UNIVERSITAT POLITÈCNICA DE VALÈNCIA
ESCOLA TÈCNICA SUPERIOR D'ENGINYERIA INFORMÀTICA
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Adaptació massiva de models de llenguatge en transcripció de video-xarrades

Projecte final de carrera - Enginyeria Informàtica

Santiago R. Piqueras Gozalbes

Supervisat per:
Dr. Alfons Juan Císcar
Dr. Jesús Andrés Ferrer

29 de juliol de 2013

A Leo i Lines

Índex

Capítol 1

Introducció

1.1 Motivació

És difícil negar que visquem en un món de informació. Els avançaments científics produïts en les últimes dècades han facilitat el accés a la mateixa a gran part de la població.

Dins de l'espiral de informació, i especialment en l'àmbit educatiu, les videoxarrades estan emergent com un poderós mitjà de difusió. Es poden comptar ja per centenars els portals web que oferixen gratuïtament continguts acadèmics de moltes i diverses temàtics en aquest format.

És en aquest context on un treball en transcripció automàtica de la parla pren sentit. La transcripció automàtica no només elimina les barreres existents pel fet de que la informació sigui audible, sinó que pot facilitar, per exemple, la extracció automàtica de dades de les xarrades o la traducció dels continguts. Es considera doncs de especial interès que aquesta transcripció sigui de la màxima qualitat possible.

Com és d'esperar, la tasca de aconseguir una transcripció automàtica propera o fins i tot comparable a la que puga obtindre una persona no és senzilla. Per això sembla raonable restringir aquesta tasca, per exemple, a la transcripció de vídeos de una temàtica concreta, amb la intenció de focalitzar els esforços i obtindre millors resultats.

La adaptació dels models emprats en el reconeixement de la parla és part d'aquesta focalització. En aquest treball centrarem els esforços en el model de llenguatge.

1.2 Reconeixement de formes

El reconeixement de formes és la branca de la informàtica, i més en concret de la intel·ligència artificial, que s'encarrega del desenvolupament de sistemes complexos perceptors d'objectes. En el context del reconeixement automàtic, s'entén com a percepció l'assignació d'una etiqueta, entre un conjunt (finit o infinit) d'etiquetes permeses, al objecte. Com a exemple, podríem tindre com a objecte el àudio d'una frase en concret, i com a etiqueta la transcripció que torne el nostre sistema.

A l'hora de dissenyar un classificador, hem de tindre en compte que les mostres deuran passar tres fases ben diferenciades:

1. **Preprocès** A partir del objecte, s'adquirix el senyal, es filtra per eliminar (o reduir) soroll i es prepara per a la següent fase.
2. **Extracció de característiques** Del senyal processat, s'adquirix la informació rellevant per al sistema i es computa un vector de característiques, tractant d'extreure tanta informació discriminant com ens sigui possible.
3. **Classificació** Les mostres ja estan preparades per la classificació. Amb l'ajuda dels models prèviament entrenats, s'assigna una etiqueta al objecte d'entrada. Aquesta etiqueta serà l'eixida de la classificació.

A la Figura 1.1 podem fer-se una idea de com funciona un sistema reconeixedor.

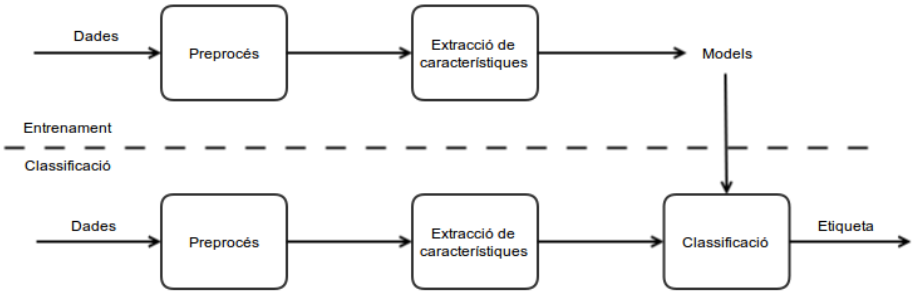

Figura 1.1: Sistema reconeixedor

1.3 Reconeixement automàtic de la parla

Dins del reconeixement de formes, el camp que es tracta en aquest projecte és el del reconeixement automàtic de la parla o ASR (del anglès, *Automatic Speech Recognition*). En aquest capítol es vol donar una visió global del problema que originà aquesta branca.

L'objectiu del ASR és processar un senyal acústic (entrada del sistema) que conting a veu i obtindre la transcripció en text (eixida) del que s'està dient. Formalitzarem

aquest problema com l'obtenció d'una seqüència de paraules $w = w_1 w_2 ... w_N$ a partir de un seguit d'observacions acústiques $x = x_1 x_2 ... x_T$ que maximitze la probabilitat a posteriori $p(w|x)$ [Jel97], com veiem a la fórmula 1.1.

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

On $\hat{w}$ serà la eixida desitjada. Pel teorema de Bayes, podem descompondre aquesta probabilitat com:

$$\hat{w} = arg \max_{w} \frac{p(x|w)p(w)}{p(x)} \tag{1.2}$$

Com, donada una mostra x, $p(x)$ és constant per a tot w, podem eliminar-ho del denominador, quedant-nos la següent fórmula.

$$\hat{w} = arg \max_{w} p(x|w)p(w) \tag{1.3}$$

Com és evident, si per a un llenguatge donat coneguérem aquestes probabilitats, el problema estaria resolt. Com podem imaginar, aquest cas ideal no es dona en la realitat. El nostre objectiu a l'ASR, i en general a tot el reconeixement de formes, serà aproximar aquestes probabilitats mitjançant l'entrenament de models amb dades prèviament etiquetades. Per a aquesta tasca, en l'ASR s'empren dos models:

Model acústic S'encarrega d'estimar $p(x|w)$. El models més estesos amb aquesta finalitat són els models de markov de capa oculta (*Hidden Markov Models* o HMM). L'entrenament d'aquestos models requerix les mostres de àudio processades (és a dir, que han passat ja per les fases de preprocés i extracció de característiques) i el text corresponent, aixina com una transcripció fonètica del mateix.

Model de llenguatge S'encarrega de l'estimació de $p(x)$. El model tradicional per al modelat de llenguatge són els n-grames, i és el que gastarem en aquest projecte, encara que treballs recents estan explorant l'ús de xarxes neuronals. En aquest cas, no és necessari res més que textos en el idioma a transcriure, ja que aquesta probabilitat no és depenent de la seqüència de vectors acústics. Com és objectiu d'aquest projecte l'adaptació d'aquests models, a la secció 1.4 els veurem en profunditat.

Com ja hem vist a la figura 1.1, l'obtenció d'aquests models a partir de mostres requerix una fase d'entrenament. Que aquest entrenament es faça amb les tècniques adequades és tan o més important que les dades emprades en el mateix.

Per a l'entrenament de models acústics, es farà servir el algorisme de Viterbi, un algorisme recursiu per trobar la seqüència més probable d'estats (camí) en un graf a partir d'una seqüència d'observacions [YEG+02]. Pel que respecta a models de llenguatge basats en n-grames, l'entrenament es fa a partir de comptes d'n-grames 1.4.1 i suavitzat 1.4.2.

En ambdós casos, si es disposa de dades de distintes fonts i/o de distint domini, es sol entrenar models individuals per a cada corpus i després gastar tècniques d'interpolació per combinar les eixides . Freqüentment, la interpolació es gasta per fer adaptació de models, mitjançant l'ajustament dels pesos.

A la figura 1.2 podem veure un diagrama que recull el procés que seguix l'entrenament i la classificació en un sistema de ASR. Aquesta figura no recull una possible interpolació de models acústics i/o de llenguatge.

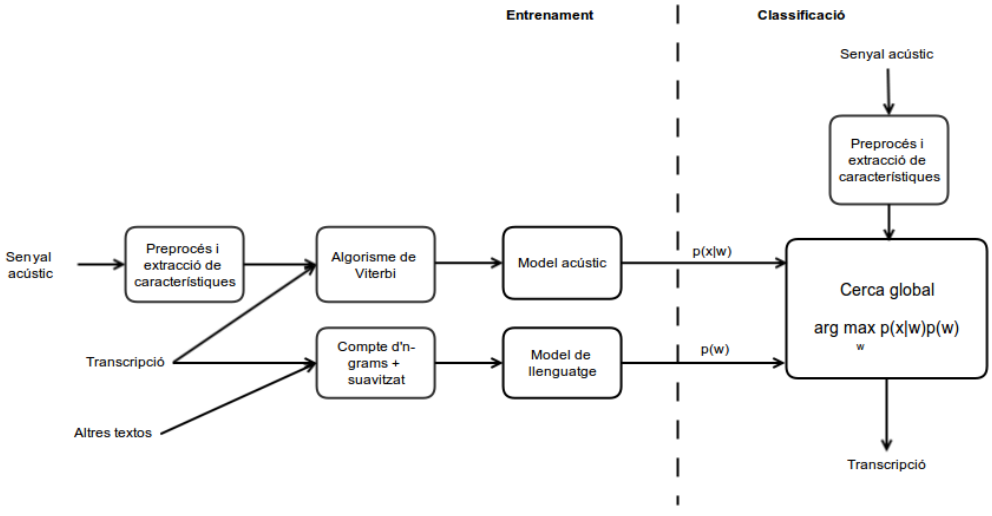

Figura 1.2: Entrenament i classificació d'un sistema d'ASR

1.4 Model de llenguatge

Com hem dit en la secció anterior, el objectiu d'un model de llenguatge en el ASR és el de puntuar les possibles transcripcions d'una frase concreta. Donada una frase, la puntuació assignada serà la probabilitat de que el nostre model genere eixa frase.

Suposant que la frase s de longitud N està composta de les paraules $w_1 w_2 \dots w_N$, podem computar la probabilitat $p(s)$ com

$$p(s) = \prod_{i=1}^{N} p(w_i | w_1^{i-1}) \tag{1.4}$$

Intentar aplicar aquesta fórmula directament implicaria que suposem que la probabilitat de aparició de la última paraula d'una frase es depenent de tota la resta de paraules de dita frase. Depenent del llenguatge, aquesta suposició pot no semblar massa raonable. Una millor aproximació semblaria considerar que la aparició d'una

paraula concreta depèn d'un context limitat, és a dir, de les paraules que apareixen properes d'ella.

Els models d'n-grames intenten captar aquesta idea de context limitat. Així, a l'hora de fer el còmput, la probabilitat de que una paraula concreta aparega en la posició i només dependrà de les paraules que hagen aparegut en les n posicions anteriors, on n es un paràmetre fixe del model. Per tant, la fórmula per computar la probabilitat d'una frase ens queda així

$$p(s) \approx \prod_{i=1}^{n} p(w_i | w_{max(1,i-n+1)}^{i-1}) \tag{1.5}$$

1.4.1 Estimació

Abordem ara la qüestió de com obtindre un model de n-grames d'un llenguatge concret. Per solucionar aquest problema, necessitem adquirir un corpus d'entrenament, és a dir, frases que (suposem) ben construïdes del llenguatge objectiu. Obtingut aquest corpus, podem aproximar la probabilitat que busquem com

$$p(w_i | w_{i-n+1}^{i-1}) \approx \frac{c(w_{i-n+1}^{i})}{c(w_{i-n+1}^{i-1})} \tag{1.6}$$

On c és la funció que ens diu el nombre de voltes que ha aparegut un n-grama al corpus. Aquesta aproximació es coneix com **estimació per màxima versemblança**.

1.4.2 Suavitzat

El principal problema que sorgix de l'aproximació per màxima entropia és que es produïx una distribució de probabilitats que assigna zeros als esdeveniments que no s'han donat en el corpus d'entrenament. Com la probabilitat d'una frase és el producte de la probabilitat dels n-grames, un sol terme igual a zero provocarà que la probabilitat de la frase sencera sigui zero. Com podem imaginar, és molt poc realista imaginar que en les dades que es posseïxen es donen tots els possibles n-grames d'un llenguatge.

Les tècniques de suavitzat sorgixen per contrarestar aquest problema. Prenen el seu nom del fet de que pretenen, literalment, suavitzar la distribució de probabilitats per tal d'eliminar les probabilitats nul·les. Un estudi sobre les distintes tècniques disponibles per a fer el suavitzat d'n-grames es pot veure en [CG99].

Quan volem calcular la probabilitat de una paraula després d'un n-grama, la majoria de les tècniques de suavitzat actuals empren n-grames de ordres menors per dur a terme la tasca. Segons com s'empren aquestos n-grames, distingirem dos tipus de funcions de suavitzat:

Backoff La probabilitat d'un n-grama que es troba al corpus d'entrenament només depèn dels comptes calculats per aquest n-grama. És a dir, si $c(w_{i-n+1}^{i}) \neq 0$, no gastarem els comptes de n-grames de menor ordre.

Interpolació Els comptes de n-grames de ordres inferiors s'empren en el còmput de la probabilitat, tant si el compte del n-grama és zero o superior.

Les dos tècniques que s'empraran en aquest projecte són *Kneser-Ney Smoothing* (o KN Smoothing) i *Modified KN Smoothing*, les dos descrites al article [CG99].

1.4.3 Interpolació

La interpolació de models de llenguatge (no confondre amb la interpolació que es gasta a les tècniques de suavitzat) busca la combinació de distints MLs en uno nou. Aquesta tècnica ens permet crear models independents amb dades que proveïsquen de distintes fonts i/o de distinta natura i posteriorment fusionar-los per obtindre un model apte per al reconeixement.

La interpolació lineal és una de les tècniques més esteses per a interpolar models de llenguatge. Donat un conjunt de models, li assignarem a cadascun d'ells un paràmetre λ_i al que anomenarem pes en la interpolació. Per tal de computar la probabilitat que donats $w_{i-n+1} \dots w_{i-1}$ aparega la paraula w_i, emprarem la equació 1.7.

$$p^{INT}(w_i|w_{i-n+1}^{i-1}) = \sum_i \lambda_i p_i(w_i|w_{i-n+1}^{i-1}) \tag{1.7}$$

Per a que p^{INT} sigui una distribució de probabilitats, els paràmetres λ han de ser zero o superiors i estan restringits per la següent equació.

$$\sum_i \lambda_i = 1 \tag{1.8}$$

El valor dels paràmetres λ_i de la interpolació deu ser ajustat segons la tasca a solucionar. Una de les solucions més comuns es la optimitzar els paràmetres per tal de minimitzar la perplexitat del model front a un corpus de dades del domini del model de llenguatge final, mitjançant tècniques de optimització com el algorisme EM. A aquest corpus se li sol donar el nom de *development* (desenvolupament). A la secció 1.6 veurem com fer el còmput amb SRILM.

1.5 Avaluació dels resultats

Per a avaluar els resultats obtinguts, anem a emprar dos mesures clàssiques del reconeixement de la parla, la perplexitat (**PPL**) i el Word Error Rate (**WER**). Com veurem, aquestes dos mesures tendixen a estar relacionades entre si.

Tal i com es descriu a [YEG+02], podem entendre la perplexitat com el nombre mitjà de paraules que poden aparèixer després d'un prefix donat. Donats dos models

de llenguatge amb el mateix vocabulari, el model que tinga una perplexitat menor "dubtarà menys"a l'hora de avaluar un text concret. S'ha comprovat experimentalment que, en general, els models amb menor perplexitat oferixen millors resultats, encara que alguns articles mostrem com aquesta correlació no sempre es dona [CBR98].

Per a computar la perplexitat de una frase necessitem la probabilitat que té el model de produir eixa frase. Obtindrem la perplexitat mitjançant la següent fórmula:

$$PPL = 2^{-\frac{1}{m} log_2 p(w_1, w_2, \ldots, w_m)} \tag{1.9}$$

On $\{w_1, w_2, \ldots, w_m\}$ són les paraules de la frase i m és la llargària o nombre de paraules. El exponent d'aquesta fórmula també es coneix com entropia per paraula.

La perplexitat d'un model front a un text concret T ens la donarà la fórmula 1.10.

$$PPL = 2^{-\frac{1}{N} \sum_{s \in T} log_2 p(s)} \tag{1.10}$$

L'altra mesura que emprarem per a l'avaluació serà el WER, una mesura clàssica en l'avaluació de sistemes reconeixedors de la parla. És una mesura derivada de la distància mínima d'edició o distància de Levenshtein i pretén quantificar la diferència entre la transcripció tornada del sistema i una referència supervisada que suposem correcta. La fórmula 1.11 ens indica com calcular-ho.

$$WER = \frac{I + D + S}{N} \tag{1.11}$$

On I, D, S són el nombre mínim d'operacions de insercions, esborraments o substitucions que s'han de fer per a transformar la referència en l'eixida del nostre sistema, respectivament; i N és el nombre de paraules total a la referència.

Com ja s'ha dit, les dos mesures emprades solen estar relacionades, però són fonamentalment distintes. El còmput del WER es molt més costós i requerix que es dispose de dades en forma de àudio i d'un sistema reconeixedor complet. El sistema reconeixedor emprat en este projecte es descriurà a la secció 4.3.4.

Per altra banda, la perplexitat és un comput que només afecta al model de llenguatge desenvolupat i que pot no ser indicadora d'un resultat millor en la transcripció. El seu extens ús ve motivat de que no requerix de sistemes externs i posseïx un cost computacional molt més reduït.

En aquest treball gastarem la perplexitat al llarg de tota la experimentació, però sense perdre de vista que l'objectiu final és que la millora obtinguda per les tècniques d'adaptació es trasllade a una millora de WER.

1.6 Eines informàtiques

S'han emprant al llarg d'aquest projecte dos toolkits informàtics per al reconeixement de la parla: *SRI Language Modeling Toolkit* (SRILM) i *transLectures-UPV toolkit* (TLK).

1.6.1 SRILM

SRILM [Sto02] és un conjunt de ferramentes desenvolupat pel *Speech Technology and Research (STAR) Laboratory* del institut de recerca *SRI International*, a California, EEUU. Permet la construcció, modificació i aplicació de models de llenguatge. Consistix en una sèrie de programes i utilitats implementades sobre una base de llibreries en el llenguatge C++ i disponibles per a sistemes UNIX i Windows.

Els programes i scripts de SRILM dels que s'han fet us són:

ngram Principal ferramenta per a la utilització de LM. Ens servirà per calcular les entropies necessàries per a la selecció (ampliat a la secció 4.2), les perplexitats per a la avaluació de resultats (secció 1.5) i la interpolació de models (secció 1.4.3 per a la descripció, 4.3.3 per a l'aplicació).

ngram-count Principal ferramenta per a la construcció de LM. Ens permet estimar models de llenguatge a partir del text i ajustar distints paràmetres d'aquesta construcció, com la grandària dels n-grames, restriccions en el vocabulari o la tècnica de suavitzat a emprar (secció 1.4.2)

compute-best-mix Script que permet computar els pesos òptims de interpolació lineal de un seguit de models de llenguatge. Per pesos òptims s'entén aquells que minimitzen la perplexitat total del model de llenguatge interpolat front a un text donat. El gastarem a la secció 4.3.3 per calcular estos pesos.

1.6.2 TLK

TLK [tUT] és un toolkit desenvolupat per la Universitat Politècnica de València, en el marc del projecte transLectures, per a la construcció i ús de sistemes de reconeixement automàtic de la parla. Consistix de una serie de utilitats i programes sobre una llibreria base en C. S'oferix per a sistemes UNIX, en codi obert i llicència Apache 2.0. La primera release oberta va veure la llum al Maig de 2013.

Gràcies al sistema de ASR desenvolupat per la UPV per al projecte transLectures, que funciona sobre TLK, es podrà avaluar el funcionament de les tècniques desenvolupades sobre un sistema transcriptor competitiu. El interès de això és el de atorgar robustesa als resultats, pel que sabrem que les millores obtingudes poden aplicar-se a sistemes reals, una part que a voltes és ignorada per altres articles i treballs d'investigació.

Com el sistemes de ASR són complexos i no és objectiu de aquest projecte descriure'ls en detall, no es va a fer una descripció en profunditat de les ferramentes de TLK emprades. Sí que es vol destacar, per la seua relació en els models de llenguatge, que el format que gasten SRILM i TLK per a l'emmagatzemament de LM no és el mateix. SRILM gasta el estàndard ARPA, mentre que TLK gasta un format propi. El programa **tLlmformat** ens facilitarà la conversió.

Capítol 2

Descripció dels corpus

2.1 Introducció

En aquest capítol es descriuen els corpus (les dades) que s'han emprat en el projecte. Podem distingir dos possibles classificacions: corpus per a entrenament, ajustament de paràmetres i experimentació; o bé corpus del domini i fora del domini. Anem a fer servir esta segona classificació per a descriure els corpus, ja que, com veurem en la capítol 3, és una distinció necessària a l'hora de aplicar tècniques de selecció.

L'última secció d'aquest capítol la dedicarem a descriure les operacions de preprocés que s'han dut a terme per al entrenament de models.

2.2 Corpus del domini

El corpus de frases del domini està format per transcripcions disponibles de vídeos de la plataforma poli[Media] [Uni12]. poli[Media] és una plataforma web, creada i suportada per la UPV, per a la producció i distribució de continguts multimèdia educatius, i especialment vídeos. Va ser llençada al 2007 i posteriorment ha sigut implementada en altres universitats, com la Universitat Autònoma de Barcelona o la Universitat de Sao Paulo.

El total de dades disponible prové de la transcripció de 739 vídeos, que cobrixen prop de 114 hores de àudio. El corpus poli[Media] es gastarà tant per a entrenament com per a les proves de testing. Amb aquest objectiu, s'han dividit les dades en tres subcorpus, als que anomenarem *train*, *dev* i *test*:

train S'emprarà com a corpus del domini per a la selecció, i s'inclourà en la interpolació final.

dev Ens permetrà mesurar la eficàcia de la selecció i ajustar els pesos λ de la interpolació.

test Ens permetrà fer una valoració final dels resultats del mètode.

En la taula 2.1 podem veure un resum de la divisió realitzada. La columna "Grandària vocabulari"ens indica el nombre de paraules úniques que hi ha al corpus.

Taula 2.1: Dades del domini

Corpus	Nº de vídeos	Nº de frases	Nº de paraules	Grandària vocabulari
train	690	39.3K	0.9M	26.9K
dev	26	1401	35.4K	4.7K
test	23	1139	30.6K	4.2K

Del corpus poli[Media] es disposa a més de les dades acústiques que donaren lloc a les transcripcions. Açò ens permetrà, una volta acabada l'experimentació amb els models de llenguatge, comprovar si les millores obtingudes es traslladen a un sistema de transcripció de la parla real.

2.3 Corpus externs

Les dades disponibles fora del domini provenen de 9 fonts distintes de variada natura. Cadascun d'aquestos corpus es tractarà per separat per tal de discriminar aquells que siguen més rellevants per a la tasca que ens ocupa i donar-los un paper més important al model de llenguatge.

Les fonts que composen els distints corpus de dades són:

elperiodico Diari de temàtica general espanyol El Periódico. Aquest diari és bilingüe (espanyol i català), nosaltres només emprarem els textos que estiguen en espanyol. Cal dir que s'ha extret el text a partir de documents en pdf amb la ferramenta pdf2tex, pel que aquest corpus pot contindre frases que no són útils per a l'entrenament de models, com puga ser la cartellera.

EPPS Extractes del *European Parliament Plenary Sessions*, compost de centenars de intervencions als parlament europeu i espanyol. El total de hores transcrites és de 100, 38 de les quals provenen del parlament espanyol i les 62 restants del europeu. En aquest últim cas, algunes de les transcripcions s'han generat a partir d'intervencions fetes directament en espanyol, mentre que altres són transcripcions de gravacions d'intèrprets.

europarl Extractes del *European Parliament Proceedings*, pel que és similar a EPPS.

news Articles periodístics de fonts diverses obtinguts de Internet. Aquest corpus ho distribuïx el *Workshop on Statistical Machine Translation* [WMT13].

news-commentary Idèntic al corpus anterior, també està compost de notícies recopilades per Internet i distribuït per el WMT [WMT13].

TED Transcripcions de video-xarrades de la web TED. Aquesta web oferix xarrades de temàtiques diverses, fent èmfasi en la difusió d'idees. Com les xarrades originals són en anglès, farem servir seran les traduccions supervisades de les mateixes. Tant les traduccions com les transcripcions han sigut aportades per usuaris no experts.

tt Documents de distint tipus (jurídics, legals, acordes institucionals) produïts al Parlament Europeu.

UnitedNations Documents de distints tipus produïts a les Nacions Unides, extrets de la seua Web. Proveït per el WMT.

Els corpus TED, EPPS i europarl provenen de intervencions orals, que podria ser rellevant per la tasca que ens ocupa. La grandària dels corpus s'especifica a la taula 2.3.

Taula 2.2: Dades fora del domini

Corpus	Nº de frases	Nº de paraules	Grandària vocabulari
elperiodico	2.6M	44.5M	197.9K
EPPS	130K	828K	26.9K
europarl	2.1M	53.8M	133.1K
news	8.6M	213.3M	401.5K
news-commentary	183K	4.4M	77.5K
TED	315K	2.3M	69.4K
tt2	448K	10.1M	81.3K
UnitedNations	9.9M	312.6M	288.1K

Depenent de la font, és probable que les dades sense processar no siguen òptimes per a l'entrenament de models de llenguatge per al reconeixement. Als fitxers de dades podrem trobar signes de puntuació, nombres, capitalitzacions diverses i fins i tot paraules i frases en altres llenguatges. A la secció 2.4 s'explica amb més detall el procés de neteja que s'ha fet dels corpus per tal de preparar-los per a l'entrenament.

2.3.1 Google n-grams

El corpus *Google n-grams*[MSA+11] és una col·lecció de dades de comptes de n-grames provinents de llibres digitalitzats. Es disposen de dades de unigrames a 5-grames de 7 idiomes distints. Per fer-se una idea de com massiu és aquest corpus, els autors afirmen que 'comprèn prop del 4% dels llibres mai publicats'. La versió que s'ha fet servir en aquest projecte és la 1, publicada en Juliol de 2009. El corpus també conté més informació dels n-grames, com la data en la que aparegueren o el nombre de volums distints que el contenen. Aquesta informació serà filtrada, ja que no ens interessa per a construir els models.

S'ha volgut dedicar una secció a banda a aquest corpus per dos motius:

- No es disposen dels textos originals que han donat lloc als comptes. A la secció 3.2 veurem que això ens limitarà les tècniques a aplicar.
- La grandària del corpus és, comparat amb la resta de corpus disponibles, enorme. La informació sintàctica que obtinga el nostre model de llenguatge del corpus Google serà, per tant, molt més fiable que la que donen la resta de corpus. En el capítol de selecció per diferència d'entropia creuada, secció 4.3.3, aprofitarem aquest fet per a l'adjust de paràmetres a la selecció.

Taula 2.3: Dades de Google n-grams

Corpus	Nº de frases	Nº de paraules	Grandària vocabulari
Google	-	44.8G	2.2M

2.4 Neteja de corpus

Tots els corpus, tant del domini com de fora del domini (amb excepció de *Google n-grams*), han de passar per un pas previ de 'neteja' on les dades quedaran preparades per a l'entrenament de models. Algunes de les operacions que conté aquesta fase són:

- Separació del corpus en frases.
- Eliminació de símbols de puntuació (punts, comes, guions, etc.).
- Eliminació de les frases repetides que no aportaran res al model. Un exemple el trobem al corpus **UnitedNations**, on la paraula "Acta"apareixia moltes voltes com si fos una frase sencera.
- Conversió de tot el text a minúscules.

Així doncs, la següent frase extreta del corpus **europarl**:

```
Señora Presidenta, ¿se ha contabilizado mi voto, que no ha podido ser
 realizado electrónicamente, porque no tengo la tarjeta?
```

Quedarà així:

```
señora presidenta se ha contabilizado mi voto que no ha podido ser
realizado electrónicamente porque no tengo la tarjeta
```

Aquest procés es coneix al anglès com *tokenization*. L'objectiu és eliminar tota la informació que no sigui imprescindible per al model i que puga causar confusió. A l'hora de fer proves de transcripció, a la referència disponible també li aplicarem les mateixes operacions de neteja.

Amb els corpus degudament processats, ja podem passar a l'entrenament i adaptació de models de llenguatge.

Capítol 3

Selecció aleatòria

3.1 Introducció

En aquest capítol veurem una introducció descripció a la selecció de dades per a l'entrenament de models de llenguatge, i es desenvolupa una de les tècniques disponibles per a la tasca, la selecció aleatòria. Està dividit en quatre seccions: introducció a la selecció, selecció aleatòria, experimentació realitzada i una breu conclusió.

3.2 Introducció a la selecció

La selecció és una tècnica d'adaptació que consistix en l'extracció (intel·ligent) de dades d'un o més corpus per a l'entrenament de models de llenguatge. Ve motivada per dos objectius:

- Millorar els resultats de classificació. Les dades de fora del domini, al ser habitualment més nombroses, proporcionen de robustesa sintàctica al models de llenguatge finals, a canvi de poder introduir soroll. Si seleccionem les dades adequades podem donar-les un pes més rellevant en els nostres models. Discriminant les que no es consideren aptes per a la tasca, minimitzarem el soroll.

- Reduir el cost, en temps i grandària, del entrenament i la classificació. Entrenar models amb menys dades dona lloc a models més xicotets i fàcils de emprar. Aquesta reducció té interès sobretot en sistemes amb recursos limitats.

Hem de fer notar, però, que en molts sistemes de reconeixement de la parla ja s'apliquen tècniques amb objectius que es solapen amb els de la selecció. Per exemple, les tècniques de ajust de paràmetres per a interpolació lineal[JM80] de models de llenguatge atorguen major pes a les distribucions que millors resultats atorguen, fent servir un corpus del domini de prova. Al capítol 4 veurem com aquest solapament pot afectar als resultats obtinguts.

Per altra banda, les tècniques de poda de models de llenguatge, com la descrita en [Sto00], busquen reduir la grandària del model de llenguatge. La diferència fonamental entre les tècniques de selecció i les de poda és que les de poda treballen sobre el model de n-grames, mentre que la selecció treballa sobre les dades d'entrenament sense processar. Els toolkits de modelat com *SRILM* tenen la opció de realitzar distints tipus de poda.

El fet de que la selecció sigui una tècnica a aplicar a nivell de frase limita el seu ús a quan les dades d'entrenament estiguen sense processar,és a dir, es dispose de les frases que donaran lloc als models de llenguatge. Com hem vist a la secció 2.3.1, del corpus *Google n-grams* només disposem dels comptes de n-grames, i no dels textos que han donat lloc als comptes. Per tant, per a aquest corpus no podrem aplicar-li tècniques de selecció.

3.3 Selecció aleatòria

La tècnica de selecció aleatòria consistix en elegir aleatòriament les frases del corpus fora del domini que formaran part del subcorpus del model. És una tècnica ràpida i senzilla d'implementar. Per altra banda, no es valora en cap moment si les frases elegides aporten un benefici al model o no, pel que no tindrem garanties a priori de que el model millore.

3.4 Experimentació

En aquesta secció es mostren els resultats de la experimentació amb selecció aleatòria, amb l'objectiu de valorar la utilitat de la tècnica. Es varen entrenar models de llenguatge amb SRILM per a cada corpus i grandària de selecció desitjada i es va computar la perplexitat que donaven d'aquestos models amb el corpus de desenvolupament *dev*.

Al llarg de tot el projecte s'han fet servir models de trigrames, tant a l'hora de fer la selecció com a l'hora de provar els subcorpus seleccionats. Aquesta decisió ve motivada per la cerca d'un equilibri entre error del models i velocitat de l'entrenament i ús dels mateixos.

Per altra banda, i si no es diu el contrari, a l'hora de suavitzar models de llenguatges emprarem el suavitzat de *Kneser-Ney modificat* amb interpolació, amb vista dels bons resultats que presenta [CG99].

Per trobar la grandària òptima de selecció s'ha decidit seguir una aproximació logarítmica, decrementant a cada iteració la grandària del subcorpus seleccionat a la meitat. Els resultats els tenim a la Figura 3.1.

Com podem apreciar, la grandària ideal de selecció per a tots els corpus és del 100%, és a dir, tindrem resultats òptims si gastem els corpus sencers. També podem observar que en els corpus que gasten menys dades (com **epps** o **ted**) la perplexitat es dispara més ràpidament al fer la selecció.

D'aquest experiment podem concloure que la selecció aleatòria no dona bons resultats per a l'aprenentatge de models de llenguatge.

3.5 Conclusions

Els resultats són clars: la selecció aleatòria no és una bona tècnica a aplicar si pretenem reduir el error del model a entrenar. Aquest resultat és independent de les característiques del corpus d'entrenament, com la grandària o la proximitat al domini. En el següent capítol veurem una altra tècnica de selecció molt més interessant per a la nostra tasca.

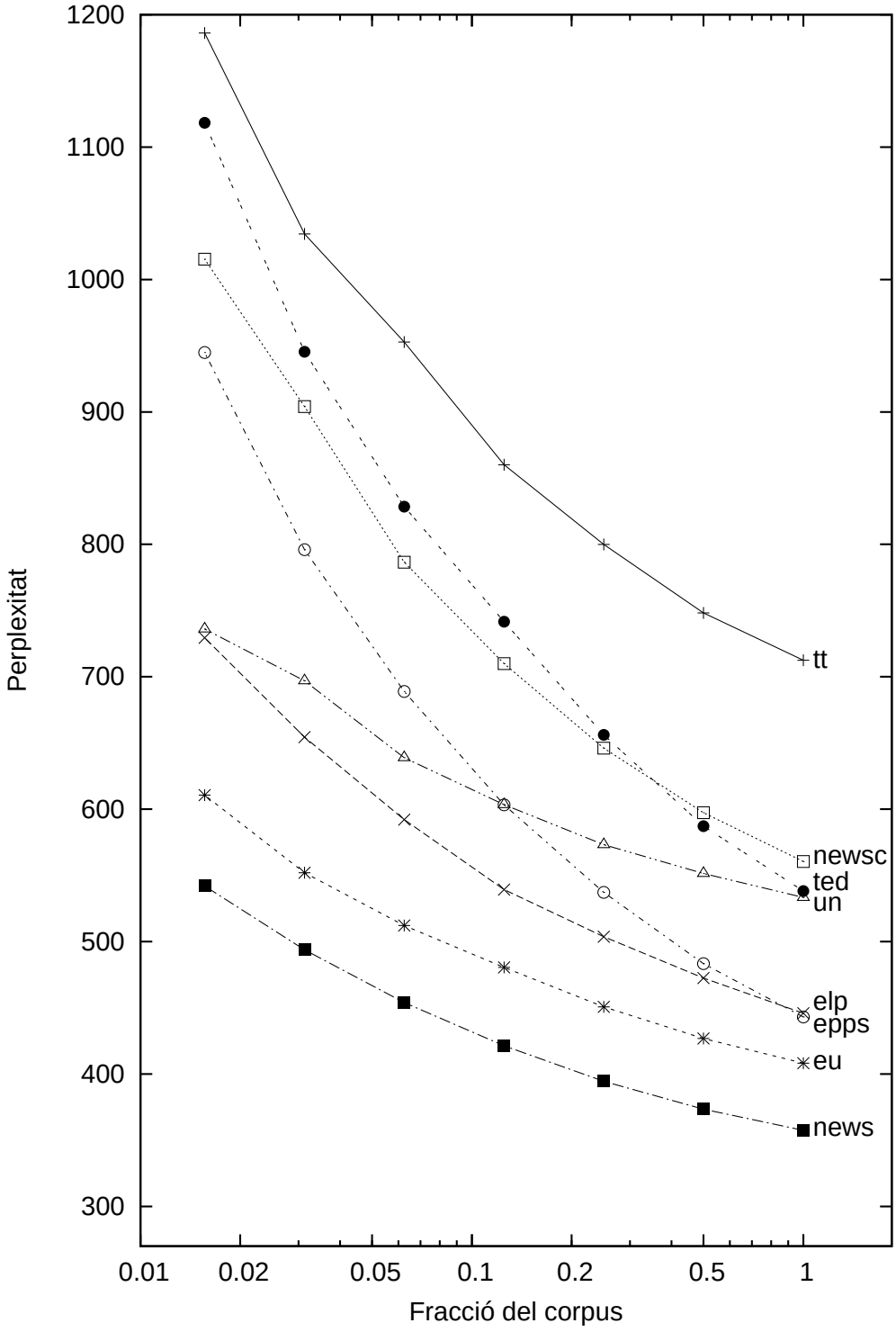

Figura 3.1: PPL en *dev* de cada corpus seleccionat aleatòriament

Capítol 4

Selecció per diferència de entropia creuada

4.1 Introducció

En aquest capítol es descriu una altra tècnica de selecció de dades, que coneguem com selecció per diferència d'entropia creuada (*cross-entropy difference selection* [ML10], selecció CED), així com els experiments duts a terme. La idea és aplicar aquesta tècnica d'adaptació a un cas real de transcripció de video-xarrades. Per a aquesta tasca, es disposa de les dades descrites al capítol 2.

El capítol es troba dividit en dos seccions: una primera en la que descriurem la selecció de CED amb la justificació teòrica, i una segona d'experimentació. Aquesta segona secció es subdividix en cinc parts:

1. Aplicació de la selecció als corpus de manera individual. Ens permetrà valorar si la selecció funciona com esperem.

2. Interpolació dels corpus seleccionats i comparació amb models entrenats amb totes les dades.

3. Aplicació de la selecció als corpus, tenint en compte la interpolació. Veurem si els resultats canvien respecte a la selecció individual.

4. Interpolació dels corpus amb la nova selecció i comparació amb models entrenats amb totes les dades.

5. Resultats de PPL i de WER del millor sistema.

Tanca el capítol una breu conclusió.

4.2 Selecció per diferència de entropia creuada

La selecció de models de llenguatge per diferència de entropia creuada fou descrita per Robert Moore i William Lewis en el seu treball *Intelligent Selection of Language Model Training Data*[ML10]. En esta secció es va a detallar en profunditat com funciona aquesta tècnica i la seua justificació teòrica.

La assumpció sobre la que treballa el mètode CED és que existix un subcorpus dins del corpus fora del domini que té les característiques estocàstiques del corpus del domini, és a dir, que ha sigut estret de la mateixa distribució del corpus del domini. El objectiu de la selecció és la extracció d'aquest subcorpus.

Donada una frase del corpus fora del domini, volem calcular la probabilitat de que pertanyi al subcorpus que volem. L'objectiu és puntuar totes les frases disponibles amb aquest criteri i després construir el subcorpus amb les que hagen obtingut una puntuació major. Pel teorema de Bayes, podem formalitzar aquesta probabilitat així:

$$p(N_I|s,N) = \frac{p(s|N_I,N)p(N_I|N)}{p(s|N)} \tag{4.1}$$

On N el corpus fora del domini, N_I és el subcorpus que busquem i s la frase actual. Com N_I és un subcorpus de N, podem simplificar $p(s|N_I,N)$ com

$$p(N_I|s,N) = \frac{p(s|N_I)p(N_I|N)}{p(s|N)} \tag{4.2}$$

Ara bé, la assumpció de partida és que N_I i I, on I és el corpus del domini, han sigut estrets de la mateixa distribució de probabilitat. Per tant, podem substituir $p(s|N_I)$ per $p(s|I)$ i tindrem

$$p(N_I|s,N) = \frac{p(s|I)p(N_I|N)}{p(s|N)} \tag{4.3}$$

Podem calcular $p(s|I)$ i $p(s|N)$ sense problemes, ja que disposem de ambdós models de llenguatge. Ens quedaria calcular la probabilitat de que el subcorpus N_I pertanya a N. Però aquesta probabilitat no és necessària per dos motius: és constant per a tota frase s i tampoc volem realment calcular $p(N_I|s,N)$. El objectiu d'aquest còmput és ordenar les frases per a l'extracció del subcorpus, i la inclusió o no del factor $p(N_I|N)$ no modificarà aquest ordre.

Ara podem moure'ns al domini dels logaritmes de forma que tinguem $log(p(s|I)) - log(p(s|N))$. L'últim pas és normalitzar aquest còmput per la longitud de la frase s. Això ens evitarà afavorir les frases més curtes, que tenen tendència a tindre valors superiors de $log(p(s|I)) - log(p(s|N))$ que les frases més llargues. Per tant, el còmput final per a puntuar serà $\frac{log(p(s|I)) - log(p(s|N))}{len(s)}$, que és precisament la diferència de entropies creuades $H_I(s) - H_N(s)$.

Com sabem, la entropia creuada $H(s)$ està relacionada amb la perplexitat de la forma $PPL = b^{H(s)}$, o b és la base front a la qual es vol calcular la perplexitat (habitualment 2). Per tant, a l'hora d'aplicar el mètode CED, hem de tindre en compte que la diferència només donarà un bon resultat de ordenació si les entropies creuades són comparables.

La normalització del vocabulari és un pas essencial per tal d'aconseguir aquesta tasca. És habitual que un corpus especialitzat continga paraules que els corpus més generals desconeguin, i a l'inrevés. La normalització ajuda a minimitzar el efecte nociu que poden tindre les paraules OOV (*out of vocabulari*, o fora del vocabulari) en el còmput de la entropia.

El problema addicional que tindrem, però, prové del fet de que es disposa de moltes mes dades de fora del domini que del domini a transcriure, que ens allunya de la comparabilitat desitjada. Per tractar de minimitzar-ho, es computaran els dos models de llenguatge (del domini i fora del domini) amb una grandària de corpus comparable. Amb aquest objectiu, a l'hora d'entrenar els models, seleccionarem aleatòriament un subcorpus del corpus fora del domini que tinga una grandària similar al corpus del domini.

4.3 Experimentació

Es descriu a continuació l'experimentació duta a terme amb la selecció CED. Com s'ha dit a la secció 3.4, s'han emprat en tots els experiments models de trigrames i suavitzat de *Kneser-Ney modificat* amb interpolació si no es diu el contrari.

4.3.1 Selecció dels corpus per separat

En aquesta secció es descriuen els resultats obtinguts de aplicar la tècnica de selecció CED als corpus fora del domini descrits a la secció 2.3. L'objectiu és doble: determinar quines frases de cada corpus són les apropiades per ser seleccionades i quin és la grandària òptima del subcorpus a seleccionar.

S'han seguit els següents passos per a cada corpus:

1. Extracció d'un subcorpus aleatori del corpus a seleccionar d'una grandària comparable al *train*.

2. Còmput del vocabulari conjunt del corpus *train* i del corpus a seleccionar. Això ens permetrà crear dos models de llenguatge amb el mateix vocabulari.

3. Entrenament dels models de llenguatge.

4. Còmput de la entropia creuada de cadascuna de les frases del corpus fora del domini.

5. Còmput de la diferència de entropia creuada de les frases i ordenació.
6. Extracció de subcorpus de distinta grandària, seguint la ordenació calculada.
7. Entrenament de un ML per cada subcorpus estret.
8. Càlcul de les perplexitats de cadascun dels ML amb el corpus *dev*.

En la Figura 4.1 podem veure els resultats obtinguts de la selecció CED.
A la Taula 4.1 podem veure de manera més clara els percentatges òptims de selecció computats, és a dir, aquells que han obtingut una perplexitat menor.

Taula 4.1: Percentatge òptim de selecció

	Corpus	Percentatge òptim
Corpus xicotets (<10M paraules)	EPPS	50%
	news-commentary	25%
	TED	25%
Corpus mitjans (entre 10M i 100M paraules)	elperiodico	12.5%
	europarl	25%
	tt	12.5%
Corpus grans (>100M paraules)	news	12.5%
	UnitedNations	6.25%

Per tal de millorar els resultats de Moore, es varen aplicar les següents modificacions:

- Proves en els distintes tècniques de suavitzat de MLs (en els models de llenguatge del pas 3). El suavitzat de Kneser-Ney amb backoff fou el que mostrà els millors resultats, encara que la diferència amb mKN interpolat fou menor d'un 1%.
- Repeticions. Per a puntuar cada frase s'aplica la tècnica CED varies voltes, modificant el subcorpus escollit aleatòriament (pas 1) com a representatiu del corpus fora del domini. Les puntuacions obtingudes es sumen i s'empren per a ordenar el corpus.
- Reducció successiva de la grandària. Aquesta tècnica s'ha de combinar amb les repeticions. Com farem diverses repeticions per puntuar una frase, podem reduir la grandària del subcorpus a seleccionar de manera progressiva. Així, per exemple, podríem realitzar 4 repeticions, puntuar i ordenar totes les frases, reduir el corpus a la meitat amb aquesta puntuació, i continuar el procés. Açò permet, per una banda, accelerar les repeticions, i per l'altra, fer una selecció millor, ja que al reduir les dades el subcorpus computat serà més representatiu del corpus sencer.

Els millors resultats es varen obtindre combinant les tres modificacions. La diferència de perplexitat front a la selecció descrita per Moore fou de fins un 3%.

4.3.2 Interpolació dels models

Es va provar a interpolar els models amb les grandàries òptimes calculades. A les Taules 4.2 i 4.4 es mostren els valors dels pesos computats per als models complets i seleccionats i la variació del pes produïda per la selecció, afegint o no a la interpolació un model entrenat amb *Google n-grams*. Com s'ha dit a la secció 1.6.1, aquest còmput s'ha realitzat amb **compute-best-mix**.

Taula 4.2: Pesos a la interpolació sense Google, selecció inicial

Corpus	λ sense selecció	λ amb selecció	Canvi de pes
train	0.5024	0.4219	$\downarrow$
EPPS	0.0453	0.0435	$\approx$
news-commentary	0.0031	0.0021	$\approx$
TED	0.0188	0.0072	$\downarrow$
elperiodico	0.0283	0.0444	$\uparrow$
europarl	0.0901	0.0801	$\downarrow$
tt2	0.0040	0.0087	$\uparrow$
news	0.2038	0.2648	$\uparrow$
UnitedNations	0.1042	0.1271	$\uparrow$

Taula 4.3: Pesos a la interpolació amb Google, selecció inicial

Corpus	λ sense selecció	λ amb selecció	Canvi de pes
train	0.4564	0.4001	$\downarrow$
EPPS	0.0468	0.0389	$\downarrow$
news-commentary	0.0026	0.0009	$\downarrow$
TED	0.0154	0.0072	$\downarrow$
elperiodico	0.0198	0.0243	$\uparrow$
europarl	0.0632	0.0533	$\downarrow$
tt2	0.0037	0.0039	$\approx$
news	0.0994	0.1580	$\uparrow$
UnitedNations	0.0464	0.0679	$\uparrow$
Google	0.2462	0.2446	$\approx$

Podem apreciar com els corpus més grans, que també són els que hem seleccionat més agressivament, tenen més pes a la interpolació amb la selecció que amb el models entrenats amb els corpus complets. Es pot justificar pel fet de que aquests corpus probablement tinguen més dades no relacionades amb el domini i que només afegixen soroll al model. Aquest fet també es dona en la interpolació amb Google.

A la Taula 4.4 podem veure el resultat de computar la perplexitat del corpus *dev* amb els models interpolats, sense i amb selecció, sense i amb Google.

Taula 4.4: PPL en *dev* dels models interpolats, selecció inicial

Corpus	PPLDev
Sense selecció	194.217
Amb selecció	197.964
Sense sel. + Google	173.383
Amb sel. + Google	174.89

Els resultats foren descoratjadors. Havent reduït la perplexitat de cada model per separat, el model interpolat resultava tindre una perplexitat menor si gastàvem els corpus de partida. En les següents apartats veurem com resoldre aquest problema.

4.3.3 Selecció amb interpolació

Es decidix passar a una altra estratègia per valorar la idoneïtat de la grandària de la selecció del cada corpus. En comptes de computar la perplexitat amb un model entrenat només amb el subcorpus corresponent, s'entrena aquest model i posteriorment s'interpola amb un model entrenat a partir del *train* (que és del domini) i un model generat a partir dels n-grames de Google. És amb aquest nou model interpolat amb el que calculem la perplexitat en *dev*.

La justificació d'aquest còmput és la següent: si calculem la perplexitat només amb un model entrenat amb el subcorpus seleccionat, no estarem tenint en compte que després el que volem és gastar aquest corpus interpolat amb altres. Això ens pot dur a que, per fer una selecció molt agressiva, estiguem perdent dades que sí són útils per al model i mantenint informació que poden aportar els altres corpus de la interpolació.

Encara que, idealment, deuríem provar la interpolació de cada possible grandària de cada subcorpus amb tota la resta de grandàries de tots els altres subcorpus, és completament inviable plantejar un experiment d'aquesta magnitud. Els corpus de *train* i Google s'han fet servir en aquest apartat per que representen dos extrems en els models:

- El corpus *train* és xicotet però conté molta informació del domini, pel que aportarà informació semàntica rellevant per a la tasca.
- El corpus Google és massiu i de domini molt general, per la qual cosa donarà robustesa sintàctica al model final.

A més, restringir la interpolació a tres models accelerarà un procés ja de per si prou costós, ja que requerix que es computen els pesos λ per a cada grandària de cada subcorpus a provar. Per reduir més el cost, no s'han provat percentatges menors que $\frac{1}{16}\%$, ja que donaren pitjors resultats en totes les proves realitzades.

Els resultats obtinguts amb aquesta estratègia es mostren a la Figura 4.2.

La Taula 4.5 ens mostra, a mode de resum, les grandàries òptimes calculades amb la nova estratègia.

Taula 4.5: Percentatge òptim de selecció amb interpolació

	Corpus	Percentatge òptim
Corpus xicotets (<10M paraules)	EPPS	100%
	news-commentary	100%
	TED	100%
Corpus mitjans (entre 10M i 100M paraules)	elperiodico	25%
	europarl	50%
	tt	50%
Corpus grans (>100M paraules)	news	25%
	UnitedNations	25%

D'aquesta taula podem treure que, tal i com passava en els experiments dels corpus per separat (taula 4.1), són altra volta els corpus més grans els que es beneficien de grandàries de selecció més xicotetes. En aquest cas, per als tres corpus que tenen menys de 10 milions de paraules, la grandària de selecció òptim és del 100%, és a dir, que hem de gastar el corpus complet. Cal destacar també que, sense excepció, les grandàries de selecció òptimes computades ara són superiors a les computades pel mètode sense interpolar.

Una volta (re)obtinguts les grandàries òptimes per als corpus, procedim a fer la interpolació amb dels corpus seleccionats i comparar-los amb els no seleccionats.

Per fer la interpolació, i de cara a fer testing amb un sistema reconeixedor, s'ha limitat el vocabulari dels models de llenguatge. El vocabulari final estarà composat de:

- Les 80000 paraules més freqüents que apareixen en els 8 subcorpus fora del domini (no inclou Google).
- Totes les paraules del corpus *train*, ja que està format de dades del domini i es considera interessant mantindre-les.
- Les 50000 paraules més freqüents de Google n-grams, si pertoca[1]. Açò permet reduir la grandària d'aquest corpus i fer factible treballar amb ell.

Adicionalment, cuidarem que el vocabulari no contiga nombres escrits en xifres (p.e. 12345). El nombre de paraules úniques final, és a dir, la grandària del vocabulari, és de 88289 per al cas sense Google, i de 95288 per als MLs que contenen Google.

Es llisten a continuació els resultats de calcular els valors λ òptims per interpolar el corpus *train* amb els corpus fora del domini, excloent Google n-grams. Es mostren

[1] Les paraules no s'han afegit al vocabulari quan no s'interpola amb aquest corpus

els resultats dels corpus seleccionats i sense seleccionar i si el pes ha augmentat o disminuït al seleccionar.

Taula 4.6: Pesos a la interpolació sense Google

Corpus	λ sense selecció	λ amb selecció	Canvi de pes
train	0.5024	0.4457	$\downarrow$
EPPS	0.0453	0.0426	$\downarrow$
news-commentary	0.0031	0.0014	$\approx$
TED	0.0188	0.0046	$\downarrow$
elperiodico	0.0283	0.0347	$\uparrow$
europarl	0.0901	0.0660	$\downarrow$
tt2	0.0040	0.0046	$\approx$
news	0.2038	0.2559	$\uparrow$
UnitedNations	0.1042	0.1445	$\uparrow$

Podem veure ara els resultats de optimitzar la interpolació per a MLs de tots els corpus d'entrenament disponibles. Cal recordar que el corpus Google n-grams no s'ha pogut seleccionar.

Taula 4.7: Pesos a la interpolació amb Google

Corpus	λ sense selecció	λ amb selecció	Canvi de pes
train	0.4564	0.4199	$\downarrow$
EPPS	0.0468	0.0417	$\downarrow$
news-commentary	0.0026	0.0007	$\downarrow$
TED	0.0154	0.0042	$\downarrow$
elperiodico	0.0198	0.0214	$\approx$
europarl	0.0632	0.0506	$\downarrow$
tt2	0.0037	0.0028	$\approx$
news	0.0994	0.1573	$\uparrow$
UnitedNations	0.0464	0.0758	$\uparrow$
Google	0.2462	0.2257	$\downarrow$

4.3.4 Resum de resultats

En la Taula 4.8 podem veure els resultats amb els models interpolats, tant en perplexitat com en WER, amb els corpus de Dev i Test. També s'han inclòs el nombre de bigrames i trigrames de cada model, que ens permet valorar la millora de grandària que s'obté amb la selecció.

Per al còmput del WER, s'ha fet servir el sistema de reconeixement de la parla del castellà desenvolupat per la UPV per al projecte transLectures, del mes T12. Aquest sistema és competitiu e inclou tècniques d'adaptació de models acústics com CMLLR [GB01].

Taula 4.8: Resum de resultats

Corpus	Nº Bigrames	Nº Trigrames	PPL Dev	WER Dev	PPL Test	WER Test
Sense selecció	21.5M	28.4M	194.217	24.08	241.929	25.23
Amb selecció	10.5M	9.5M	190.322	23.81	234.485	25.18
Sense sel. + Google	23.7M	56M	173.383	23.07	218.97	24.59
Amb sel. + Google	13.6M	42.3M	171.427	23.00	215.438	24.56

A la taula podem apreciar que, per al cas dels models sense Google n-grams, la selecció aporta una moderada millora en el WER amb una gran reducció de la grandària. El model amb selecció conté un terç dels trigrames del model complet i obté millors resultats tant en PPL com en WER, en *dev* i *test*.

En el cas dels models amb Google, la millora es menor però existent. Al no poder aplicar la selecció a Google, la reducció de grandària és molt menor. Encara i tot, els resultats són positius. Donem per fet que, si es disposés de les dades originals, veuríem una reducció de PPL i WER més gran.

4.4 Conclusió

S'ha presentat la tècnica de selecció per diferència d'entropia creuada i s'ha pogut comprovar experimentalment la utilitat d'aquesta tècnica en l'entrenament de models de llenguatge per a sistemes de ASR. S'ha vist que computar la perplexitat de cada model amb un conjunt de development per calcular la grandària òptima de selecció no proporciona els resultats esperats.

S'ha desenvolupat i provat una altra aproximació per mesurar aquesta grandària, que implica la interpolació amb un model entrenat amb dades del domini i un altre que aporte robustesa sintàctica. Aquesta aproximació ha mostrat millors resultats, i s'ha aconseguit reduir tant la grandària com la perplexitat final del model interpolat.

S'han provat els models sense selecció i amb selecció en un sistema reconeixedor de la parla competitiu. Així s'ha comprovat que la millora que hem vist en perplexitat es trasllada a una millora de WER.

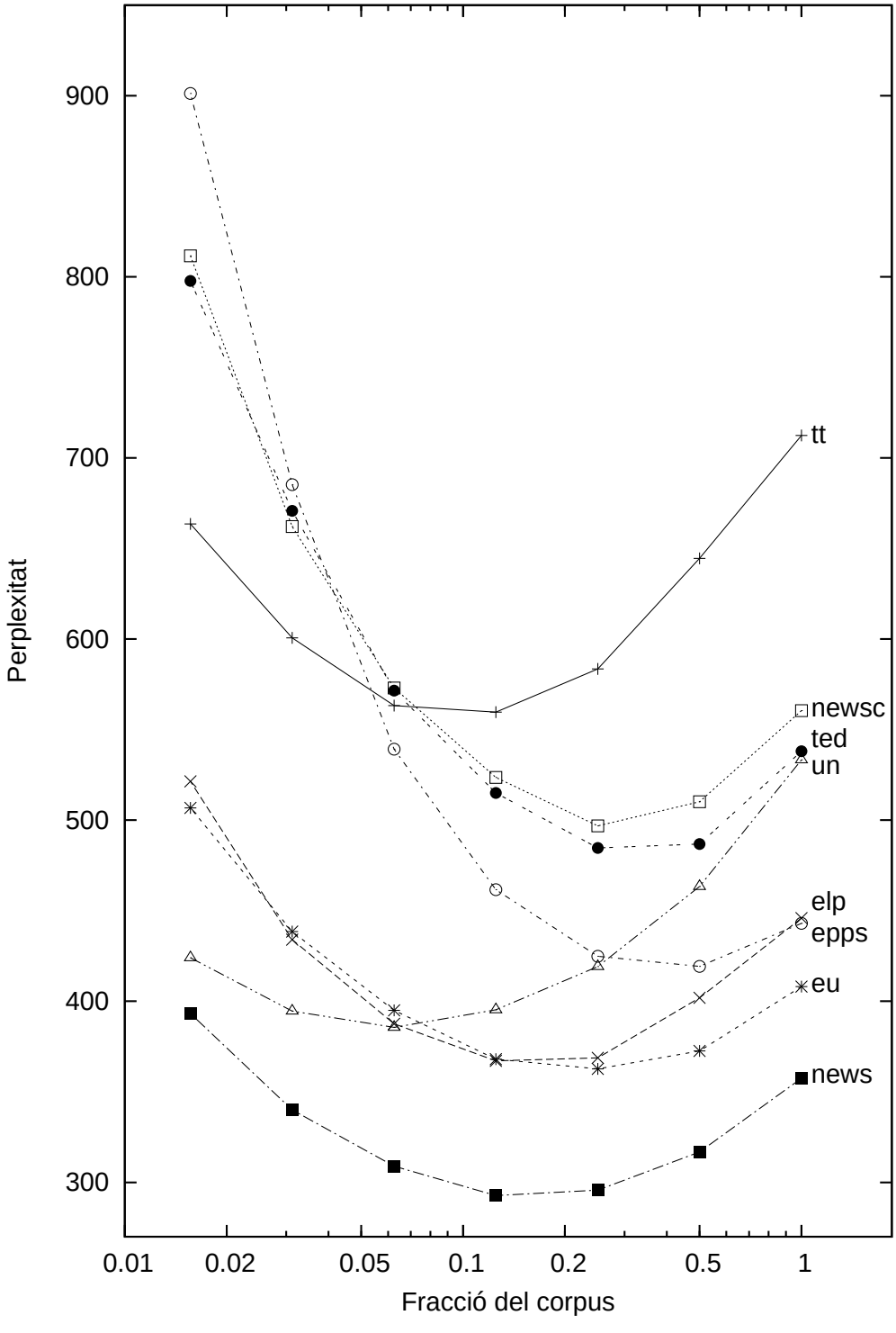

Figura 4.1: PPL en *dev* de cada corpus per selecció CED

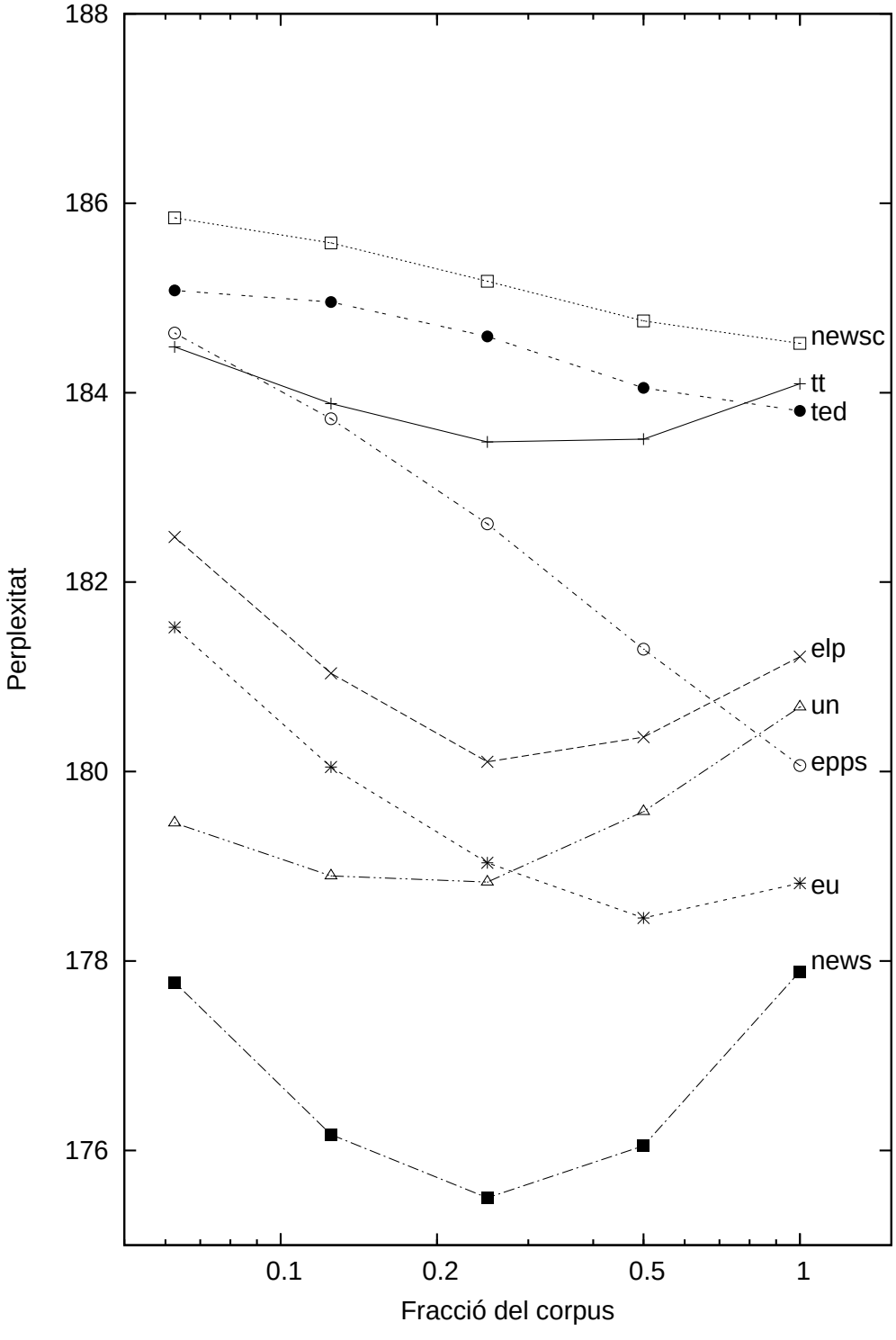

Figura 4.2: PPL en *dev* de selecció CED interpolant amb *train* + Google

Capítol 5

Conclusions i treball futur

5.1 Conclusions

L'objectiu inicial del projecte era el de provar tècniques de adaptació de models de llenguatge per a millorar la transcripció de video-xarrades. Es volia explotar que aquestes xarrades pertanyen a un àmbit concret, el acadèmic, per modificar els MLs, optimitzant-los per a la tasca.

Per tal de donar validesa als resultats, tota l'experimentació s'ha fet sobre una tasca de transcripció real, la transcripció de un subconjunt (23 en total) de vídeos de la plataforma poli[Media]. S'ha emprat un corpus del domini consistit en transcripcions de 690 vídeos de poli[Media] per l'entrenament de models. S'ha disposat de transcripcions d'altre subconjunt de 26 vídeos de pM per a desenvolupament. A més, s'han fet servir un total de 9 corpus externs, 8 dels quals es podran seleccionar al disposar de les dades original. En total, el volum de dades tractat és comparable al que podria gastar un sistema competitiu.

S'han emprat dos conjunts de ferramentes per a la tasca: SRILM per crear, modificar i treballar amb els models de llenguatge i TLK per provar els models obtinguts dins d'un sistema competitiu de transcripció automàtica de la parla. Aquestes ferramentes han facilitat enormement la feina realitzada.

S'ha aplicat una tècnica existent de selecció de dades, la selecció per entropia creuada o selecció CED. Sobre aquesta tècnica s'han realitzat diverses modificacions (repeticions, selecció amb interpolació, etc.) per millorar els models.

A més, hem pogut comprovar experimentalment que el fet de que un model de llenguatge done una perplexitat menor al processar un corpus concret no vol dir que al interpolar aquest model obtinguem millors resultats que els que dona un altre corpus amb major perplexitat, incloent el cas de que normalitzem els models per a que tinguen el mateix vocabulari i que ambdós models provinguen del mateix conjunt de

dades.

El resultat final ha sigut d'èxit, ja que s'ha aconseguit millorar tant el resultat de la transcripció (mesurat amb el WER) com reduir la grandària dels models de llenguatge. La diferència de WER amb els models dels corpus sense seleccionar és de 0.27 (*dev*) i 0.05 (*test*) sense incloure Google n-grams. Aquesta millora s'ha obtingut reduint a menys de la meitat el nombre de bigrames i a un terç el nombre de trigrames del model final. Incloent Google, la diferència de WER és de 0.07 (*dev*) i 0.03 (*dev*), amb una reducció al 57% del nombre de bigrames i al 75% el de trigrames.

Aquest resultat mostra que la qualitat de la transcripció no va directament correlada amb el volum de dades disponibles, que és un aspecte molt a tindre en compte en l'actualitat, on gràcies a les noves tecnologies s'ha facilitat l'accés a corpus cada volta més grans. Altres treballs com [GRST+12] han arribat a conclusions similars per a la tasca de traducció.

5.2 Futures línies de treball

Els bons resultats semblen confirmar que apostar per l'adaptació de models és garantia de millora dels sistemes reconeixedors. Es proposen ara diverses línies de treball derivades d'aquest projecte.

Hem vist com tindre en compte la interpolació a l'hora de decidir la grandària del subcorpus millorava considerablement el resultat de la selecció. Es podria intentar anar un xic més enllà e incorporar la interpolació en el scoring de les frases.

Una altra idea per trobar la grandària òptima de selecció seria prendre una aproximació greedy incremental. Així, es buscaria la grandària òptima primer del corpus que tinga un major pes a la interpolació, s'entrenaria un model amb les dades seleccionades, s'interpolaria i es passaria al següent corpus.

Per altra banda, s'està veient en els últims anys com sorgixen altres models de llenguatges no basats en n-grames com són les xarxes neuronals. Aquestes xarxes presenten en l'actualitat millors resultats que els models clàssics. Seria d'interès comprovar si l'adaptació per selecció pot aportar millores a aquestos models. Donat que la selecció és una tècnica de filtrat de dades prèvia a la construcció del model, a priori no hi ha res que ens indique el contrari.

Finalment, sembla interessant investigar si es pot aplicar la idea de la selecció a l'adaptació de models acústics. Encara que ara mateix el volum de dades que s'empra per entrenar models acústics no sol ser comparable amb el dels models de llenguatge, és possible que aquesta situació canvie en un futur proper. Per tant, les tècniques de entrenament de models deuen estar preparades per processar i filtrar les dades útils i descartar o minimitzar l'impacte de les dades no rellevants.

Bibliografia

[CBR98] Stanley F Chen, Douglas Beeferman, and Roni Rosenfield. Evaluation metrics for language models. 1998.

[CG99] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393, 1999.

[GB01] Asela Gunawardana and William Byrne. Discriminative speaker adaptation with conditional maximum likelihood linear regression. In *INTERSPEECH*, pages 1203–1206, 2001.

[GRST+12] Guillem Gascó, Martha-Alicia Rocha, Germán Sanchis-Trilles, Jesús Andrés-Ferrer, and Francisco Casacuberta. Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 152–161. Association for Computational Linguistics, 2012.

[Jel97] Frederick Jelinek. *Statistical methods for speech recognition*. MIT press, 1997.

[JM80] F. Jelinek and R. L. Mercer. Interpolated estimation of markov source parameters from sparse data. *Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–397, may 1980.

[ML10] Robert C Moore and William Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224. Association for Computational Linguistics, 2010.

[MSA+11] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[Sto00] Andreas Stolcke. Entropy-based pruning of backoff language models. *arXiv preprint cs/0006025*, 2000.

[Sto02] A. Stolcke. Srilm - an extensible language modeling toolkit. Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado, September 2002.

[tUT] The transLectures UPV Team. The translectures-upv toolkit (tlk). `http://translectures.eu/tlk`.

[Uni12] Universitat Politècnica de València. poli[media]. `https://polimedia.upv.es/`, 2012.

[WMT13] WORKSHOP ON STATISTICAL MACHINE TRANSLATION: shared task Machine Translation, 2013.

[YEG+02] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. The htk book. *Cambridge University Engineering Department*, 3, 2002.

Índex de figures

Índex de taules