UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

TREBALL FI DE MÀSTER

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Aplicació de xarxes neuronals profundes en traducció automàtica per a recursos educatius oberts

Màster Universitari en Intel·ligència Artificial, Reconeixement de Formes i Imatge Digital.

Pau Baquero Arnal

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

Director experimental:
Adrià Agustí Martínez Villaronga

11 de desembre de 2017

Índex

Capítol 1

Introducció

Abans d'entrar directes a la matèria, parlarem un poc de com està estructurada per a fer-nos una idea de què tenim entre mans. Aquesta memòria té 5 capítols, i aquest primer està pensat principalment per facilitar la lectura i comprensió de la resta. Ací introduïm al treball realitzat, posant-lo en context i explicant-ne la motivació. La idea és bastir el marc perquè es puga seguir la resta de la memòria amb facilitat. El treball gira al voltant d'abordar el problema de la traducció automàtica des d'una aproximació relativament nova: els models *sequence-to-sequence* de xarxes neuronals.

En aquest capítol introductori, la primera secció (1.1) parla de la motivació del tema i del treball. La segona (1.2) vincula el meu Treball Fi de Grau (TFG) sobre transcripció automàtica amb aquest treball. La tercera secció (1.3) fa referència a la recerca prèvia en avaluacions d'usuari en transcripció i traducció automàtiques. La quarta (1.4) explica l'evolució dels sistemes de traducció automàtica en els darrers anys i el canvi de models basats en frases (*phrase-based*) a models de xarxes neuronals (*Neural Machine Translation*, o *neural MT*). La cinquena i última secció (1.6) defineix les mètriques utilitzades en aquest treball per avaluar les traduccions automàtiques.

1.1 Motivació

Els sistemes de traducció automàtica permeten generar traduccions ràpides i barates, per bé que els millors sistemes encara tenen errors molt considerables en comparació als traductors humans. En aquest treball, l'interés en la traducció automàtica es concreta dins de la Universitat Politècnica de València per a integrar els sistemes en plataformes per a transcriure i traduir vídeos educatius automàticament.

L'objectiu pràctic és millorar l'accessibilitat de vídeos educatius, proporcionant eines automàtiques de transcripció, traducció i síntesi de veu en altres llengües. Amb aquesta finalitat, el grup de recerca *Machine Learning and Language Processing* (MLLP) desenvolupa eines automàtiques de transcripció, traducció i síntesi de veu. Aquestes eines ja ofereixen resultats excel·lents, com podreu comprovar en el capítol 3.

La tasca de traducció és, ara mateix, la més susceptible de millora de les tres dins de l'àmbit de les tecnologies del llenguatge, i d'ací ve la justificació principal d'haver escollit aquest tema per al Treball Fi de Màster.

1.2 Transcripció

Aquesta secció posarà en context el treball desenvolupat en el meu TFG [6] dins del marc d'aquest Treball de Fi de Màster. El TFG abordava, des del punt de vista de l'aplicació a vídeos educatius, el problema previ a la traducció: la transcripció automàtica d'àudio, també conegut com a "reconeixement de la parla" o ASR (*Automatic Speech Recognition*). El treball comparava els *toolkits* TLK [10] i Kaldi [22]: el primer com a l'eina utilitzada pel grup MLLP, i el segon per ser un dels més utilitzats globalment en ASR.

Si bé no té sentit entrar ara en detall del funcionament dels *toolkits*, sí que és apropiat que revisitem el treball de forma resumida, sobretot els resultats de la comparació i les conclusions, com a treball previ en un àmbit d'aplicació molt relacionat amb aquest TFM.

Per entendre les possibles causes de diferències en resultats entre TLK i Kaldi, primer cal tindre una idea de quins tipus de models es gasten en els sistemes d'ASR entrenats. Fonamentalment, es pren la hipòtesi que maximitza l'equació 1.1, on w és una seqüència de paraules, x és la seqüència de característiques acústiques de l'àudio, i $\hat{w}$ el resultat obtingut.

$$\hat{w} = \arg\max_{w} p(x|w) \cdot p(w) \tag{1.1}$$

Ací, $p(w)$ és el model de llenguatge (modelitza la probabilitat d'una frase), i $p(x|w)$ és el model acústic (modelitza la probabilitat d'una seqüència de sons donada una frase). El model de llenguatge va ser entrenat mitjançant una eina externa, SRILM [27], així que era el mateix TLK i Kaldi. Per tant, les diferències es poden explicar per dos motius: els models acústics o la cerca entre possibles hipòtesis.

En els dos casos vam entrenar sistemes amb un procediment tan semblant com vam poder. Aquest procediment constava dels següents pasos:

1. Entrenament del model acústic de gaussianes
2. Adaptació del model acústic al locutor (CMLLR)
3. Entrenament del model acústic amb xarxes neurals profundes (DNN, *Deep Neural Networks*)

Tots els models acústics estan basats en models ocuts de Markov (HMM, de *Hidden Markov Models*); la diferència està en com es modelitzen les probabilitats d'emissió. L'adaptació CMLLR no afecta al tipus de model, sinó a les característiques. En resum, això vol dir que poden haver-hi models estàndards i adaptats, amb gaussianes i amb xarxes neuronals. Pel funcionament dels *toolkits*, TLK genera els quatre tipus de

model mentre que Kaldi no genera el model de xarxes sense adaptació com a resultat final.

La mètrica principal amb la qual vam avaluar els sistemes és la del WER (*Word Error Rate*), que es defineix en l'equació 1.2, on N són el nombre de paraules i I, D i S són els errors comesos bé per inserció (*Insertions*), eliminació (*Deletions*) o substitució (*Substitutions*) de paraules.

$$\text{WER} = \frac{I + D + S}{N} \tag{1.2}$$

Vam comparar el WER en la tasca de poliMedia-català de tots els models obtinguts per les dues eines. Els resultats estan representats en la taula 1.1, que representa el WER en funció de *toolkit* i de tipus de model per a aquesta tasca.

Taula 1.1: Comparació de l'error del reconeixement en cada etapa

	Gaussianes		Xarxes	
Eina	Estàndard	Adaptat	Estàndard	Adaptat
Kaldi	37.3%	28.8%	—	21.1%
TLK	43.6%	34.4%	25.3%	22.2%

En aquesta taula podem comprovar que hi ha una distància lleugera entre els models de xarxes, però considerable entre els models de gaussianes. Tenint en compte que els models de gaussianes s'utilitzen per inicialitzar els models de xarxes, això pareixia indicar que l'eina TLK era susceptible de millora si es refinava la modelització de gaussianes per a HMMs.

Per a mesurar l'eficiència temporal, vam utilitzar la mètrica de l'RTF (*Real Time Factor*), definida en l'equació 1.3, on t_p és el temps de processament i t_v és la duració del vídeo o àudio.

$$RTF = \frac{t_p}{t_v} \tag{1.3}$$

Els resultats de RTF estan indicats en la taula 1.2, on apareixen els resultats de Kaldi amb i sense *rescoring*.

Taula 1.2: Comparació de RTF

Sistema	RTF
TLK	9.1
Kaldi (inicial)	0.8
Kaldi (amb rescoring)	2.2

Aquestes mesures mostren que el sistema de TLK va significativament més lent que el de Kaldi, i que hi ha molt a millorar especialment en el pas previ al *rescoring*. Tot semblava indicar que la cerca d'hipòtesis en Kaldi estava més optimitzada.

Aquest treball va ser útil per al grup MLLP en el seu desenvolupament de TLK. Amb aquestes febleses ja identificades, es va treballar sobre elles i les últimes versions de TLK ja tenen uns models de gaussianes i algorismes de cerca comparables als de Kaldi.

1.3 Avaluació sobre casos d'ús reals

El grup MLLP desenvolupa sistemes automàtics de transcripció i traducció per a aplicar-los sobre vídeos educatius. Dins del context d'aquest TFM, cal parlar de l'avaluació d'aquesta aplicació en casos d'ús real, que va resultar en un article publicat [17]. Aquest treball va aplicar sistemes automàtics adaptats de transcripció i traducció a dos casos reals: la plataforma de MOOCs EMMA, i el repositori de vídeos educatius (poliMedia) *UPV media*; amb l'objectiu de generar subtítols multilingües automàtics de bona qualitat a un baix cost.

El treball posava de relleu la importància dels sistemes de transcripció i traducció, mesurant l'impacte que tenen els seus resultats sobre el món real. Féiem tres tipus d'avaluació: de qualitat de resultats, de temps humà requerit per revisar-los, i d'impacte sobre els casos d'ús. Encara que aquest tema ocupa un capítol sencer d'aquest TFM (el capítol 3), en aquesta secció resumirem els resultats obtinguts a partir d'aquest treball també com a treball previ relacionat i que motiva el TFM i justifica treballar en traducció automàtica.

Pel que fa a qualitat de resultats, els sistemes adaptats van tindre uns errors prou baixos com per fer-los útils per a la revisió. Els sistemes de transcripció van obtindre resultats majoritàriament per sota del 25% de WER (veure secció 1.2), i els sistemes de traducció per sota del 50% de TER (veure secció 1.6). Aquests resultats són més bons que els obtinguts per proveïdors comercials: un 38% millors que els subtítols de YouTube i un 17% millors que les traduccions de Google Translate. Això es deu, principalment, a l'adaptació dels sistemes acústics al locutor i l'adaptació al domini en el cas de la traducció.

Quant al temps de revisió, els resultats del treball indiquen que una persona pot estalviar-se des del 30% al 70% del temps dedicat a la revisió dels subtítols, i del 25% al 75% del temps dedicat a revisar les traduccions, comparat amb partir des de zero. No només això sinó que es va proposar un sistema de regressió simple per a aproximar el temps de revisió en funció de l'error, per a trancripció i per a traducció.

Per últim, l'impacte que va tindre en els casos d'ús va ser notable. La disponibilitat de subtítols multilingües va provocar, en la plataforma EMMA, un augment del 70% en les inscripcions d'alumnes. En la plataforma *UPV media* no es va replicar aquest augment per les condicions del repositori i de l'alumnat de la UPV, però es va millorar l'accessibilitat dels vídeos per a persones amb dificultat auditiva i estudiants no nadius, i els subtítols van permetre desenvolupar funcions d'indexació i cerca de vídeos.

1.4 Traducció: de *phrase-based* a *neural MT*

La traducció automàtica és un problema obert, i s'hi han donat diferents formes d'encarar-lo des de punts de vista molt distints. En aquest treball, ens centrem en l'aproximació estadística. Així, la traducció automàtica es pren com a un problema de classificació, on hem de prendre la traducció per a una certa frase amb màxima probabilitat.

Fins a l'irrupció de les xarxes neuronals en traducció automàtica (*neural MT*), els sistemes estat-de-l'art eren els basats en frases (*phrase-based*). En aquesta secció, veurem les idees que fonamenten els sistemes *phrase-based*, per veure com funcionen els sistemes més avançats de què disposa la Universitat Politècnica de València. També explorarem les bases del *neural MT*, on hi ha esperances de què puguen obtindre's millors resultats.

1.4.1 Phrase-based

Els sistemes *phrase-based* es basen en la idea de segmentar una frase, traduir segment a segment i reordrenar els segments per a generar la frase traduïda. Consten de tres models:

1. Model de **traducció de segments**: indica com de probable és que un segment es traduïsca per un altre. Pren forma de taula amb parells de segments i probabilitats.

2. Model de **reordenació**: indica la probabilitat de cada reordenació possible dels segments.

3. Model de **llenguatge**: indica la probabilitat *a priori* que una certa frase es done en eixa llengua.

Aquests tres models es combinen per donar una probabilitat a cada traducció possible, i entre totes les possibles traduccions l'objectiu és trobar la de màxima probabilitat.

Els sistemes *phrase-based* van ser una millora substancial respecte als anteriors, basats en paraules, i han representat l'estat-de-l'art de la traducció automàtica durant molt de temps. En els últims anys, s'han millorat introduïnt elements de xarxes neuronals en els seus models però els seus fonaments segueixen sent els mateixos.

Un dels *toolkits* més utilitzats en el camp de la traducció automàtica per a la creació de sistemes competitius, MOSES [13], utilitza aquesta aproximació. Hem entrenat i avaluat un sistema de traducció amb aquesta eina, per establir un *baseline* amb què comparar els resultats obtinguts al llarg del treball.

1.4.2 Neural MT

Els sistemes de *neural MT* no estan formats per una combinació de diversos models, sinó que aborden directament el problema de la traducció amb un únic model de xarxa

neuronal. Aquest model, pot constar de diferents parts, però s'entrenen totes alhora amb un únic criteri d'entrenament (sistema *end-to-end*).

Els primers sistemes de *neural MT* funcionaven construïnt un vector que representa tot el significat de la frase d'entrada, i a partir d'aquest vector s'obté la traducció. Tenim, així, un model d'*encoder-decoder*. Aquesta forma de treballar està representada en la figura 1.1.

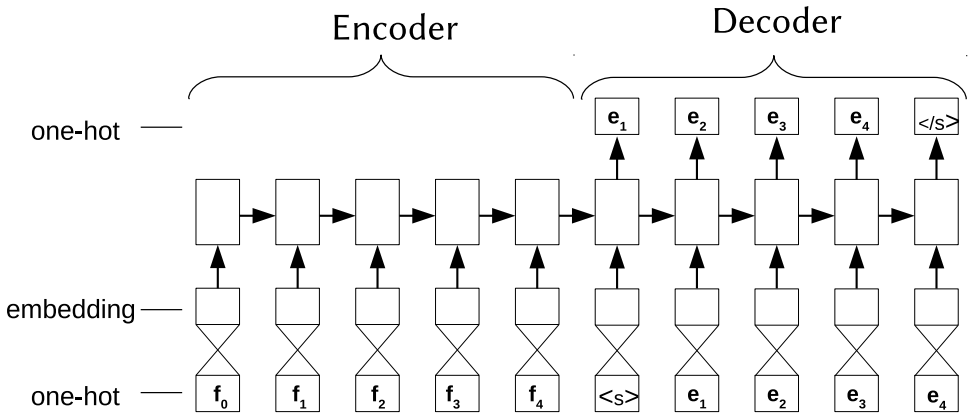

Figura 1.1: Sistema bàsic de *neural MT*

Com que treballem amb seqüències, l'estructura natural són les xarxes recurrents (RNN, de *Recurrent Neural Network*): la RNN *encoder* llegeix l'entrada, i en acabant li passem l'estat com a inicialització de la RNN *decoder*. Aquest estat ha de representar tot el significat de la frase d'entrada en un únic vector. Tant l'entrada com l'eixida es representen com una seqüència de vectors, on cada vector és una paraula. Amb una capa d'*embedding* es converteix cada paraula en la seua representació en un espai continu.

L'eixida del decoder, en cada instant de temps t és un vector de probabilitats que indica com de probable és cada paraula del vocabulari en eixa posició (e_t). L'entrada del decoder és, en cada cas, la paraula emesa en $t-1$.

Encara que funcionen sorprenentment bé, aquests sistemes tenen les seues limitacions, sobretot per a les frases més llargues, on el vector no és capaç de representar tot el contingut i a més el gradient va decaient al llarg de la xarxa recurrent, i per molt que utilitzem LSTMs això fa que el significat de les primeres paraules que entren a l'*encoder* difícilment arribarà al final del *decoder*. Per tant, aquests sistemes de *neural MT* serveixen com a idea base però no són capaços de produir resultats ni tan sols semblants als dels sistemes *phrase-based*.

Per resoldre aquest problema, es va introduïr el mecanisme d'atenció [5], que permet construir un vector de context c_t per a cada instant t a partir d'una suma ponderada dels estats interns de l'*encoder*. Aquest c_t és utilitzat per a predir la següent paraula. Aquest esquema de treball està reflexat en la figura 4.1

Segons aquesta idea, l'*encoder* pot ser unidireccional o bidireccional. En la figura

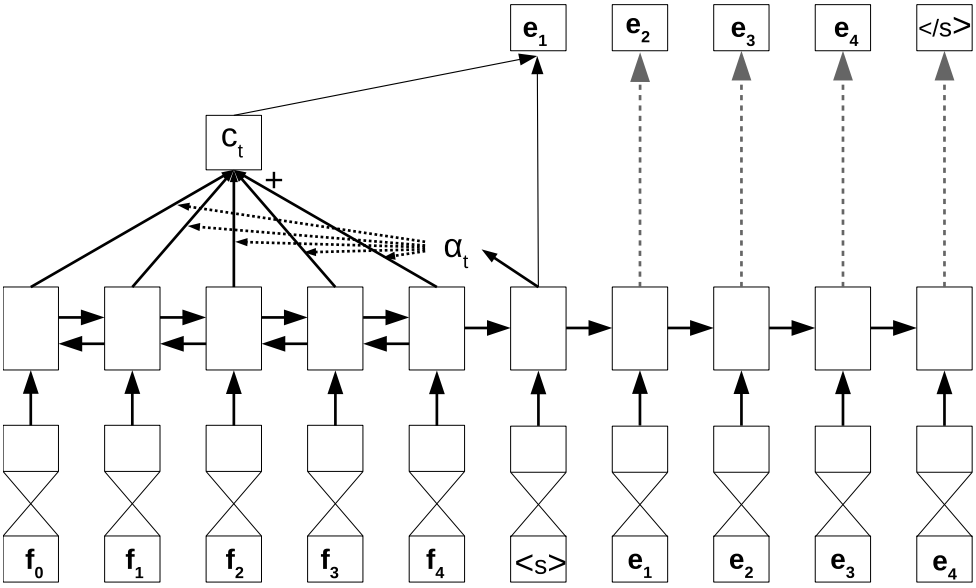

Figura 1.2: Mecanisme d'atenció per a *neural MT*

està representat el cas bidireccional, que hem usat en aquest TFM. Un *encoder* bidireccional és capaç de codificar informació de tota la frase en cada estat intern; la idea és que represente en major grau el significat i context de cada paraula.

Una volta feta la codificació, la decodificació es fa paraula per paraula. Cada paraula en l'eixida correspon a un estat del *decoder*, que s'obté de la següent manera: es puntuen els estats de l'*encoder* i es normalitzen, per obtindre els pesos α_t amb què es fa la suma ponderada dels estats de l'*encoder*. El resultat és el vector de context c_t. Això permet focalitzar l'"atenció" de la decodificació sobre una part concreta de l'entrada. A partir del vector de context c_t es calculen les probabilitats de cada paraula del diccionari. Aquest procés està descrit amb major detall en la secció 4.2, incloent com es calculen els pesos α_t i com es calcula la probabilitat de cada paraula a partir de c_t i l'estat corresponent del *decoder*.

Els resultats dels sistemes de *neural MT* amb mecanisme d'atenció són tan bons que en els últims anys s'han convertit en els sistemes que millor funcionen ara mateix. Aquest TFM està basat sobre aquesta idea, amb la intenció de millorar els sistemes de traducció del grup MLLP.

1.5 Eines informàtiques

Per realitzar aquest TFM, les eines informàtiques que s'han utilitzat són:

- **Giza++**[19]. Una eina de còmput dels alineaments entre frases de dos llengües distintes, necessària per entrenar sistemes *phrase-based*.

- **Moses**[14]. Un *toolkit* lliure per a l'entrenament de sistemes de traducció automàtica, amb focus sobre els sistemes *phrase-based*. Dins de Moses també tenim l'*script* `multi-bleu.perl`, que calcula la mètrica BLEU de qualitat de les traduccions automàtiques, que abordem en la següent secció.

- **Tensorflow**[3]. Una biblioteca de codi obert per a l'aprenentatge automàtic, molt utilitzada per a xarxes neuronals artificials. Els sistemes de *neural MT* utilitzats en aquest TFM estan implementats en Tensorflow.

1.6 Avaluació de resultats: TER i BLEU

La qualitat de les traduccions automàtiques es pot avaluar de distintes formes. Les més convenients són les automàtiques, que se solen basar en corpus paral·lels. En aquest treball utilitzem les més conegudes: TER i BLEU.

El TER [26] (*Translation Error Rate*) s'inspira en la distància de Levenshtein i mesura la quantitat d'edicions necessàries per convertir l'eixida del sistema en la referència. El TER està definit per a una o múltiples referències, i en aquest cas es pren la distància d'edició a la referència més propera. El TER es calcula així:

$$\text{TER} = \frac{\text{quantitat d'edicions}}{\text{mitjana de paraules en la referència}} \tag{1.4}$$

El BLEU [20] (*Bilingual Evaluation Understudy*) és una mètrica que intenta aconseguir una alta correlació amb les mesures de qualitat subjectives fetes per humans. Està pensat per extreure una aproximació de la qualitat en tot el corpus: els seus resultats no s'han de tindre en compte frase a frase. El BLEU és un valor entre 0 i 1, on un resultat igual a 1 indica que el resultat és idèntic a la referència. La mètrica és una modificació de la precisió (fracció de paraules en l'eixida que estan en la referència) tenint en compte no només paraules sinó també n-grames.

Capítol 2

Conjunts de dades utilitzats

En aquest capítol es presenten els dos conjunts de dades que s'han utilitzat per a entrenar i avaluar els sistemes de traducció automàtica. Aquests dos corpus són el WMT per a Alemany→Anglés i el poliMedia per a Castellà→Anglés. Ambdós són corpus paral·lels, és a dir tenen cada frase en les dues llengües; recordem que tot el treball se circumscriu al camp de l'aprenentatge supervisat.

2.1 WMT: Alemany→Anglés

El WMT (*Workshop on Machine Translation*) és una conferència anual internacional de traducció automàtica estadística. Originalment un *workshop*, l'any 2016 es va convertir en conferència mantenint les sigles. Les dades que publica per a entrenament i avaluació són públiques, i són les utilitzades en aquest TFM.

De les tasques que proposa WMT, hem abordat la de traducció en l'àmbit de les notícies en el parell de llengües Alemany→Anglés. Cada any es publica un nou conjunt d'avaluació en aquesta tasca, des del 2013. El conjunt d'entrenament és el mateix. Les dades vénen ja tokenitzades tant en el conjunt d'entrenament com els conjunts d'avaluació.

La taula 2.1 mostra estadístiques bàsiques sobre aquests conjunts de dades: el d'entrenament, que rep el nom de `train`, i els d'avaluació, que reben el nom de `newstest[any]`. Aquestes estadístiques són: per una banda, la quantitat de frases del conjunt, i per l'altra, la mitjana de paraules per frase en cada llengua, que pot ser diferent perquè una frase no té necessàriament la mateixa quantitat de paraules en una llengua i en una altra.

En la taula 2.1 podem comprovar que la quantitat de frases del conjunt d'entrenament és molt superior, per ordres de magnitud, a la dels conjunts d'avaluació. Això no és d'estranyar, perquè per a l'entrenament, com més dades millor, mentre que per a l'avaluació amb una quantitat suficient ja n'hi ha prou. L'altre fet remarcable és que la quantitat de paraules per frase és consistentment major en anglés que en alemany. Això s'explica perquè l'alemany és una llengua aglutinant, i per tant algunes de les seues paraules estan formades per dues o més que en anglés van per separat.

Taula 2.1: Estadístiques bàsiques sobre el corpus Alemany-Anglés de WMT

		Paraules per frase (mitjana)	
Conjunt	# de frases	Alemany	Anglés
`train`	4 472 674	24.2	25.5
`newstest2013`	3 000	21.4	21.9
`newstest2014`	3 003	21.5	23.0
`newstest2015`	2 169	20.7	21.9
`newstest2016`	2 999	21.5	21.9

2.2 poliMedia: Castellà→Anglés

Els vídeos poliMedia són vídeos educatius de la UPV. Si bé l'interés d'entrenar un sistema amb dades de competicions i conferències internacionals té sentit en el marc de traducció automàtica com a objectiu científic, l'interés d'entrenar un sistema per al domini de vídeos educatius de la UPV és més directe i aplicat, amb la intenció de generar subtítols per als vídeos poliMedia de la UPV.

El grup MLLP ja ha desenvolupat sistemes de traducció automàtica per a poliMedia, com descriu el capítol 3, però aquests sistemes han estat sempre *phrase-based* i aquest TFM explora la possibilitat de fer servir sistemes NMT per substituir-los. Per tant l'ús d'un conjunt de dades del domini específic de poliMedia és natural, i de fet s'ha utilitzat el mateix corpus per a entrenar els sistemes *phrase-based* que aquests NMT que han estat l'objecte del TFM.

La taula 2.2 mostra les mateixes estadístiques sobre aquest corpus que sobre l'anterior. Els tres conjunts són: `train`, per a entrenament, `dev` per a exploració de paràmetres i `test` per a avaluació.

Taula 2.2

		Paraules per frase (mitjana)	
Conjunt	# de frases	Castellà	Anglés
`train`	6 005 688	24.0	22.2
`dev`	1 401	27.0	27.6
`test`	1 139	28.3	28.2

En la taula 2.2 podem observar que, tal i com passava amb el cas anterior, hi ha moltes més dades d'entrenament que d'avaluació, per ordres de magnitud i pels mateixos motius. També podem comprovar que, en contrast, la mitjana de paraules per frase no és consistentment superior en cap de les dues llengües.

Capítol 3

Avaluació sobre casos d'ús reals

L'ús de recursos oberts com els MOOC (*Massive Open Online Courses*) no han parat de créixer ràpidament des del 2011, amb més de 35 milions d'estudiants i 4000 cursos que s'oferien a principis de 2016, xifres que dupliquen les del 2015 [25]. Per tal que els MOOC, o els OER (*Open Educational Resources*) arriben a una audiència internacional, haurien d'estar en format multilingüe. Per això un dels focus de recerca del grup MLLP és produir vídeos educatius multilingües automàticament.

Com ja hem dit en la introducció, en la secció 1.3, dins del marc de treball anterior relacionat amb aquest TFM cal parlar d'un article publicat [17] que se centra en avaluar com s'han aplicat els sistemes de transcripció i traducció a casos d'ús reals, que han estat els projectes europeus transLectures [2] i EMMA [1]. En aquest capítol resumim l'article i els seus resultats.

Per generar subtítols en la llengua original, s'utilitzen els sistemes de transcripció, o ASR (*Automatic Speech Recognition*). Una possible conclusió és que la tecnologia ASR actual ja permet generar subtítols automàtics d'alta qualitat, pràcticament llestos per publicar, sempre que s'utilitze la tecnologia estat-de-l'art i s'adapte a la tasca particular. Aquesta adaptació permet aconseguir unes millores de vora un 40% sobre els proveïdors de referència (concretament, YouTube).

Per obtindre subtítols en altres llengües, s'utilitzen els sistemes de traducció, o MT (*Machine Translation*). Aquesta aproximació s'ha provat amb èxit en els dos projectes estudiats; una conclusió és que la qualitat d'aquests sistemes és prou bona per ser utilitzada conjuntament amb la post-edició. Això vol dir que és més eficient revisar les traduccions automàtiques que traduir a mà des de zero. L'adaptació dels sistemes a la tasca també millora molt la qualitat de les traduccions: en comparació amb la referència (Google Translate), aquests sistemes adaptats han produit millores relatives de vora un 20%.

Per obtindre subtítols multilingües, el procés és el següent: s'obtenen transcripcions originals mitjançant ASR, es revisen manualment aquestes transcripcions, es tradueixen automàticament amb MT, i després es revisen les traduccions. Com més

bons siguen els sistemes d'ASR i MT, més ràpid és el procés i més bons els resultats.

Aquest capítol està estructurat cinc seccions. La secció 3.1 presenta els dos casos d'estudi. La secció 3.2 exposa mesures objectives de la qualitat dels resultats dels sistemes de transcripció i traducció en aquests casos. La secció 3.3 relaciona la qualitat dels reaultats amb el temps de revisió manual. La secció 3.4 descriu l'impacte dels sistemes sobre els casos d'estudi, amb estadístiques detallades. Per últim, la secció 3.5 tanca el capítol amb les conclusions més rellevants de l'estudi.

3.1 Casos d'estudi

Aquesta secció introdueix els dos casos d'estudi abordats en l'avaluació: EMMA i TransLectures.

El projecte europeu **EMMA** (febrer 2014 - juliol 2016) va involucrar 12 socis, amb un total de més de 30 MOOCs multilingües en diferents temes. El multilingüisme és una característica particular de la plataforma EMMA, ja que proporciona transcripció i traducció automàtiques per a tots els continguts de vídeo i text. Això inclou transcripció en 7 llengües (anglés, italià, castellà, holandés, francés, portugués i estonià) i traducció a l'anglés, el castellà i l'italià. Els recursos automàtics són revisats pels professors abans de ser publicats. La major part dels cursos són oferits en bilingüe (llengua original més anglés) o trilingüe (amb el castellà, francés o italià).

El projecte **TransLectures** treballa sobre el repositori **UPV media** [28]. Aquest repositori és un servei per crear, gestionar i difondre vídeos educatius, que reben el nom de poliMedia. Els poliMedia proporcionen una breu visió general d'un tema concret i tenen una duració mitjana de deu minuts. La taula 3.1 mostra estadístiques bàsiques del repositori agrupades per llengua, descartant llengües molt minoritàries.

Taula 3.1: Hores de vídeo de poliMedia per llengua

Llengua	Vídeos	Hores	Locutors
Català	434	52	80
Anglés	1221	173	203
Castellà	15013	2709	1572

3.2 Qualitat de transcripció i traducció

Per a desenvolupar els sistemes automàtics, es van utilitzar dues eines: TLK per a ASR, i Moses per a MT. Aquests sistemes van ser adaptats mitjançant les tècniques descrites en [16] i [4]. Aquests sistemes s'han integrat en els casos d'ús fent servir TLP [9]. El resultat d'aquest treball es pot provar a través de la *Transcription and Translation Platform* (TTP) [23].

La qualitat de la transcripció es va avaluar amb el WER (veure secció 1.2). La taula 3.2 mostra la quantitat de vídeos, duració i WER (± desviació estàndard) per a

cada llengua transcrita. Els vídeos en castellà i anglés són del repositori *UPV media*, i els vídeos en italià, neerlandés i francés són de MOOCs en la plataforma EMMA. Hi ha un gran nombre de vídeos en castellà degut a que la gran majoria dels vídeos de *UPV media* són en castellà.

Taula 3.2: Vídeos, duració i WER (± desviació estàndard) per llengua

Llengua	Vídeos	Hores	WER
Castellà	207	24.7	18.4 ± 6.4
Italià	13	1.2	25.7 ± 6.4
Anglés	25	3.5	21.9 ± 8.5
Neerlandés	11	6.9	29.4 ± 9.2
Francés	21	2.1	23.2 ± 8.3

La duració mitjana dels vídeos, excepte els que estan en neerlandés, és de menys de 10 minuts. Els vídeos en neerlandés duren més de 35 minuts de mitjana i el format de la presentació és diferent de la de les altres llengües: són entrevistes entre, normalment, dos persones; mentre que en la resta hi ha un únic locutor parlant front a la càmera.

Dels resultats en la taula 3.2, podem veure que la qualitat de la transcripció és més bona (menys WER) en castellà, seguit per anglés i francés, amb tots ells per sota del 25%. L'italià està just per sobre del 25% i el neerlandés té el màxim WER que segueix per sota del 30%. Un dels motius d'aquests resultats en neerlandés pot ser la presència de dos locutors, que dificulta l'adaptació acústica.

La qualitat de la traducció automàtica es va avaluar amb el TER (veure secció 1.6). La taula 3.3 mostra el nombre de vídeos, duració i TER (± desviació estàndard) per a cada parell de traducció. Tots els vídeos es van traduir automàticament i es van revisar. Els vídeos en castellà són part de *UPV media* i els van revisar els professors. Els vídeos Anglés→Castellà són de dos MOOCs d'EMMA originalment en italià, després traduïts a l'anglés i per al treball traduïts al castellà. De la mateixa manera, els vídeos Anglés→Italià són de MOOCs d'EMMA originalment en castellà. En total, en el conjunt d'avaluació hi ha quatre MOOCs dispibles en trilingüe (italià, anglés i castellà). Els vídeos traduïts des del neerlandés i francés també són de MOOCs d'EMMA.

Dels resultats de la taula 3.3, queda clar que la qualitat és prou bona per a que siguen útils per a la post-edició, excepte el parell Francés→Anglés. La qualitat de la traducció Francés→Anglés va resultar pitjor d'allò esperat, per dos motius principals. Primer: el procediment de dos passades utilitzat pels revisors fa que les traduccions finals siguen diferents; i segon: creiem que el sistema de MT no es va adaptar bé al domini específic dels cursos en francés.

El treball també va incloure una comparació entre els sistemes adaptats i els dels principals proveïdors comercials, en concret els dels subtítols automàtics de YouTube i els dels sistemes automàtics de Google Translate. Per a això, es va crear un altre conjunt d'avaluació amb vídeos dels MOOCs d'EMMA. La taula 3.4 mostra estadístiques

Taula 3.3: Vídeos, duració i TER (± desviació estàndard) per parell de traducció

Parell de traducció	Vídeos	Hores	TER
Castellà → Anglés	101	10.8	33.2 ± 14.4
Anglés → Castellà	29	2.5	27.0 ± 19.9
Italià → Anglés	14	1.6	37.5 ± 8.5
Anglés → Italià	121	6.5	33.8 ± 8.0
Neerlandés → Anglés	5	3.5	30.7 ± 13.4
Francés → Anglés	8	0.9	58.9 ± 5.2

de quantitat de vídeos, hores i WER obtinguts per TTP i pels subtítols automàtics de YouTube.

Taula 3.4: Vídeos, duració i WER per llengua de TTP i de YouTube

Llengua	Vídeos	Hores	TTP	YouTube
Castellà	23	3.5	14.8	22.5
Italià	3	4.0	17.1	31.6
Anglés	9	0.4	39.2	65.9
Neerlandés	2	1.1	24.5	41.1
Francés	18	2.3	20.6	32.0

Dels resultats de la taula 3.4 podem concloure que el WER de YouTube és més alt que el de TTP per a totes les llengües, i concretament la millora és de quasi un 70% de mitjana. El principal motiu és que YouTube fa servir sistemes ASR de propòsit general, mentre que els sistemes integrats en TTP estan adaptats automàticament a la tasca.

Aquest mateix conjunt d'avaluació es va ampliar per a comparar traduccions. La taula 3.5 mostra, agrupat per parell de traducció, nombre de vídeos, duració i el TER obtingut amb TTP i Google Translate. Són tot casos de vídeos que van ser primer transcrits i revisats per després ser traduïts. Per als casos de Francés→Anglés i Castellà→Francés es van utilitzar els mateixos de la taula 3.4.

Una conclusió en general de la taula 3.5 és que els sistemes de Google Translate produeixen més error que els de TTP, excepte per al parell Francés→Anglés, on mostren resultats semblants. De mitjana, les xifres de TER de Google són un 14% més altes. Com en el cas de la transcripció, els sistemes de TTP tenen millors resultats gràcies a l'adaptació al domini.

Taula 3.5: Vídeos, duració i TER per llengua de TTP i Google Translate

Parell de traducció	Vídeos	Hores	TTP	Google
Castellà → Anglés	250	13.9	33.9	44.3
Anglés → Castellà	9	0.4	35.8	42.4
Italià → Anglés	11	1.1	33.4	39.2
Anglés → Italià	81	5.4	39.7	43.3
Neerlandés → Anglés	2	1.2	42.5	45.0
Francés → Anglés	18	2.3	52.8	52.6

3.3 Temps de revisió

El temps que cal dedicar per post-editar les transcripcions i traduccions automàtiques es mesura en termes de RTF (veure secció 1.2). Si un vídeo dura 6 minuts i es tarda una hora (60 minuts) en revisar-lo, el RTF és de 10. En general, la RTF per a la transcripció manual és de 10 [24][18].

Per a revisió manual de traducció automàtica, és complicat donar una xifra concreta del temps que hi cal dedicar, però podem aproximar un mínim de 22.5 RTF i una mitjana de 30 RTF, tenint en compte que els locutors pronuncien de mitjana 150 paraules per minut i que com a molt un especialista pot traduir entre 400 i 1000 paraules per hora [21].

La taula 3.6 mostra els WER i RTF mitjans per cada llengua, i models de regressió per predir el RTF en funció del WER. Es van provar tres models de regressió: lineal, arrel quadrada i logarítmic. La taula 3.6 conté informació dels tres models només per al castellà; i la figura 3.1 mostra un gràfic de dispersió de RTF contra WER per a cada vídeo en castellà (representat en un punt), amb els tres models de regressió ajustats. Per a la resta de llenguatges mostrem només els detals del model logarítmic.

Taula 3.6: WER i RTF mitjans (± desviació estàndard) i models de regressió per llengua

Llengua	WER	RTF	Model	R^2	β	sig
Castellà	18.4	3.3 ± 1.2	WER	0.87	0.17	$< 10^{-15}$
			$\sqrt{\mathrm{WER}}$	0.90	0.78	$< 10^{-15}$
			$\ln \mathrm{WER}$	0.91	1.17	$< 10^{-15}$
Anglés	21.9	5.3 ± 1.7	$\ln \mathrm{WER}$	0.92	1.76	$< 10^{-14}$
Italià	25.7	3.9 ± 1.4	$\ln \mathrm{WER}$	0.90	1.20	$< 10^{-6}$
Neerlandés	29.4	5.8 ± 2.5	$\ln \mathrm{WER}$	0.85	1.75	$< 10^{-14}$
Francés	23.2	6.7 ± 0.8	$\ln \mathrm{WER}$	0.98	2.17	$< 10^{-15}$

A partir d'aquests resultats es pot concloure que disposar de transcripcions automàtiques redueix el temps humà dedicat a la transcripció entre un terç i dos terços: el

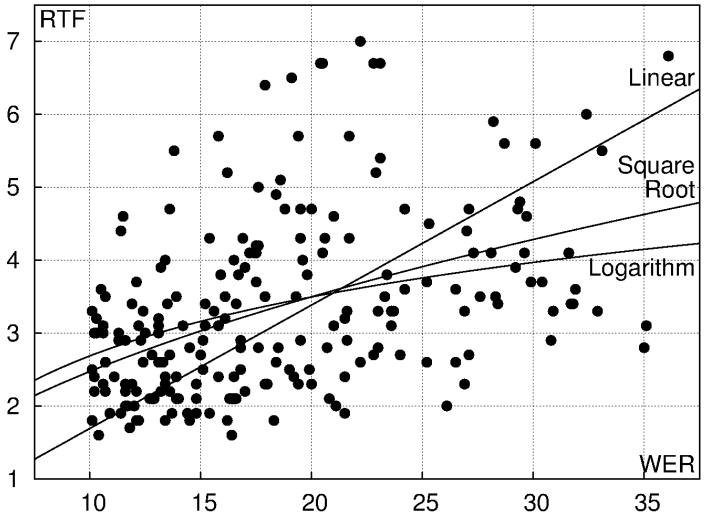

Figura 3.1: RTF vs. WER per als vídeos en castellà, i models de predicció

RTF està entre 3 i 7 si partim de transcripcions automàtiques comparat amb el RTF de 10 si no les tenim. Una altra conclusió és que el millor model predictiu de regressió és el logarítmic en termes de significació estadística. Per a totes les llengües, l'ajust és estadísticament significant($sig < 10^{-4}$) i una gran quantitat de la variabilitat s'explica pel model ($R^2 \geq 0.85$). Aquest model explica per què els usuaris tendeixen a ignorar les transcripcions automàtiques quan el WER és massa alt i prefereixen partir de zero.

La taula 3.7 mostra, per parell de traducció, TER i RTF mitjans, més models de regressió per a predir RTF en funció del TER. Els resultats estan mostrats de la mateixa forma que per a la transcripció, en la taula 3.7 i la figura 3.2

De forma pareguda a la transcripció, el primer resultat és que el temps de revisió es redueix aproximadament a un terç quan es disposa de traduccions automàtiques, excepte per al cas de Francés→Anglés. A més, el millor model de regressió torna a ser el logarítmic. La part de variabilitat de les dades que explica el model (valor R^2) no és tan alta com en les transcripcions. Això es pot explicar degut a que la tasca de traducció és més complexa que la de transcripció.

Fins ara, hem comprovat que per a predir RTF a partir de l'error d'un sistma en una llengua concreta —o un parell, en el cas de traducció—, el millor model de regressió que es pot ajustar és el logarítmic.

La següent qüestió natural a abordar és si un únic model pot predir l'RTF independentment de la llengua. La figura 3.3 és el resultat de la regressió sense separar les llengües: el gràfic de dalt mostra RTF i WER per a totes les llengües, amb un únic model de regressió logarítmica per a totes les dades, i el de baix és l'equivalent però per a TER i parells de llengües. Per al WER, el model logarítmic ajustat

Taula 3.7: TER i RTF mitjans (± desviació estàndard) i models de regressió per llengua

Llengua	WER	RTF	Model	R^2	β	sig
Castellà→Anglés	33.2	9.1 ± 4.9	TER	0.75	0.25	$< 10^{-15}$
			$\sqrt{\text{TER}}$	0.80	1.61	$< 10^{-15}$
			$\ln \text{TER}$	0.80	2.71	$< 10^{-15}$
Anglés→Castellà	27.0	7.8 ± 4.9	$\ln \text{TER}$	0.82	2.67	$< 10^{-11}$
Italià→Anglés	37.5	11.3 ± 4.2	$\ln \text{TER}$	0.89	3.15	$< 10^{-7}$
Anglés→Italià	33.8	9.6 ± 5.3	$\ln \text{TER}$	0.77	2.76	$< 10^{-15}$
Neerlandés→Anglés	30.7	9.5 ± 3.9	$\ln \text{TER}$	0.91	2.89	$< 10^{-2}$
Francés→Anglés	58.9	23.2 ± 8.0	$\ln \text{TER}$	0.90	5.67	$< 10^{-4}$

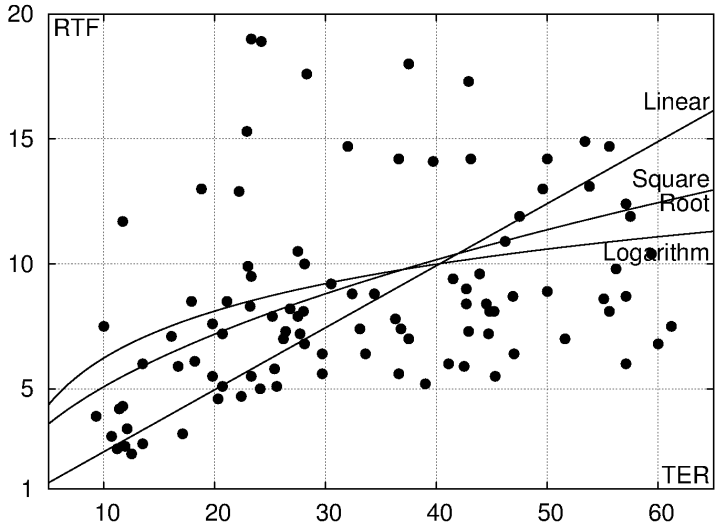

Figura 3.2: RTF vs. TER per als vídeos en castellà, i models de predicció

($R^2 = 0.87, \beta = 1.34$) és estadísticament significatiu ($sig < 10^{-15}$). Això confirma que el RTF depén molt de la qualitat de la transcripció, molt més que de la llengua en particular.

Per a traducció, on hem utilitzat el TER, el model logarítmic ajustat ($R^2 = 0.78, \beta = 2.90$) també és estadísticament significatiu ($sig < 10^{-15}$). Per tant, podem confirmar també que el RTF depén més de la qualitat de la traducció automàtica que del parell de llengües en concret.

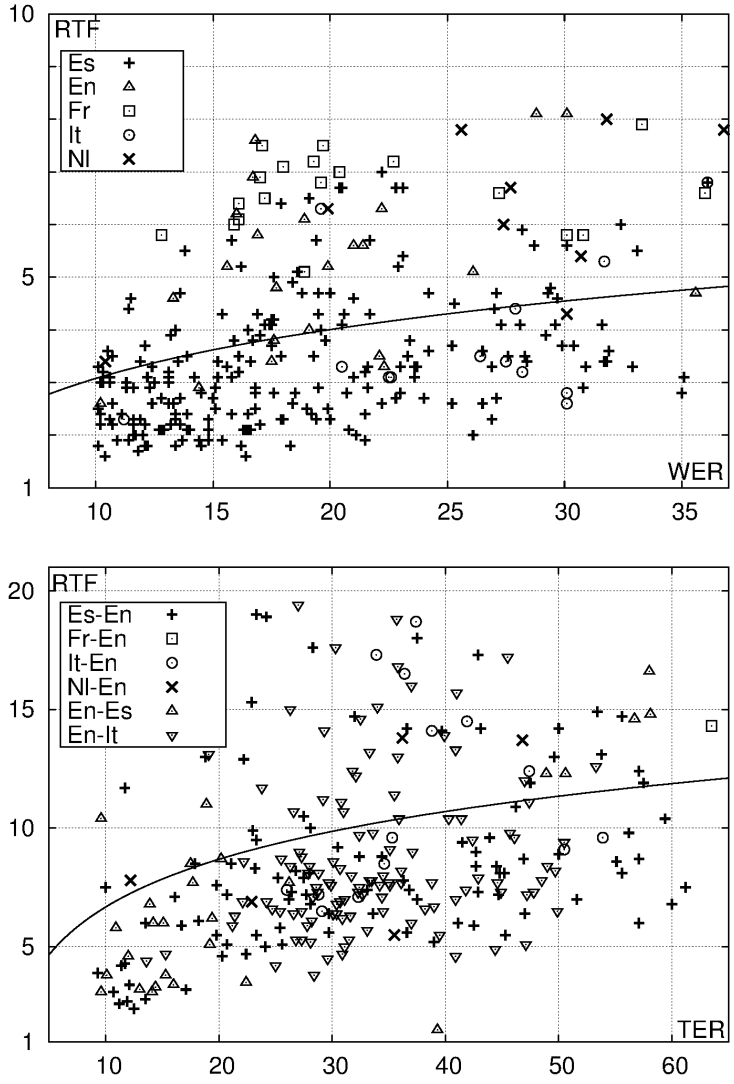

Figura 3.3: RTF vs. TER i RTF vs. WER

3.4 Impacte en els casos d'estudi

Aquesta secció presenta estadístiques recollides al llarg de dos anys sobre el consum de dades multilingües en els dos casos d'estudi descrits en la secció 3.1: la plataforma EMMA i el repositori *UPV media*.

Per a la plataforma EMMA, la taula 3.8 mostra la quantitat d'estudiants nadius i no nadius inscrits en els MOOCs agrupats per llengua de curs, i l'increment relatiu

en quantitat d'estudiants gràcies a les traduccions. Els estudiants no nadius poden seguir els cursos gràcies a la integració de TLP dins d'EMMA.

Taula 3.8: Inscripcions en MOOCs d'EMMA

Llengua	Estudiants nadius	Estudiants no nadius	Increment relatiu
Castellà	161	547	340
Francés	983	879	89
Italià	609	259	43
Neerlandés	501	104	21
Anglés	351	27	8

Els resultats de la taula 3.8 vénen donats en ordre decreixent d'increment relatiu d'estudiants. Els millors resultats són els dels cursos originalment en castellà, seguits per 161 alumnes castellanoparlants. Com que aquests cursos també s'oferien en anglés i italià, 547 alumnes no castellanoparlants s'hi van inscriure, xifra que representa el 340% dels alumnes castellanoparlants. Els pitjors resultats són els dels cursos en anglés, que van ser traduits al castellà i això va provocar només un augment del 8% en inscripcions, cosa que pot ser deguda a que molts alumnes no realment nadius poden seguir els cursos originals en anglés. En conjunt, podem dir que les versions traduïdes dels MOOCs van aconseguir atraure alumnes no nadius, augmentant les inscripcions en un 70% en total.

Per al repositori *UPV media*, la taula 3.9 mostra el nombre de reproduccions de vídeos i de subtítols, en milers, per llengua des de juny de 2015 fins a maig de 2016.

Taula 3.9: Milers de reproduccions de vídeo i de subtítols, per llengua

		Repr. de subtítols	
Llengua	Reproduccions	Castellà	Anglés
Castellà	629	6.9	1.1
Anglés	63	1.3	0.5
Total	692	8.2	1.6

La conclusió principal de la taula 3.9 és que, de mitjana, els subtítols es van gastar en l'1.4% de les reproduccions de vídeo. No obstant això, aquesta xifra representa una gran quantitat de reproduccions: quasi 700K en un any. És interessant remarcar que un 2.5% dels vídeos en anglés es van veure amb subtítols, en contrast amb un 1.3% dels vídeos en castellà. Això és perquè la gran majoria dels estudiants de la UPV entenen el castellà sense problemes, i no tan bé com l'anglés.

A banda de la disponibilitat dels subtítols, les transcripcions han permés indexar i cercar dins del repositori per paraules concretes: l'eina de cerca del repositori permet als estudiants trobar vídeos on es pronuncia una paraula en concret.

3.5 Conclusions

En aquest capítol, hem explicat gran part de l'experiència guanyada de produir subtítols multilingües de baix cost i de qualitat publicable, per a MOOCs i OER. A més de descriure els sistemes, eines i components d'integració, hem proporcionat una avaluació a fons dels resultats obtinguts des de tres punts de vista: la qualitat de les transcripcions i traduccions automàtiques a partir de sistemes adaptats, el temps humà requerit per revisar-les, i l'impacte que ha tingut en una plataforma MOOC la disponibilitat de subtítols multilingües per a un gran repositori de vídeos educatius.

La qualitat dels resultats automàtics ha estat en gran part per sota de 25% WER per a transcripció i 50% TER per a traducció. Això significa que són utils per a editar-los i obtindre subtítols de qualitat publicable, val més la pena això que partir de zero. Els resultats dels nostres sitemes adaptats han estat més bons que els obtinguts a partir d'eines proporcionades per proveïdors comercials: un 38% millors que els subtítols de YouTube i un 17% respecte a Google Translate.

Pel que fa al procés de revisió, hem vist que un professor pot estalviar-se des del 30% fins al 70% del temps que li dedicaria a transcriure i entre un 25% i un 75% del temps que li dedicaria a traduir, respecte a partir de zero. Hem proposat un sistema de regressió per a inferir el RTF en funció de l'error, WER o TER depenent de si és transcripció o traducció.

La disponibilitat de subtítols multilingües en els vídeos va tindre un gran impacte en els casos d'estudi. En la plataforma EMMA, va provocar un augment de les inscripcions d'un 70%. En el repositori *UPV media*, a banda de millorar l'accessibilitat dels vídeos per a una minoria de persones amb discapacitat auditiva i estudiants no nadius, els subtítols van permetre desenvolupar funcionalitats d'indexació i cerca de vídeos.

Capítol 4

Sistemes de traducció

Aquest capítol explica els sistemes de traducció entrenats: *phrase-based* (secció 4.1 i *neural MT* (secció 4.2), i conté també una comparativa entre les dues aproximacions (secció 4.2) en termes de qualitat de resultats i d'eficiència espacial i temporal dels sistemes entrenats.

4.1 Sistemes *phrase-based*

Els sistemes *phrase-based* són l'evolució dels models basats en paraules, i consisteixen en un conjunt de models la combinació dels quals distribueix la massa de probabilitat entre totes les possibles traduccions. Com el seu propi nom indica, el model més característic d'aquests sistemes és el model de traducció de frases. El concepte de frase ací no és tota la seqüència de paraules d'entrada o d'eixida, sinó més bé un tros de l'entrada o de l'eixida. Així, l'entrada se segmenta en "frases" d'una, dos o tres paraules aproximadament, que es tradueixen per blocs i es reordrenen.

Les idees bàsiques dels sistemes *phrase-based* han estat visitades en la secció 1.4 de la introducció, i en aquesta secció entrarem en més detall en el seu funcionament i en la implementació utilitzada en aquest TFM. Per evitar conflictes semàntics, al llarg d'aquesta secció la paraula "frase" farà referència al segment de l'entrada o eixida i no a la frase en conjunt.

4.1.1 Models

Els sistemes *phrase-based* són una aproximació generativa al problema de la traducció. Si considerem que l'objectiu dels models és representar $p(e|f)$, on e és la frase de destí i f és la frase d'origen; aquesta probabilitat es descomposa de la següent forma:

$$p(e|f) = \frac{p(e,f)}{p(f)} = \frac{p(f|e)p(e)}{p(f)} \tag{4.1}$$

Com que ens interessa el màxim en e d'aquestes probabilitats, i $p(f)$ és independent de e:

$$\arg\max_e p(e|f) = \arg\max_e p(f|e)p(e) \tag{4.2}$$

On $p(f|e)$ rep el nom de model de traducció i $p(e)$ rep el nom de model de llenguatge. En els sistemes *phrase-based* entren en joc distints models, entre els quals aquests dos, a l'hora d'avaluar una eixida:

Model de llenguatge

El model de llenguatge indica la probabilitat *a priori* d'una frase en la llengua donada: $p(e)$. Serveix per assegurar que l'eixida és una frase amb sentit en la llengua de destí, informant sobre l'elecció de paraules i el seu ordre.

Hi ha molts tipus de models de llenguatges; en aquesta memòria ens centrarem sobre els n-grames i els models de xarxes neuronals. Els dos tipus de model subdivideixen aquesta probabilitat seguint la regla de la cadena:

$$p(e_0^n) = \prod_{t=1}^{n} p(e_t|e_0^{t-1}) \tag{4.3}$$

On e_0 és un símbol especial que indica inici de frase. És a dir, la subdivisió és en la probabilitat d'una paraula donada la seua història, o dit d'una altra manera la probabilitat d'una paraula, e_t sabent la seqüència de paraules que l'han precedit, e_0^{t-1}.

Els models de n-grames es basen en una de les idees més bàsiques que es poden aplicar al problema: el recompte. Compten quantes vegades, donada eixa història, la següent paraula és justament e_t, i la proporció de les vegades que resulta ser eixa paraula és la probabilitat. L'únic problema és que per a històries prou llargues potser no hi ha aparicions, i en tot cas segur que per moltes dades de què es dispose no és possible aconseguir bones estimacions.

L'aproximació de n-grames consisteix a fer els recomptes sobre grups de n paraules. Així, un model de trigrames fa els recomptes sobre grups de tres paraules, i aproxima tota la història amb només les últimes dues paraules:

$$p(e_t|e_0^{t-1}) \approx p(e_t|e_{t-2}^{t-1}) \tag{4.4}$$

L'assumpció que la probabilitat d'una paraula depén només de les n paraules prèvies és una assumpció Markoviana: hi ha un nombre finit d'estats, format pels grup de $n-1$ paraules possibles, i la probabilitat depén únicament de l'estat i no de tota la història. En la pràctica, els models d'n-grames consisteixen en una ponderació de models d'unigrames, bigrames, trigrames, quatrigrames... fins al límit en què es decidisca.

Els models d'n-grames s'entrenen simplement amb el recompte de les aparicions, representat en l'equació 4.5, i per tant són un model dels més simples que es puguen implementar. La clau del seu èxit està en el suavitzat d'aquest recompte, tema que no abordarem en aquesta memòria per quedar fora de l'àmbit del TFM però un bon article de referència el podeu trobar en [8].

$$p(e_t|e_{t-2}^{t-1}) \approx \frac{c\left(e_{t-2}, e_{t-1}, e_t\right)}{\sum_w c\left(e_{t-2}, e_{t-1}, w\right)} \tag{4.5}$$

L'altre model de llenguatge és el model de xarxes neuronals, concretament xarxes recurrents. Les xarxes recurrents han demostrat ser especialment útils com a model de llenguatge, sobretot des de l'aparició de les LSTM. Les LSTM són una variant de les xarxes recurrents que és capaç d'aprendre dependències de llarg termini, i estan abordades amb major detall en la següent secció, que se centra en *neural MT*. Cal recordar que el fet d'utilitzar xarxes, o components amb xarxes, en algun o alguns dels models d'un sistema *phrase-based* no fa que deixe de ser un sistema d'aquest tipus. De fet els sistemes *phrase-based* estat-de-l'art utilitzen molts components neuronals en els seus models.

Els models de llenguatge de xarxes recurrents no fan cap simplificació en les assumpcions, i representen directament tota la història e_0^{t-1} en el seu estat (c_t, en el cas de les LSTM — aneu a la secció 4.2.1 per a informació més detallada) en la mesura en què tinguen els paràmetres entrenats i la talla del vector ho permeta. L'estat en una xarxa neuronal recurrent és un estat continu i no discret, sense repectar per tant l'assumpció markoviana, fet que representa una diferència important respecte als n-grames.

En la pràctica s'utilitza una ponderació entre el model d'n-grames i el model LSTM, tots dos tenen aportacions rellevants (el suavitzat, per una part; i la dependència a llarg termini, per l'altra) que molt difícilment es poden tindre en compte descartant algun dels dos models.

Model de traducció

El model de traducció és l'essència d'un sistema *phrase-based* i d'on pren el seu nom. Aquest model també pren el nom de "model lèxic" i no és més que una taula on cada entrada té tres parts:

- f, Una seqüència d'una o més paraules en la llengua d'origen
- e, Una seqüència d'una o més paraules en la llengua de destí
- g, Un *score* o puntuació per a l'entrada, que pot ser qualsevol valor dels reals.

I aquestes entrades representen que f es tradueix per e amb la puntuació g. Com que g segons aquesta definició no té per què estar normalitzat, no podem parlar de probabilitat.

El model de traducció també s'entrena per mètode de recompte, on entren en joc els alineaments. Els alineaments proporcionen informació sobre quina paraula d'origen correspon a quina paraula de destí per a cada parell d'entrada-eixida. Aquests alineaments idealment es farien a mà, però la quantitat de dades ho fa impracticable. Hi ha eines informàtiques automàtiques especialitzades en el problema dels alineaments de corpus paral·lels. En la secció 1.5 de la introducció fem referència a una d'aquestes eines, Giza++, que és la que s'ha utilitzat en aquest TFM. El problema

del càlcul automàtic dels alineaments queda fora de l'àmbit d'aquest TFM i per tant no li dedicarem més discussió.

Tornant al mètode de recompte: bàsicament, si un conjunt de paraules e correspon sistemàticament a un conjunt f, s'afegeix una entrada (f, e, g) a la taula on:

$$g = \log \frac{c(e, f)}{c(e)} \tag{4.6}$$

Ací $c(e, f)$ representa les vegades que e s'ha correspost a f, i $c(e)$ representa les vegades que e apareix en les dades d'entrenament. Podeu comprovar que g es podria intepretar com a $p(f|e)$, en l'ordre invers a la traducció desitjada, seguint l'equació 4.2.

Aquest tipus de model de traducció és característic dels sistemes *phrase-based* i substitueix als models lèxics basats en paraules i als models de fertilitat, que indicaven a quantes paraules en la llengua de destí corresponia una paraula donada en la llengua d'origen.

Una característica d'aquest model de traducció és que defineix un conjunt de les eixides possibles, a partir d'una entrada. Aquest conjunt és definit pel concepte de derivació; que per explicar-lo haurem de definir primer $\mathcal{P}$, el conjunt de possibles frases d'eixida. Recordem que "frase" en aquest context és un segment d'oració. $\mathcal{P}$ es defineix a partir de l'entrada: per cada segment de l'entrada que estiga en la taula, la frase corresponent forma part de $\mathcal{P}$.

Una derivació $y = p_1 p_2 ... p_L$ és una seqüència finita de frases que ha de complir certes restriccions: cada p_k de y és un element de $\mathcal{P}$ per a l'entrada; i cada paraula de l'entrada es tradueix exactament una vegada.

Model de reordenació

El model de traducció, o model lèxic, no té informació sobre l'ordre de les frases en l'eixida. Això és perquè diferents llengües sovint tenen sistemes d'ordenar les paraules distints. El model de reordenació puntua formes de reordenar frases en el procés de traduir d'una llengua a una altra. Hi ha distintes formes d'aplicar aquest tipus de model, des de la més simple que és una restricció sobre la distància de les frases entre si en l'eixida, a un model que tinga en compte el parell de frases per saber quin tipus de continuïtat o discontinuïtat tenen respecte a les que l'envolten.

L'existència del model de reordenació es justifica mitjançant la descomposició del model de traducció, que en l'equació 4.2 representàvem com a $p(f|e)$:

$$p(f_1^n|e_1^n) = \prod_{i=1}^{n} \phi\left(f_i|e_i\right) \cdot d\left(\text{start}_i - \text{end}_{i-1} - 1\right) \tag{4.7}$$

On ϕ és una funció definida pel model lèxic, i d és una funció que varia segons l'ordre de la frase: start_i és la posició en què comença la frase en qüestió i end_{i-1} és la posició en què acaba la frase anterior en l'oració d'origen. A banda d'aquest model de reordenació basat en distància, hi ha altres models de reordenació més complexos que tenen en compte més paràmetres.

Combinació de models

Fins ara hem descrit un model estadístic que consiteix de tres sub-models: el de llenguatge, el de traducció i el de reordenament. Si anomenenm a les probabilitats p_l, p_t i p_r, el model resultant és el següent:

$$p(e|f) = p_l(e) \cdot p_t(f|e) \cdot p_r(f|e) \tag{4.8}$$

on:

$$p_l(e) = \prod_{t=1}^{|e|} p\left(e_t|e_0^{t-1}\right) \tag{4.9}$$

$$p_t(f|e) = \prod_{i=1}^{n} \phi(f_i|e_i) \tag{4.10}$$

$$p_r(f|e) = \prod_{i=1}^{n} d\left(\text{start}_i - \text{end}_{i-1} - 1\right) \tag{4.11}$$

Si oblidem per un moment que són probabilitats, perquè només ens interessa maximitzar l'expressió, podem introduir pesos als models, per a donar més importància a uns que a altres. Així, introduïm λ_l, λ_t i λ_r:

$$\hat{e} = \arg\max_e \left(p_l(e)^{\lambda_l} \cdot p_t(f|e)^{\lambda_t} \cdot p_r(f|e)^{\lambda_r}\right) \tag{4.12}$$

Aplicant logaritmes, podem reinterpretar-ho com un model log-lineal:

$$\log p(x) = \sum_{i=1}^{n} \lambda_i h_i(x) \tag{4.13}$$

On hi ha tres funcions característiques ($n = 3$), la variable x és la tupla $(e, f, \text{start}_i, \text{end}_{i-1})$, i les tres funcions característiques són:

$$h_1 = \log p_l(e) \tag{4.14}$$

$$h_2 = \log p_t(f|e) \tag{4.15}$$

$$h_3 = \log p_r(f|e) \tag{4.16}$$

Amb un model log-lineal, podem fins i tot introduïr més funcions característiques sense preocupar-nos de la descomposició del problema original $p(e|f)$! Una addició molt raonable és la d'un model de reordenament en sentit invers, per a tindre un reordenament més robust. També podem introduïr una altra característica per afavorir frases més llargues, ja que el model de llenguatge està esbiaixat a favor de les frases més curtes.

4.1.2 Resultats

S'han entrenat sistemes *phrase-based* per a les dues tasques descrites en el capítol 2: WMT Alemany→Anglés i PM Castellà→Anglés. Els resultats han estat mesurats amb el BLEU, que podeu consultar en la secció 1.6. La taula 4.1 presenta els resultats d'aquest experiment.

Taula 4.1: Mesures de BLEU de *phrase-based* per a la tasca WMT Alemany→Anglés

Dades	BLEU `dev`	BLEU `test`
bilingües	24.3	25.8
+ monolingües	27.0	30.1

La taula 4.1 té dues entrades: la primera correspon a un model entrenat amb les dades bilingües presentades en el capítol 2, i la segona són els resultats d'introduir grans quantitats de dades monolingües en el model de llenguatge. Podem comprovar que la millora és molt significativa, amb 2.7 punts de BLEU en `dev` i 6.3 punts en `test`.

Per a comparar els resultats dels sistemes *neural MT* amb aquests, la comparació "justa" seria amb els resultats obtinguts amb les mateixes dades, que corresponen a la primera entrada de la taula. No obstant això, també incloem els altres resultats perquè a la llarga és l'objectiu a superar si voelm obtindre els millors sistemes de traducció automàtica possibles amb *neural MT*.

[[resultats pm es->en]]

4.2 Sistemes *neural MT*

Igual que amb els sistemes *phrase-based*, les idees bàsiques de *neural MT* ja han estat introduïdes en la secció 1.4 de la introducció. Ara ho abordarem amb més detall, inclent una perspectiva de la seua implementació concreta i els sistemes entrenats. Primer descriurem la topologia exacta del model utilitzat, i revisarem els processos d'entrenament i de cerca amb aquests models. A continuació exposarem els experiments i resultats obtinguts, i els refinaments que s'han aplicat per a reduïr la talla del vocabulari.

4.2.1 Topologia dels models

Tal i com hem vist en la secció 1.4, tots els models *neural MT* contemplats en aquest TFM es basen en xarxes recurrents. Entre les possibles topologies recurrents, hem escollit utilitzar LSTMs (*Long Short-Term Memory*), que és una de les opcions que han demostrat ser capaces d'aprendre dependències de llarg termini un poc millor. A continuació expliquem les equacions de les LSTMs, comparades amb les xarxes recurrents més bàsiques per entendre per què és així.

La xarxa recurrent més bàsica possible està definida per la següent equació:

$$h_t = \sigma\left(Wx_t + Uh_{t-1} + b\right) \tag{4.17}$$

On σ és una funció que sol ser la tangent hiperbòlica; h_t és l'eixida en el moment t; i x_t és l'entrada en el moment t. W i U són les matrius de pesos, i b el *bias*. En teoria, fins i tot un model així de senzill (dos matrius i un vector) és capaç de representar dependències de llarg termini, per exemple de x_4 amb h_{12}, encara que en la pràctica no ho fan degut a que el gradient ha de passar moltes voltes per la funció σ durant l'entrenament, i cada vegada que hi passa s'esvaeix una mica. Aquest problema ha estat explorat en [11] i [7].

Les LSTMs estan pensades expressament per evitar eixe problema. Van ser introduïdes en [12], i en treballs posteriors se'n van fer modificacions i variants. Les equacions d'una LSTM bàsica són les següents:

$$f_t = \sigma_g\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{4.18}$$
$$i_t = \sigma_g\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{4.19}$$
$$o_t = \sigma_g\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{4.20}$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c\left(W_c x_t + U_c h_{t-1} + b_c\right) \tag{4.21}$$
$$h_t = o_t \circ \sigma_h\left(c_t\right) \tag{4.22}$$

On f_t, i_t i o_t, definits en les tres primeres equacions, són vectors que actuen com a "portes": estan pensats per a ser multiplicats valor a valor amb altres vectors (producte Hadamard, denotat per l'operador $\circ$), "deixant passar" el valor amb un 1 i "bloquejant el pas" amb un 0 en cada posició. En realitat els valors no són 0 ni 1, sinó valors intermedis, i per tant controlen en quina mesura "passa" cada valor. Les tres portes tenen distints paràmetres (W, U i b), però les mateixes entrades i la mateixa funció d'activació σ_g, que ha de ser una funció que situe els valors entre 0 i 1, per exemple la funció logística:

$$\sigma_g(x) = \frac{1}{1 + e^{-x}} \tag{4.23}$$

La idea bàsica d'aquesta disposició és tindre un estat (c_t) que es pot modificar segons l'entrada (x_t) i a partir del qual s'extreu l'eixida (h_t). Les portes f_t i i_t controlen l'estat segons l'equació 4.21, i la porta o_t controla com se n'extreu l'eixida (equació 4.22. f_t és la *forget gate*, i multiplica a c_{t-1} per a "oblidar"-ne parts (en la mesura en què els valors de f_t s'apropen a zero). i_t és la *input gate*, i multiplica a $\sigma_c\left(W_c x_t + U_c h_t + b_c\right)$, que es pot interpretar com a "*input*", i multiplicar per i_t és seleccionar quina part se n'escriu (se suma) en c_t. La funció d'activació d'aquest *input* σ_c no és necessàriament la mateixa que la de les portes σ_g. De fet, en la implementació estàndard, $\sigma_c = \tanh$, que és una funció sigmoide que situa els valors entre -1 i 1. Per acabar, o_t és la *output gate*, que selecciona quina part de l'estat va a l'eixida tal i com mostra l'equació 4.22, on també $\sigma_h = \tanh$ en la implementació estàndard.

Aquesta forma de treballar el que aconsegueix és separar l'estat intern (c_t) de l'eixida (h_t) en un instant donat: a l'estat només se li apliquen operacions lineals (multiplicacions i sumes), de forma controlada mitjançant les portes. En conseqüència, c_t no passa per cap funció d'activació cada instant, fet que li permet mantindre el gradient durant una quantitat en principi indeterminada d'instants, i en la pràctica molt més gran que les RNN bàsiques. Totes les xarxes recurrents d'aquest TFM són LSTMs.

Amb la topologia de les cel·les bàsiques definida, podem passar a l'estructura del model de traducció. Els sistemes *neural MT* entrenats en aquest TFM han estat basats en el mecanisme d'atenció de la secció 1.4.2. Per facilitar la lectura, a continuació tornem a representar aquest mecanisme, assumint que el lector ja ha entés la idea bàsica.

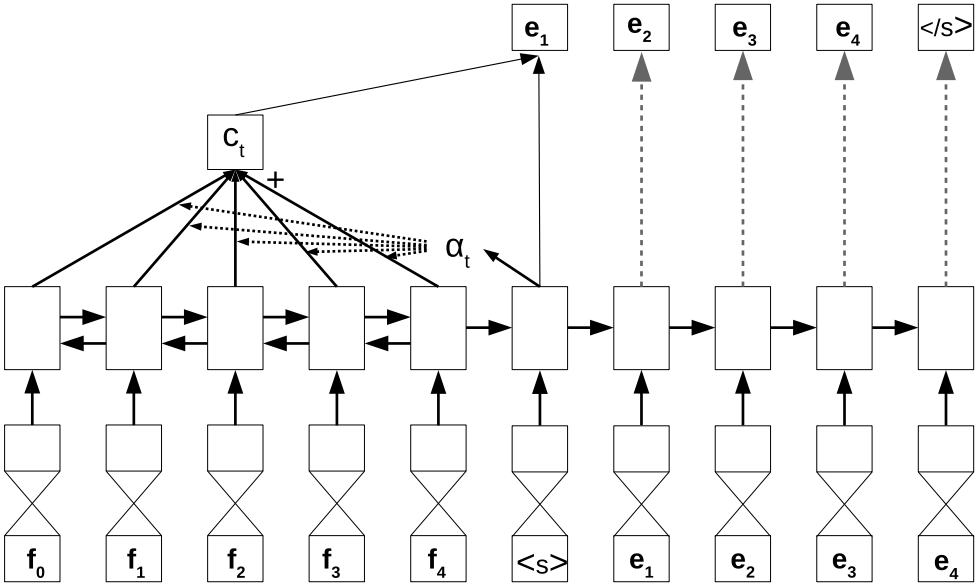

Figura 4.1: Mecanisme d'atenció per a *neural MT*

Per a l'*encoder*, hem utilitzat una xarxa recurrent bidireccional. Això és equivalent a tindre dues xarxes unidireccionals, una en cada direcció, i concatenar els dos vectors en cada instant. L'eixida de l'*encoder* és un vector $\overline{h}_s$ per a cada instant s, és a dir per a cada paraula d'entrada f_s.

El *decoder* és una xarxa recurrent unidireccional, que rep com a entrada en cada instant la paraula anterior i produeix un vector h_t. Aquest vector s'utilitza per puntuar cada $\overline{h}_s$, i generar en conseqüència el vector de context c_t, que no és més que una suma ponderada de les eixides de l'*encoder*:

$$c_t = \sum_s \alpha_t(s) \cdot \overline{h}_s \tag{4.24}$$

I d'on venen els $\alpha_t(s)$? hi ha distintes formes de computar-los a banda de l'original [5]. En aquest TFM hem utilitzat la proposta de [15], que introdueix tres variants que parteixen la idea que l'eixida del *decoder* h_t "puntua" les eixides de l'*encoder* $\overline{h}_s$:

$$\alpha_t(s) = f(h_t, \overline{h}_s) = \frac{\exp(\text{score}(h_t, \overline{h}_s))}{\sum_{s'} \exp(\text{score}(h_t, \overline{h}_{s'}))} \tag{4.25}$$

On $\text{score}(h_t, \overline{h}_s)$ es pot definir de distintes formes:

$$\text{score}(h_t, \overline{h}_s) = \begin{cases} h_t^T \overline{h}_s \\ h_t^T W_a \overline{h}_s \\ v_a^T (W_a h_t + U_a \overline{h}_s) \end{cases} \tag{4.26}$$

D'aquestes tres opcions, els nostres sistemes utilitzen la primera, que puntua els estats de l'*encoder* en funció del seu producte vectorial amb els estats del *decoder*. L'únic requisit per a que això puga funcionar és que els dos vectors siguen de la mateixa talla.

A partir de c_t i h_t calculem $\overline{h}_t$, que és un vector que li assigna una probabilitat a cada paraula e_t de ser la següent de la frase $p(e_t|e_0^{t-1})$. Aquest càlcul no és més que una capa de perceptró amb una activació softmax:

$$\overline{h}_t = \text{softmax}(W_h c_t + U_h h_t + b_h) \tag{4.27}$$

A partir d'aquest vector, s'escull la paraula següent e_t. La probabilitat de la frase serà la multiplicació de les probabilitats de cada paraula:

$$p_f(e_0^n) = \prod_{t=1}^{n} p_f(e_t|e_0^{t-1}) \tag{4.28}$$

On $p(e_0) = 1$, ja que e_0 és el *token* especial `<s>` que indica inici de frase. Si ens fixem podem veure que aquesta fórmula és molt semblant a la fórmula 4.3 que representava un model de llenguatge genèric. I en realitat, aquests models de traducció defineixen un model de llenguatge basat no en la llengua sinó en la frase d'entrada; això ho indiquem amb el subíndex f. En ser un productori, aquesta probabilitat es pot expressar també en forma logarítmica:

$$\log\left(p_f(e_0^n)\right) = \sum_{t=1}^{n} \log\left(p_f(e_t|e_0^{t-1})\right) \tag{4.29}$$

És important remarcar que, a l'hora de traduïr, no es pren sempre del *decoder* la paraula que tinga més probabilitat: potser hi ha frases amb més probabilitat que alguna de les seues paraules no era la més probable en eixe punt. Per a obtindre la frase d'eixida, s'ha d'aplicar una estratègia de cerca, explicada en la secció 4.2.3.

4.2.2 Entrenament

Amb la topologia definida, podem passar al problema de l'entrenament dels paràmetres. El primer que ens cal és un criteri d'entrenament, que en aquest cas es basa en maximitzar la probabilitat de la paraula correcta en cada instant, criteri que rep el nom de *maximum likelihood estimation*.

Les xarxes recurrents s'entrenen desenrotllant la xarxa al llarg del temps i creant un graf de computació no recurrent. Fet això, en cada instant t es calculen els errors i es retropropaguen els gradients cap a tots els paràmetres de la xarxa des de l'instant 0 fins al t.

L'entrenament es podria fer mostra a mostra, però és més pràctic fer-ho per grups anomenats *batches*, on cada *batch* conté una certa quantitat de mostres. Un *batch* situa els vectors en forma de matriu:

$$X = \begin{bmatrix} x_0 & x_1 & ... & x_n \end{bmatrix} \tag{4.30}$$

Aquesta idea fa ús de la propietat de la multiplicació de matrius segons la qual:

$$WX = \begin{bmatrix} Wx_0 & Wx_1 & ... & Wx_n \end{bmatrix} \tag{4.31}$$

Per tant podem tractar l'entrada de la mateixa manera si és un vector o si és un *batch*, sense canviar res. En compte de fer Wx (premultiplicar W a un vector), fem WX (premultiplicar W a una matriu). Treballar per *batches* té molts avantatges, el més notable del qual és que permet accelerar els càlculs amb GPUs, que són capaces de realitzar multiplicacions de matrius en coma flotant aprofitant al màxim el paral·lelisme de les operacions. A més, tindre en compte el gradient del *batch* i no de cada mostra redueix molt significativament el soroll en l'entrenament, movent els paràmetres cap a una direcció que redueix l'error global amb molta més precisió en cada pas.

Això causa un inconvenient menor, que ve del fet que les mostres d'entrenament són de longitud variable, i el temps de còmput per a un *batch* ve donat per la seua frase més llarga. Això pot fer que es perden molts recursos en frases curtes; la solució més efectiva consisteix a agrupar les frases per longitud: les més llargues entre elles i les més curtes també.

L'ordre de les mostres és important: s'ha demostrat que les últimes mostres tenen més impacte sobre el model resultant. Així, si les últimes mostres són tot frases curtes el model pot "oblidar" en part a traduïr frases llargues, per exemple. Alhora, per inicialitzar és més senzill fer-ho amb frases curtes. Una estratègia comuna és entrenar primer amb frases curtes i anar augmentant el tamany, encara que això pot acabar creant un model que tendeix a fer frases més llargues del que generalment caldria. En aquest TFM s'han ordrenat els *batches* aleatòriament, i comptem amb que hi ha una grandíssima quantitat de dades per a que encara que li coste un poc més inicialitzar el resultat final siga més equilibrat.

4.2.3 Cerca

En la cerca, el nostre objectiu és obtindre la frase d'eixida de màxima probabilitat donada la frase d'entrada. Com hem mencionat adés, no podem simplement prendre la paraula de màxima probabilitat en cada instant, o ens arrisquem a acabar amb una frase subòptima: potser la frase més probable té alguna paraula que puntualment no ho era. Per aconseguir-ne la millor cal seguir un procediment de cerca.

La traducció amb xarxes recurrents va paraula per paraula. En cada pas, predim la paraula següent computant una distribució de probabilitats entre totes les paraules del vocabulari. A partir d'ací, es poden definir estratègies de cerca en l'espai de les possibles frases en l'idioma objectiu. L'estratègia de prendre sempre la paraula la de màxima probailitat rep el nom de cerca voraç.

L'alternativa més usual és el *beam search*, o cerca en feix, que manté actives les n hipòtesis més probables i les va expandint, generant-ne de noves i eliminant les de menor probabilitat. La idea és simple: en comptes d'escollir una primera paraula de la frase, creem una llista de les n paraules més probables. A partir de cada paraula, es generen les paraules més probables en la segona posició, i ens quedem amb els n parells de major probabilitat. En cada pas, s'acumulen les probabilitats de cada hipòtesi sencera, no només la de l'última paraula. Quan alguna de les frases arriba al símbol `</s>` que indica final de frase, es guarda i s'elimina del *beam*, i es redueix en 1 el nombre de frases que se cerquen paral·lelament. La cerca s'acaba quan aquest nombre arriba a zero.

Al final, la hipòtesi més probable és l'escollida com a traducció. Cal comentar un detall, i és que tal com es calculen les probabilitats (equacions 4.28 i 4.29) les frases més llargues reben molta menys massa de probabilitat, així que en la pràctica se sol aplicar una normalització en funció de la longitud de frase per donar possibilitat a les frases més llargues de resultar escollides.

4.2.4 *Byte Pair Encoding*

En qualsevol llengua, la freqüència d'aparició de les paraules segueix una distribució desigual: unes poques paraules hi apareixen molt, i moltes paraules hi apareixen molt poc. La conseqüència és que per molt gran que siga la talla del vocabulari, sempre hi ha una quantitat significativa de paraules que se'n queden fora.

En una xarxa com les ací descrites, el temps de còmput i la memòria depenen linealment de $|V|$, la talla del vocabulari: la matriu que genera les probabilitats de cada paraula té $|V|$ files, i a més la funció $softmax$ involucra un sumatori sobre tot el vector, de talla $|V|$. En la pràctica, la majoria de models de traducció neurals restringeixen el vocabulari a entre 20 000 i 80 000 paraules aproximadament, marcant la resta com a `<unk>` o algun altre símbol especial que indica que estan fora del seu vocabulari. Això fa que els refinaments dels sistemes de NMT destinats a reduïr la talla del vocabulari tinguen un interés especial.

L'aproximació més comú és dividir les paraules en parts, en unitats més menudes que moltes paraules comparteixen entre si, per exemple paraules compostes, prefixos i sufixos, derivacions morfològiques o flexió verbal. A més, algunes llengües com l'alemany aglutinen paraules i poden beneficiar-se més d'aquest procediment.

Un mètode molt popular per a generar el conjunt de parts de paraula i paraules completes és el *Byte Pair Encoding* (BPE), que com millors resultats dóna és quan s'entrena sobre corpus paral·lels. El procés és simple: comencem amb un vocabulari de només les lletres. En cada pas, el parell més freqüent d'unitats –En el primer pas, el parell més freqüent de lletres– s'uneix. Es fan els passos que siguen necessaris fins obtindre un vocabulari de la talla desitjada, tenint en compte que a cada pas $|V|$ augmenta en 1 i que se li han d'agefir símbols especials com `<s>` i `</s>`.

4.2.5 Resultats

S'han entrenat sistemes *neural MT* per a les dues tasques descrites en el capítol 2, WMT Alemany→Anglés i PM Castellà→Anglés, igual que amb els *phrase-based*. Els resultats es presenten en dues taules.

[[Paràmetres de xarxa i configuració]]

La taula 4.2 presenta els resultats dels experiments per a la tasca WMT Alemany→Anglés.

Taula 4.2: BLEU de *neural MT* per a la tasca WMT Alemany→Anglés

Sistema	BLEU `dev`	BLEU `test`
base	18.3	19.1
+ BPE	20.9	25.3

La taula 4.2 té dues entrades. La primera correspon a un sistema base entrenat amb un vocabulari de 30 000 paraules; i la segona correspon a un model entrenat amb un vocabulari que utilitza BPE, de la mateixa talla. Podem comprovar que hi ha una millora significativa amb l'ús de BPE, de 2.6 punts de BLEU en `dev` i 6.2 punts en `test`.

Aquesta millora es pot explicar per dos motius: el primer i més rellevant és que amb la mateixa talla de vocabulari, un sistema que utilitza BPE pot representar un vocabulari molt més gran; i el segon és que el sistema pot detectar algunes regularitats en els patrons, per exemple una part de paraula que indica el plural, o el temps verbal.

La taula 4.3 presenta els resultats dels experiments per a la tasca PM Castellà→Anglés.

Taula 4.3: BLEU de *neural MT* per a la tasca PM Castellà→Anglés

Sistema	BLEU `dev`	BLEU `test`
base	22.0	25.7
+ BPE	24.0	26.6

La taula 4.3 té dues entrades, igual que la 4.2, i corresponen de la mateixa manera a un sistema base i a un sistema amb BPE per al vocabulari. En aquest cas també

hi ha una millora, de 2.0 punts de BLEU per al `dev` i de 0.9 punts per al `test`, que és menor que la que observem en Alemany→Anglés.

Una possible causa d'aquesta diferència és que l'alemany és una llengua aglutinant, i per això en dividir les paraules en trossos el sistema pot identificar més correlacions amb paraules en anglés, cosa que es nota menys en el cas Castellà→Anglés, on cap de les dues llengües és algutinant.

4.3 Comparació dels sistemes *phrase-based* i *neural MT*

Per tancar aquest capítol, en aquesta secció comparem els resultats empírics dels sistemes entrenats *phrase-based* i *neural MT*.

Els resultats exposats al llarg del capítol estan resumits en la taula 4.4.

Taula 4.4: Resum dels resultats de tots els sistemes per a les dues tasques

		phrase-based		***neural MT***	
Tasca	Sistema	`dev`	`test`	`dev`	`test`
Alemany→Anglés	base	24.3	25.8	18.3	19.1
	+ dades/BPE	27.0	30.1	20.9	25.3
Castellà→Anglés	base	—	—	22.0	25.7
	+ dades/BPE	—	—	24.0	26.6

La taula 4.4 mostra un resum dels resultats obtinguts, per facilitar la comparació. Agrupa els resultats per tasca, i dedica dues entrades a cada una. La primera és el sistema base, i la segona és el refinament aplicat, que en els sistemes *phrase-based* consisteix a alimentar el model de llenguatge amb grans quantitats de dades monolingües, i en els sistemes *neural MT* consisteix a substituir el vocabulari per un vocabulari BPE.

Els dos refinaments no són equiparables, ja que introduir grans quantitats de dades que l'altre sistema no ha tingut fa la comparació injusta, així i tot s'han introduït perquè representen la tecnologia dels millors sistemes de traducció de què disposem, i als que aspirem a curt o mitjan termini a superar mitjançant la nova tecnologia de *neural MT*. La comparació justa és la del sistema base de *phrase-based* amb el sistema BPE de *neural MT*, que representa el millor resultat que hem pogut obtindre gastant les mateixes dades.

En tots els casos estudiats en aquest TFM, els sistemes *phrase-based* han donat resultats més bons que els *neural MT*. En el cas d'Alemany→Anglés, els sistemes base han tret uns resultats pitjors en 6.0 punts en `dev` i 6.7 en `test`. No obstant això, si comparem el sistema base del *phrase-based* amb el sistema BPE de *neural MT*, obtenim uns resultats que estan a només 3.4 punts en `dev` i a 0.5 punts en `test`.

En conclusió, a igualtat de condicions, els millors sistemes *neural MT* entrenats al llarg d'aquest TFM s'han aproximat als resultats dels millors sistemes *phrase-based*

que podem entrenar, amb distàncies de 0.5 punts en el millor cas. Això no representa els millors sistemes que podem entrenar amb *neural MT*, ja que hi ha tècniques que no hem utilitzat en aquest TFM, com el *ensemble* o l'entrenament utilitzant dades monolingües, amb les que ens podríem posar al nivell i probablement superar els resultats obtinguts pels nostres millors sistemes *phrase-based*, tenint en compte la poca diferència en els resultats obtinguts.

Capítol 5

Conclusions

Aquest capítol tanca la memòria del TFM amb un resum de tot el que s'ha vist al llarg de tota la memòria i les conclusions més rellevants de cada part, per donar una visió general del treball. A més, s'hi presenten les vies de treball futur per aconseguir superar els resultats de *phrase-based* amb sistemes *neural MT*.

En aquest treball, hem abordat la traducció automàtica des d'un enfocament recent que trenca amb la tecnologia que s'ha estat utilitzant fins ara. Hem motivat aquesta elecció amb la formació científica i amb l'aplicació pràctica a plataformes de vídeos educatius, en la generació ràpida i barata de subtítols multilingües.

En la introducció, també hem lligat el TFM amb el TFG que el precedeix, que tractava sobre transcripció automàtica. Aquest lligam uneix els dos treballs que es circumscriuen a l'àmbit de les tecnologies del llenguatge i les seues aplicacions. També hem vinculat el TFM a una publicació científica relacionada, en l'avaluació de les seues aplicacions en casos d'ús reals en plataformes educatives. Hem introduït el pas de *phrase-based* a *neural MT* que està tenint lloc en els darrers anys, presentant les bases de les dues tecnologies i les seues diferències. També hem explicat les eines utilitzades per realitzar aquest TFM i dues mètriques per avaluar els resultats obtinguts.

En el capítol 2, hem donat detalls dels conjunts de dades que s'han utilitzat per entrenar i avaluar els sistemes de traducció automàtica. Els corpus han sigut els del WMT per a Alemany→Anglés, i el corpus poliMedia per a Castellà→Anglés. Hem mostrat estadístiques de quantitat de frases i de mitjana de paraules per frase per cada llengua i en cada conjunt de dades de les dues tasques.

Hem dedicat el capítol 3 a l'avaluació dels sistemes de transcripció i traducció automàtiques sobre casos d'ús reals, treball que va conduir a la publicació d'un article científic en una revista internacional. Hem proporcionat una avaluació dels resultats d'aquesta aplicació des de tres punts de vista: la qualitat dels resultats, el temps humà requerit per revisar-los, i l'impacte que han tingut sobre les plataformes MOOC.

La qualitat dels resultats ha estat en nivell útils per a la seua aplicació, i de fet han estat millors que proveïdors comercials de multinacionals com YouTube i Google Translate, per l'adaptació dels sistemes a la tasca. Pel que fa a temps humà de revisió, hem vist que un professor pot estalviar-se al voltant del 50% del temps, en molts casos

més, i hem proposat un sistema de regressió per inferir el temps que cal dedicar a revisar a partir de l'error dels sistemes. En l'avaluació de l'impacte dels sistemes, hem vist que la disponibilitat de subtítols multilingües va tindre un gran ressò sobre els casos d'estudi, amb un augment de les inscripcions d'un 70% en EMMA i millores en l'accessibilitat i funcionalitat afegida a partir dels subtítols en poliMedia.

El capítol 4 ha explicat els sistemes entrenats, la seua tecnologia i els resultats obtinguts en les dues tasques abordades. Pel que fa als sistemes *phrase-based*, hem vist que estan formats de tres models: el model d llenguatge, el model de traducció o model lèxic, i el model de reordenament. Hem vist que un punt fort de la seua tecnologia és la combinació de models, que alguns es poden entrenar amb més dades i que, vistos com a model log-lineal, se li poden afegir més funcions característiques, enfortint el sistema resultant.

Quant als sistemes *neural MT*, hem detallat a fons la topologia de les xarxes i els motius pels que és així, des de l'organització d'una cel·la recurrent al funcionament en concret del mecanisme d'atenció, donant una visió rigorosa de cada part i una visió general del conjunt. També hem exposat l'ús d'aquests models durant l'entrenament i durant la cerca en l'espai de les possibles traduccions. Hem completat la presentació amb un dels refinaments aplicables a aquest tipus de sistema, el *Byte Pair Encoding*, i hem comprovat la seua eficàcia de forma experimental.

Els resultats obtinguts pels sistemes *neural MT* han estat relativament pitjors que els dels sistemes *phrase-based*, però aquesta diferència es reduïa a distàncies gens insalvables, que han arribat als 0.5 punts de BLEU en algun cas si féiem la comparació justa. En conclusió, a igualtat de condicions els sistemes *neural MT* s'aproximen als resultats dels millors sistemes *phrase based*. Com que podem entrenar sistemes *neural MT* amb tècniques que no s'han aplicat en aquest TFM, una conclusió directa del treball és que estem prop de superar els resultats dels sistemes *prase-based*, tenint en compte la poca distància que en separa els resultats.

Bibliografia

[1] Emma project. http://platform.europeanmoocs.eu.

[2] Transcription and translation of videolectures project. http://www.translectures.eu.

[3] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[4] Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[6] Pau Baquero-Arnal. Comparació de les eines informàtiques TLK i Kaldi per al desenvolupament de sistemes de reconeixement de la parla en català. Treball Fi de Grau, Universitat Politècnica de València, 2016.

[7] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, Mar 1994.

[8] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359 – 394, 1999.

[9] M. A. del Agua, A. Giménez, N. Serrano, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan. *The Translectures-UPV Toolkit*, pages 269–278. Springer International Publishing, Cham, 2014.

[10] MA del Agua, A Giménez, N Serrano, Jesús Andrés-Ferrer, J Civera, A Sanchis, and A Juan. The translectures-upv toolkit. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 269–278. Springer, 2014.

[11] Sepp Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. 04 1991.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[13] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[14] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[15] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025, 2015.

[16] Adrià Martínez-Villaronga, Miguel del Agua, Jesús Andrés-Ferrer, and Alfons Juan. Language model adaptation for video lectures transcription. In *Proceedings of ICASSP*, pages 8450–8454, 2013.

[17] Juan Daniel Valor Miró, Pau Baquero-Arnal, Jorge Civera, Carlos Turró, and Alfons Juan. Multilingual videos for moocs and oer. *Journal of Educational Technology & Society*, 2017. in print.

[18] Juan Daniel Valor Miró, Joan Albert Silvestre-Cerdà, Jorge Civera, Carlos Turró, and Alfons Juan. Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories. *Speech Communication*, 74(Supplement C):65 – 75, 2015.

[19] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[21] Mirko Plitt and François Masselot. A productivity test of statistical machine translation post-editing in a typical localisation context. 93:7–16, 01 2010.

[22] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[23] Alejandro Manuel Pérez González de Martos, Joan Albert Silvestre Cerdà, Juan Daniel Valor Miró, Jorge Civera Saiz, and Alfons Juan Císcar. MLLP Transcription and Translation Platform. In *10th European Conference on Technology Enhanced Learning (EC-TEL 2015)*, Toledo, Spain, 2015. Springer International Publishing.

[24] Dennis Reidsma, D.H.W. Hofs, and N. Jovanovic. *Designing Focused and Efficient Annotation Tools*, pages 149–152. Noldus Information Technology, 2005. Imported from HMI.

[25] Dhawal Shah. By the numbers: Moocs in 2015. https://www.class-central.com/report/moocs-2015-stats/, 2015.

[26] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 2006.

[27] Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904, 2002.

[28] Carlos Turro, Miguel Ferrando-Bataller, Jaime Busquets, and Aristóteles Cañero. Polimedia: a system for successful video e-learning. In *International Conference EUNIS 2009*, 2009.

Índex de figures

Índex de taules