Platform* as an evolution of the first version presented in Chapter 5. TLP has been publicly released as open-source software and it is free to download and use. Similarly, the preliminary automatic and user evaluations in the poliMedia repository presented in Chapter 5 were extended, showing how the overall transcription and translation quality of a media repository can be enhanced over time by means of introducing technological upgrades into the ASR and MT systems integrated into TLP. Also, we have proven that massive adaptation techniques provide significant improvements in transcription and translation quality. Furthermore, user evaluations reflected that using automatic transcriptions or translations as a start point to generate perfect subtitles saves about two thirds of the total time that would be needed to do that from scratch. Finally, we presented the MLLP's Transcription and Translation Platform, a cloud service grounded on TLP that is serving high-quality automatic subtitles to several institutions in Spain and Europa.

In summary, the main contributions of this thesis are the following:

- An explicit conditional phrase length modelling approach for SMT that provide systematic and significant improvements over strong baselines for different language pairs.
- A simple yet powerful and efficient approach for AS to detect speech segments in audio

signals.

- A free and open-source solution to integrate ASR and MT technologies into large video lecture repositories, capable of generating cost-effective high-quality multilingual subtitles.
- An extensive evaluation of several ASR and MT systems in different languages to gauge the positive effect of massive adaptation techniques in video lecture repositories.
- A new approach to video lecture recommendation for content-based RS using automatic speech transcripts.
- A new language model adaptation technique for ASR that yields significant WER improvements over solid baselines.

9.2 Publications

Most of the work of this thesis has directly yielded articles in international conferences and journals. In this section we enumerate these contributions to the scientific community, highlighting their relationship with the chapters of this thesis.

The proposed Explicit Length Models for Statistical Machine Translation in Chapter 3 led two publications: one international conference and one journal article, being this latter an extension of the former:

- *Explicit Length Modelling for Statistical Machine Translation*. Joan Albert Silvestre-Cerdà, Jesús Andrés-Ferrer and Jorge Civera. Pattern Recognition and Image Analysis, vol. 6669, pp. 273-280, Springer Berlin Heidelberg, 2011. ISBN 978-3-642-21256-7, DOI 10.1007/978-3-642-21257-4_34.
  - Contribution: first author, main contributor of the published work.
- *Explicit Length Modelling for Statistical Machine Translation*. Joan Albert Silvestre-Cerdà, Jesús Andrés-Ferrer and Jorge Civera. Pattern Recognition, vol. 45, no. 9, pp. 3183-3192, Elsevier, 2012. ISSN 0031-3203, DOI 10.1016/j.patcog.2012.01.006.
  - Contribution: first author, main contributor of the published work.

The AS system described in Chapter 4 participated in the Audio Segmentation competition from the Albayzin 2012 Evaluations, achieving the 2nd position out of six systems and five participants:

- *Albayzin Evaluation: The PRHLT-UPV Audio Segmentation System*. Joan Albert Silvestre-Cerdà, Adrià Giménez, Jesús Andrés-Ferrer, Jorge Civera and Alfons Juan. Online proceedings of the VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop (IberSpeech 2012), Madrid (Spain), pp. 596-600. 2012. ISBN 84-616-1535-2.
  - Contribution: first author, main contributor of the published work.

In Chapter 5, the description of the system architecture of the first version of our transLectures Platform, as well as the first automatic and human evaluations carried out to assess the benefits of the whole platform, resulted in three publications:

- *A System Architecture to Support Cost-Effective Transcription and Translation of Large Video Lecture Repositories.* Joan Albert Silvestre-Cerdà, Alejandro Pérez, Manuel Jiménez, Carlos Turró, Alfons Juan and Jorge Civera. Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC 2013), Manchester (UK), pp. 3994-3999, 2013. DOI 10.1109/SMC.2013.682.

  – Contribution: first author, main contributor of the published work.

- *TransLectures*. J. A. Silvestre-Cerdà, M. A. 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. Online proceedings of the VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop (IberSpeech 2012), Madrid (Spain), pp. 345-351. 2012. ISBN 84-616-1535-2.

  – Contribution: first author, main contributor of the published work.

- *Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories*. Juan Daniel Valor Miró, Joan Albert Silvestre-Cerdà, Jorge Civera, Carlos Turró and Alfons Juan. Speech Communication, vol. 74, pp. 65-75, Elsevier, 2015. ISSN 0167-6393, DOI 10.1016/j.specom.2015.09.006.

  – Contribution: co-author, contributed to generate the automatic transcriptions of the lectures to be reviewed by users, and to set up the experimental conditions needed to perform all user trials.

The video lecture recommender system developed for the VideoLectures.NET site in collaboration with Alejandro Pérez González de Martos and presented in Chapter 6 yielded the following publication:

- *Using Automatic Speech Transcriptions in Lecture Recommendation Systems*. A. Pérez-González-de-Martos, J.A. Silvestre-Cerdà, M. Rihtar, A. Juan and J. Civera. Online proceedings of VIII Jornadas en Tecnología del Habla and IV Iberian SLTech Workshop (IberSpeech 2014), Las Palmas de Gran Canaria (Spain), pp. 149-158, 2014. ISBN 978-84-617-2862-6.

  – Contribution: co-author, responsible for the analysis, design and implementation of the whole Recommender System except for the *regular* and *occasional* update modules.

The language model adaptation technique described in Chapter 7, in collaboration with Adrià Martínez Villaronga, was published in the following collection:

- *Language model adaptation for lecture transcription by document retrieval*. A. Martínez-Villaronga, M. A. Del-Agua, J. A. Silvestre-Cerdà, J. Andrés Ferrer and A. Juan. Lecture Notes in Computer Science, vol. 8854, pp.129-137, Springer International Publishing, 2014. ISBN 978-3-319-13622-6, DOI 10.1007/978-3-319-13623-3_14.

  – Contribution: co-author, responsible for the design and implementation the proposed document retrieval module; also contributed to the design and execution of the experiments.

Finally, the second user evaluations presented in Chapter 8, in collaboration with Juan Daniel Valor Miró, resulted in an international conference paper:

- *Efficient Generation of High-Quality Multilingual Subtitles for Video Lecture Repositories*. Juan Daniel Valor Miró, Joan Albert Silvestre-Cerdà, Jorge Civera, Carlos Turró and Alfons Juan. Design for Teaching and Learning in a Networked World, Lecture Notes in Computer Science, vol. 9307, pp. 485-490, Springer International Publishing, 2015. ISBN 978-3-319-24257-6, DOI 10.1007/978-3-319-24258-3_44.

  – Contribution: co-author, contributed to generate the automatic transcriptions and translations of the lectures to be reviewed by users, and to set up the experimental conditions needed to perform all user trials.

9.3 Future Work

The work done for this thesis has revealed several technological and scientific opportunities that can be tackled as future work.

Regarding the explicit length models for SMT presented in Chapter 3, we will carry out a full Viterbi-like iterative training procedure that may outperform the proposed Viterbi-based estimation method. Moreover, we would also study the combination of the Viterbi extracted counts with those heuristically extracted as a smoothing technique. Finally, we will also explore alternative optimisation methods to MERT such as MIRA [1].

Although the AS system described in Chapter 4 already provides excellent performance when detecting speech segments, it could be interesting to explore the adoption of the hybrid DNN-HMM approach, that is, to replace the emission probabilities of HMMs by DNNs instead of GMMs.

With regard to the RS presented in Chapter 6, our simple click-based evaluation suggested that the previous RS was slightly more used than ours. For this reason, we intend to evaluate and compare both RS using other evaluation approaches that truly measure the suitability of their recommendations. Besides, we plan to retrain the RS with English transcriptions of better quality, in order to study the impact of the quality of the speech transcriptions on the quality of the recommendations. Furthermore, since TLP can provide multilingual subtitles, we would like to extend this approach in order to also provide recommendations of related lectures in other languages.

With respect to our new language model adaptation technique proposed in Chapter 7, we plan to perform a comparative study between our document retrieval method and alternative

methods proposed by other authors [2, 3, 5]. Also, we will study how to adapt language models to the vocabulary of the speaker in order to disambiguate certain words and expressions frequently used by the lecturer that are not necessarily related to the topic of the lecture.

Finally, in relation to *The transLectures-UPV Platform* presented in Chapters 5 and 8, we plan to extend its functionalities in order to give full support to transcription and translation of Massive Open On-line Courses (MOOCs). Also, we will be open to export TLP to other fields such as television or cinema, since only cosmetic changes are needed by TLP to fulfil the needs of these areas. Finally, we will explore the possibility of reactivating the intelligent interaction mode of the TLP Player using speaker-adapted word confidence measures [4] with our greatly improved ASR and MT systems.

Bibliography

[1] D. Chiang, Y. Marton, and P. Resnik. Online large-margin training of syntactic and structural translation features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 224–233, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[2] C. Munteanu, G. Penn, and R. Baecker. Web-based language modelling for automatic lecture transcription. In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, pages 2353–2356, 2007.

[3] I. Rogina and T. Schaaf. Lecture and presentation tracking in an intelligent meeting room. In *4th IEEE International Conference on Multimodal Interfaces (ICMI 2002), 14-16 October 2002, Pittsburgh, PA, USA*, pages 47–52, 2002.

[4] I. Sánchez-Cortina, J. Andrés-Ferrer, A. Sanchís, and A. Juan. Speaker-adapted confidence measures for speech recognition of video lectures. *Computer Speech & Language (In Press)*.

[5] T. Schlippe, L. Gren, N. T. Vu, and T. Schultz. Unsupervised language model adaptation for automatic speech recognition of broadcast news using web 2.0. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 2698–2702, 2013.

LIST OF FIGURES

LIST OF TABLES