UNIVERSITAT POLITÈCNICA DE VALÈNCIA
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Modelat Explícit de la Longitud en Traducció Automàtica Estadística

Treball d'Investigació - Doctorat en Informàtica

Joan Albert Silvestre Cerdà

Supervisat per:
Dr. Jorge Civera Saiz
Dr. Jesús Andrés Ferrer

8 de setembre de 2011

Resum

El modelat explícit de la longitud és un problema que ha estat explorat prèviament en diferents tasques de reconeixement de formes, oferint bons resultats. En aquest treball, es presenten dos models de longitud juntament amb dos mètodes d'estimació i dos parametritzacions alternatives per a traducció automàtica estadística (TAE). Concretament, hem incorporat els models de longitud com a característiques addicionals al model logarítmic-lineal d'un sistema de TAE estat de l'art basat en seqüències de paraules, amb l'objectiu d'estudiar la contribució de la informació de la longitud de les seqüències de paraules en el procés de traducció. Es mostren els resultats dels experiments que hem dut a terme en una tasca de referència de la TAE, els quals posen en relleu els beneficis del modelat explícit de la longitud.

Índex

Capítol 1

Introducció

1.1 Traducció Automàtica Estadística

La Traducció Automàtica Estadística (TAE) és una àrea de recerca relacionada amb la lingüística computacional i el reconeixement de formes que té com a objectiu proveir traduccions de texts entre dues llengües de forma automàtica, tot mitjançant l'ús de models estadístics inferits també de forma automàtica a partir d'exemples de traducció.

En TAE, el problema de la traducció es formula com la cerca de la frase destí (traducció) més probable $\hat{y}$ donada la frase origen x

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, p(y \mid x)\,. \tag{1.1}$$

La funció argmax es llegeix de la següent forma: de totes les possibles frases destí y, ens interessa aquella que maximitze el valor de probabilitat d'un model estadístic $p(y \mid x)$, és a dir, la traducció més probable $\hat{y}$. En general, modelar correctament aquesta distribució de probabilitat és complicat, motiu pel qual s'aplica la regla de Bayes per obtenir una expressió alternativa equivalent

$$\hat{y} = \underset{y}{\operatorname{argmax}} \frac{p(y)\, p(x \mid y)}{p(x)} = \underset{y}{\operatorname{argmax}}\, p(y)\, p(x \mid y)\,. \tag{1.2}$$

Aplicar la regla de Bayes ens ha permés descompondre la distribució de probabilitat original en dos termes: $p(y)$, anomenat model de llenguatge, i $p(x \mid y)$, model de traducció invers. Destacar que, donat que busquem la traducció més probable, podem prescindir del terme $p(x)$, ja que no depèn d'y i per tant no afecta al càlcul de la funció *argmax*. Aquesta aproximació es coneix com el model de la font i del canal [Brown 90], el qual defineix l'equació fonamental de la traducció automàtica estadística [Brown 93]. A la Figura 1.1 podem observar l'arquitectura general del procés de traducció automàtica estadística.

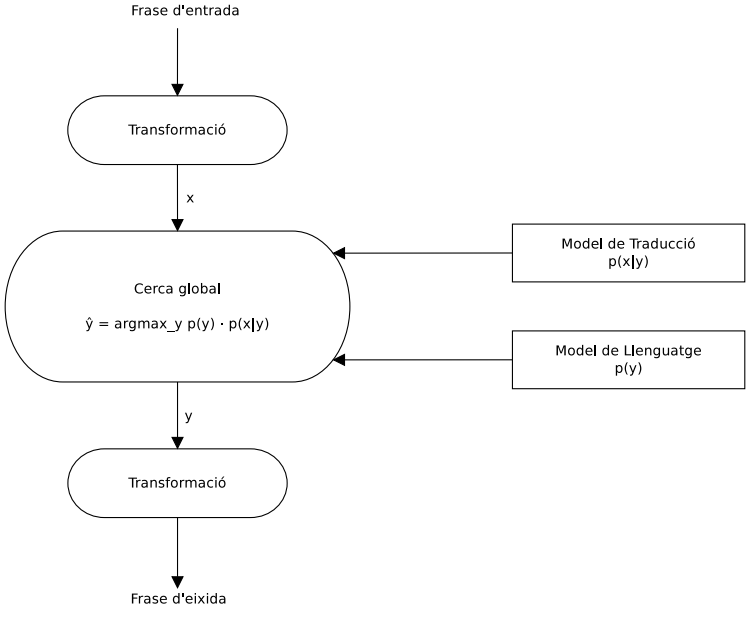


Figura 1.1: Arquitectura general del procés de traducció en TAE.

1.2 Models de llenguatge

Els models de llenguatge són un formalisme que permet modelar les propietats sintàctiques i semàntiques d'un llenguatge. D'alguna forma, un model de llenguatge restringeix les seqüències d'unitats lingüístiques permeses en el llenguatge. Així, en el context de la TAE, el model de llenguatge $p(y)$ s'encarrega d'avaluar com de probable és que la frase destí y siga gramaticalment correcta respecte al llenguatge destí. El models de llenguatge són àmpliament emprats en altres aplicacions del reconeixement de formes tals com reconeixement de la parla o reconeixement de text manuscrit.

Existeixen diverses aproximacions al modelat del llenguatge, sent la més comunament emprada la basada en n-grames. És precisament aquesta aproximació la que s'ha utilitzat en aquest treball i la que a continuació s'explica.

Models d'n-grames

L'aproximació d'n-grames als models de llenguatge basats es caracteritzen per modelar les concatenacions de paraules a través de probabilitats d'ocurrència de seqüències de paraules de longitud fixa n. Formalment, la probabilitat d'una frase o seqüència de paraules y ve donada per la següent expressió:

$$p(y) = p(y_1) \prod_{i=2}^{I} p(y_i \mid y_1^{i-1}) \,. \tag{1.3}$$

En aquesta equació, I és la longitud en paraules de la cadena y, y_i és la paraula i-èsima de la cadena y, y_1^{i-1} és la història de paraules, i $p(y_i \mid y_1^{i-1})$ és la probabilitat d'observar la paraula y_i una vegada s'ha observat la història y_1^{i-1}.

No obstant, és complicat estimar aquesta distribució de probabilitat degut a l'elevat nombre de paràmetres, que depén de la longitud de les cadenes, les quals poden ser arbitràriament llargues. És per aquest motiu que aquests models són normalment aproximats mitjançant el formalisme dels n-grames, limitant la longitud de la història considerada

$$p(y) \simeq p(y_1) \prod_{i=2}^{I} p(y_i \mid y_{i-n+1}^{i-1}) \,. \tag{1.4}$$

Acurtar la història ens permet estimar de forma més robusta els models de llenguatge, tot a costa de limitar la captura de dependències entre paraules. La decisió de quin valor adoptar per a n depén de les característiques del corpus d'entrenament emprat, principalment de les seves dimensions. Històricament, la majoria dels sistemes de TAE han emprat trigrames [Ney 95], però en l'actualitat s'entrenen models de 5-grames.

Estimació de models d'n-grames

L'estimació d'un model d'n-grames es realitza per màxima versemblança a partir d'un corpus d'entrenament, calculant la freqüència relativa d'aparició dels paràmetres del model

$$p(y_i \mid y_{i-n+1}^{i-1}) = \frac{N(y_{i-n+1}, \ldots, y_{i-1}, y_i)}{N(y_{i-n+1}, \ldots, y_{i-1})} \,, \tag{1.5}$$

on $N(\cdot)$ representa el nombre d'ocurrències de la seqüència de paraules considerada al corpus.

L'estimació dels paràmetres del model per freqüències relatives presenta un problema important: s'assigna probabilitat nul·la als esdeveniments no observats al conjunt d'entrenament. Per aquest motiu, els models d'n-grames són suavitzats per tal d'obtindre una distribució de probabilitat similar sense probabilitats nul·les. Existeixen diferents tècniques de suavitzat. A [Manning 99] podem trobar un estudi detallat sobre tècniques de suavitzat i tècniques d'estimació de models d'n-grames.

1.3 Models de traducció

Un model de traducció $p(x \mid y)$ és un formalisme estadístic que modela la probabilitat de traducció d'una determinada frase origen donada una frase destí. Aquesta distribució de probabilitat pot ser aproximada de diferents formes, atenent a la manera de caracteritzar el procés de traducció. La primera aproximació estadística que es va concebre fou l'aproximació basada en paraules, en la que el procés de traducció es duu a terme paraula a paraula. Aquesta aproximació va definir els fonaments a partir del quals s'erigiren altres aproximacions, com són les aproximacions basades en arbres jeràrquics, transductors estocàstics, o en seqüències de paraules. En aquesta

secció descriurem en primer lloc l'aproximació basada en paraules, per a posteriorment donar a conèixer l'aproximació basada en seqüències de paraules, considerada actualment com estat de l'art i que és precisament en la que s'ha basat aquest treball.

1.3.1 Models de traducció basats en paraules

Un model basat en paraules[1], com el seu propi nom indica, modela el procés de traducció a nivell de paraula, donant lloc al que anomenem traducció lèxica. Aquests models introdueixen el concepte d'alineament [Brown 90], que és un *mapping* entre paraules de les frases d'entrada i eixida. Considerant que

$$\begin{aligned} J = |x|\,,\ I = |y|\,, \\ x \equiv x_1\,,\ldots\,,x_J \equiv x_1^J\,, \\ y \equiv y_1\,,\ldots\,,y_I \equiv y_1^I\,, \end{aligned}$$

un alineament es defineix com una funció a que relaciona la j-èsima paraula de la frase d'entrada x, amb la i-èsima paraula de la frase d'eixida y [Koehn 10]

$$a : j \to i\,.$$

Aquesta funció està completament definida: per a tota paraula de la frase d'entrada x existeix una paraula de la frase destí y amb la qual està alineada

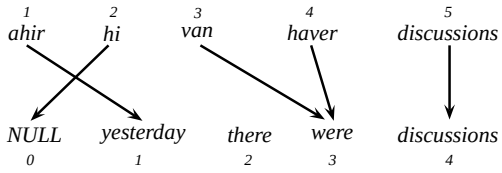
$$\exists i \in \{0,\ldots,I\} : a_j = i \qquad \forall j \in \{1,\ldots,J\}\,.$$

Aquesta restricció es compleix fins i tot quan una o més paraules d'x no es troben relacionades amb cap paraula de d'y, sent en aquest cas $a_j = 0$, on 0 representa un token especial anomenat NULL (paraula nul·la). A la Figura 1.2 es pot observar un exemple sintètic d'alineament entre una frase d'entrada en català i una frase d'eixida en anglès.

L'alineament entre paraules s'introdueix al model de traducció com una variable oculta, donat que els corpus d'entrenament disponibles es troben alineats a nivell de frase, però no a nivell de paraula

$$\begin{aligned} p(x \mid y) &= p(x, J \mid y) \\ &= p(J \mid y)\, p(x \mid J, y) \\ &= p(J \mid y) \sum_a p(x, a \mid J, y) \\ &= p(J \mid y) \sum_a p(a \mid J, y)\, p(x \mid a, J, y)\,. \end{aligned} \tag{1.6}$$

[1] *Word-Based Models*, en anglès.



$$a : \{1 \to 1, 2 \to 0, 3 \to 3, 4 \to 3, 5 \to 4\}$$

Figura 1.2: Exemple d'alineament entre dues frases en català x i en anglés y

El model de traducció original es descompon en dos termes: un model de longitud $p(J \mid y)$, que en la pràctica és simplificat assumint que la longitud de la frase d'entrada depén únicament de la longitud de la frase d'eixida, és a dir, $p(J \mid y) \approx p(J \mid I)$; i un model de traducció basat en paraules $p(x, a \mid J, y)$, que a la vegada es factoritza en un model d'alineament $p(a \mid J, y)$ i un model de traducció lèxic $p(x \mid a, J, y)$.

Models d'IBM

Amb tot, acabem de proposar un model de traducció $p(x, a \mid J, y)$ que intuïtivament proporciona la probabilitat de traduir y en x paraula a paraula, considerant únicament les relacions entre paraules definides per l'alineament a.

A partir d'aquest model es deriven els models de traducció d'IBM [Brown 93] que van revolucionar el món de la TAE. Existeixen un total de 5 models diferents, cadascun de major complexitat que l'anterior. El més senzill és el Model 1, el qual considera que la traducció es duu a terme prenent en compte únicament la traducció paraula a paraula prenent en compte l'alineament entre elles. Si factoritzem el model de traducció vist a l'Equació (1.6),

$$p(x, a \mid J, y) = \prod_{j=1}^{J} p(a_j \mid a_1^{j-1}, J, y)\, p(x_j \mid x_1^{j-1}, a, J, y)\,. \tag{1.7}$$

El Model 1 d'IBM simplifica els models lèxic i d'alineament prenent aquestes assumpcions:

$$p(a_j \mid a_1^{j-1}, J, y) \approx \frac{1}{I+1}\,,$$
$$p(x_j \mid x_1^{j-1}, a, J, y) \approx p(x_j \mid y_{a_j})\,,$$

i per tant, el model de traducció complet quedaria com segueix:

$$p(x \mid y) = p(J \mid y) \sum_{a} p(x, a \mid J, y) \approx p(J \mid I) \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(x_j \mid y_i)\,. \qquad (1.8)$$

La darrera equació ens revela una visió generativa del procés de traducció: en primer lloc, donada la longitud I de la frase d'eixida y, s'escull una longitud J de la frase d'entrada x d'acord amb la distribució $p(J \mid I)$. Aleshores, per cada posició de la frase d'entrada $1 \leq j \leq J$, s'escull una posició de la frase d'eixida a_j seguint una distribució de probabilitat uniforme, i finalment s'escull una paraula de la frase d'entrada x_j basant-nos en la distribució de probabilitat $p(x_j \mid y_{a_j})$.

El Model 1 d'IBM s'entrena mitjançant l'algoritme EM[2] [Dempster 77], ja que permet encarar el problema de la informació incompleta i de l'estimació de paràmetres de models probabilístics amb variables ocultes. Per obtenir informació més detallada sobre aquest i la resta de models d'IBM, remetem al lector a [Brown 93].

1.3.2 Models basats en seqüències de paraules

En la secció anterior hem presentat un model que es basa en la traducció de paraules aïllades. Com es pot preveure, traduir una frase paraula a paraula no és un procediment massa apropiat, donat que es perd per complet tota la informació de context disponible. Per aquest motiu, resulta més convenient escollir una unitat bàsica de traducció de major grandària que permeta capturar dita informació. Dita unitat, que pot englobar una o més paraules, rep el nom de seqüència de paraules[3].

Les seqüències de paraules s'obtenen dividint una frase completa en segments de paraules contigues i de longitud variable. La Figura 1.3 il·lustra com funcionen aquests sistemes: la frase d'eixida y és segmentada en seqüències de paraules, de forma que cada seqüència $\overline{y}_k$ és traduïda originant seqüències de paraules de la frase origen $\overline{x}_k$ que poden ser reordenades (permuta de posició en la frase) posteriorment.

Cal destacar, en primer lloc, que els sistemes de TA basats en seqüències de paraules no empren cap mètode lingüístic per segmentar una frase, i en segon lloc, que el poder obtenir segments de longitud variable permet "memoritzar" les seves respectives traduccions, podent arribar a emmagatzemar fins i tot traduccions de frases completes. No obstant, en la pràctica el nombre de paraules que pot abastar una seqüència és limitat (típicament a 7), doncs la complexitat d'aquests models (el nombre de paràmetres) creix exponencialment conforme a la longitud o grandària màxima que poden assolir les seqüències de paraules.

Definició del model de seqüències de paraules

Com hem esmentat adés, les frases origen i destí són segmentades en K seqüències de paraules

[2] *Expectation-Maximization Algorithm.*
[3] *Phrase*, en anglès.

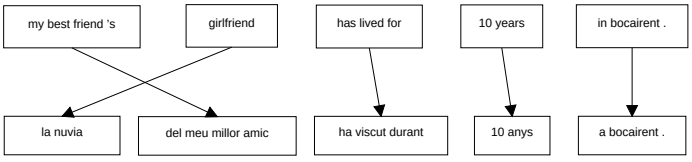


Figura 1.3: Exemple del procés de traducció automàtica en sistemes basats en seqüències de paraules.

$$x \equiv \overline{x}_1, \ldots, \overline{x}_k, \ldots, \overline{x}_K \,,$$
$$y \equiv \overline{y}_1, \ldots, \overline{y}_k, \ldots, \overline{y}_K \,.$$

Un model de traducció $p(x \mid y)$ basat en seqüències de paraules defineix la probabilitat de traduir la frase d'eixida y, segmentada en K seqüències de paraules, en la frase d'entrada x de la següent forma:

$$p(x \mid y) = \prod_{k=1}^{K} p(\overline{x}_k \mid \overline{y}_k) \,, \tag{1.9}$$

és a dir, $p(x \mid y)$ es defineix com el producte de la probabilitat de traduir cada seqüència de paraules de la frase destí $\overline{y}_k$ en la corresponent seqüència de paraules de la frase origen $\overline{x}_k$.

Extracció de seqüències de paraules

Donat un corpus de parells de frases $\{(x_n, y_n) \in C : n = 1, \ldots N\}$, es planteja el problema d'extracció dels parells de seqüències de paraules $(\overline{x}, \overline{y})$ a partir d'aquest conjunt. Una possible solució a aquest problema és considerar la segmentació de cadascuna de les frases del corpus d'entrenament com una variable oculta en el model, i estimar les segmentacions més probables per màxima versemblança aplicant un entrenament *Expectation-Maximization* [Dempster 77], tal i com ocorre als models basats en paraules d'IBM (veure Secció 1.3.1). Una altra alternativa que és àmpliament utilitzada als sistemes de traducció estat de l'art consisteix en obtindre les seqüències de paraules mitjançant l'ús de tècniques heurístiques que parteixen de la informació proveïda per els alineaments de paraules de cada parell de frases (veure Secció 1.3.1). En el nostre cas ens centrarem en la segon estratègia.

Essencialment es parteix dels alineaments a nivell de paraula en ambdues direccions de traducció, els quals es combinen de forma adient aplicant un algorisme heurístic. Posteriorment s'extrauen els parells de seqüències de paraules que siguen consistents amb els alineaments combinats.

Un parell de seqüències de paraules $(\overline{x}, \overline{y})$ és vàlid i consistent amb un alineament A si totes les paraules $x_1, \ldots, x_n \in \overline{x}$ que presenten alineaments en A troben totes les seves paraules alineades en la seqüència $\overline{y}$, i viceversa. Més formalment [Koehn 10]:

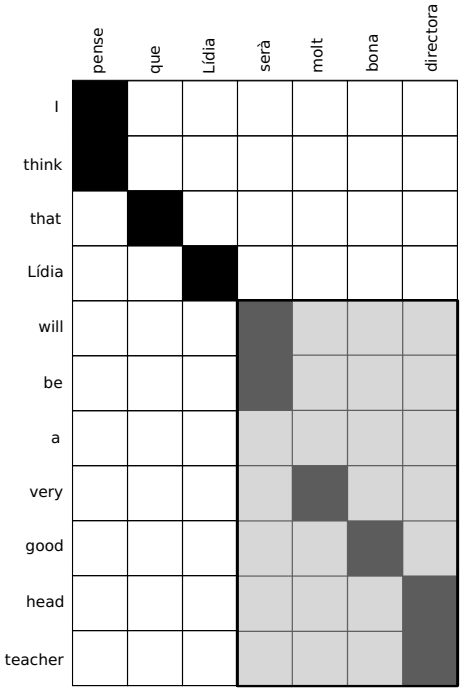


Figura 1.4: Exemple d'extracció de parells de seqüències de paraules a partir de l'alineament entre paraules d'un parell de frases.

$$
\begin{aligned}
(\overline{x},\overline{y}) \text{ és consistent amb } A \Leftrightarrow \quad & \forall x_j \in \overline{x} : (x_j, y_i) \in A \Rightarrow y_i \in \overline{y} \wedge \\
\wedge \quad & \forall y_i \in \overline{y} : (x_j, y_i) \in A \Rightarrow x_j \in \overline{x} \wedge \\
\wedge \quad & \exists x_j \in \overline{x},\, y_i \in \overline{y} : (x_j, y_i) \in A\,. \qquad (1.10)
\end{aligned}
$$

Cal notar que l'última condició determina que un parell de seqüències de paraules deu contenir almenys un alineament entre paraules. Per descomptat, es poden incloure paraules no alineades, ja que no es violaria cap de les condicions estipulades a l'Equació (1.10). Això significa que, quant menys alineaments, més possibles parells de seqüències de paraules a extraure, a excepció del cas extrem (cap alineament), doncs no es podria extraure cap seqüència de paraules al violar-se la tercera condició.

A la Figura 1.4 podem observar un exemple d'extracció de seqüències de paraules a partir d'un alineament entre paraules d'un parell de frases. En la quadrícula es mostren els alineaments existents entre les paraules que formen part de la frase en anglès *I think that Lídia will be a very good teacher* i en català *pense que Lídia serà molt bona directora*. Les zones negres denoten l'existència d'un alineament entre paraules, mentre que la zona gris (incloent les zones negres dels alineaments) determina un parell de seqüències de paraules extretes de forma vàlida (*serà molt bona directora* -

will be a very good head teacher). Fixem-nos en que les restriccions adés formulades (veure Equació (1.10)) es compleixen: d'una banda existeix almenys un alineament entre les paraules que formen part de les seqüències de paraules extretes, i d'altra banda s'abasta tot possible alineament de les paraules incloses. Aquest és només un dels nombrosos parells de seqüències de paraules que es poden extraure partint d'aquest alineament.

Cal remarcar que les seqüències de paraules que s'extrauran poden tenir una longitud variable, doncs poden abastar des de paraules aïllades fins frases senceres. No obstant, com ja hem esmentat adés, és molt convenient que la longitud de les seqüències de paraules extretes siga limitada, encara que cal tenir en compte que les seqüències de paraules llargues permeten capturar major informació de context. Ara, cal notar que les seqüències de paraules de major longitud es donen lloc amb molt menor freqüència que les seqüències més curtes, molt més freqüentment emprades per construir les traduccions de les frases d'entrada, però que presenten l'inconvenient d'aportar menys informació de context. La conclusió és que necessitarem tant de segments curts que podrem aplicar en nombroses ocasions, com segments llargs que ens permetran realitzar traduccions molt més precises.

Estimació del model

L'estimació d'un model de traducció de seqüències de paraules invers es realitza, a partir de les seqüències de paraules extretes de forma heurística a partir d'un corpus d'entrenament de parells de frases $\{(x_n, y_n) \in C : n = 1, \dots N\}$, de la següent forma [Koehn 10]:

$$p(\overline{x} \mid \overline{y}) = \frac{N(\overline{x}, \overline{y})}{\sum_{\overline{x}'} N(\overline{x}', \overline{y})}\,, \tag{1.11}$$

on $N(\overline{x}, \overline{y})$ és el nombre de vegades que s'ha extret el parell de seqüències de paraules $(\overline{x}, \overline{y})$.

Reordenament de seqüències de paraules

Com hem vist a l'Equació 1.9, el model de traducció s'ha definit de forma monòtona, sense contemplar el reordenament de les seqüències de paraules. No obstant, l'ordre en que apareixen les seqüències de paraules pot ser diferent atenent als idiomes considerats, i per tant, es mostra convenient habilitar la possibilitat de permutar la posició de les seqüències de paraules d'entrada (traducció no monòtona). En general, és preferible que no es produisquen reordenaments, o si es produeixen que siguen mínims, doncs els grans canvis de posició són infreqüents i, per tant, poc probables (encara que aquesta afirmació depèn de les característiques dels llenguatges considerats). En aquest sentit, cal penalitzar els salts llargs al reordenar les seqüències de paraules, així com premiar la monotonicitat del procés de traducció.

Per a tal fi s'introdueix un model de distorsió o reordenament que es defineix com una funció exponencial $d(\gamma) = \alpha^{\gamma}$ amb un valor apropiat del paràmetre $\alpha \in [0, 1]$,

sent γ la distància en paraules del salt efectuat per la seqüència de paraules al ser reordenada. La distància del salt γ es calcula de la següent forma:

$$\gamma = |\text{inici}_k - \text{fi}_{k-1} - 1| \,, \tag{1.12}$$

on inici_k és la posició, respecte a la frase origen, de la primera paraula de la seqüència de paraules d'eixida que es tradueix a la k-èsima seqüència de paraules d'entrada, i fi_{k-1} és la posició de l'última paraula de la seqüència de paraules d'eixida que es tradueix a la $(k-1)$-èsima seqüència de paraules d'entrada. Aleshores, si combinem aquest model amb el model de traducció monòton, obtenim un nou model proporcional al presentat a l'Equació (1.9)

$$p(x \mid y) \propto \prod_{k=1}^{K} p(\overline{x}_k|\overline{y}_k)\ d(|\text{inici}_k - \text{fi}_{k-1} - 1|) \,. \tag{1.13}$$

Aquest model de distorsió és extremadament simple, doncs no té en compte cap informació sobre les paraules i/o seqüències de paraules que intervenen en la traducció. Existeixen altres models de reordenament més complexos que permeten capturar informació lèxica i que estudiarem a la següent secció.

1.4 Moses: Un sistema de traducció estat de l'art basat en seqüències de paraules

Moses és un sistema de TA basat en seqüències de paraules estat de l'art, de codi obert, i desenvolupat per una gran comunitat coordinada per la Universitat d'Edimburg. Donat que aquest sistema implementa el que s'anomena un model logarítmic-lineal, en primer lloc introduirem a nivell general aquest formalisme per a posteriorment instanciar-lo al cas particular de *Moses*. Tota la informació que apareix en aquesta secció es pot ampliar consultant [Koehn 10].

1.4.1 Models logarítmic-lineals

Un model logarítmic-lineal[4] (log-lineal, d'ara endavant) és un formalisme que permet construir una distribució de probabilitat com a resultat d'integrar o combinar linealment dues o més distribucions de probabilitat (models), que en aquest context s'anomenem característiques o *features*. Formalment, un model log-lineal és una combinació lineal dels logaritmes de les funcions de probabilitat dels models en forma exponencial

$$p(x) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N} \lambda_i h_i(x)\right) \,, \tag{1.14}$$

on $p(x)$ és la distribució de probabilitat a modelar, N és el nombre de característiques a combinar, h_i és la característica i-èsima, els λ_i són els paràmetres del model (pesos)

[4] *Log-linear model*, en anglès.

associats a la característica i-èsima, i Z és un factor de normalització per assegurar que $p(x)$ és una distribució de probabilitat ben definida

$$Z = \sum_{x'} \exp\left(\sum_{i=1}^{N} \lambda_i h_i(x')\right) . \tag{1.15}$$

Instanciant un model log-lineal al cas de la TAE basada en seqüències de paraules, tenim que

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{k=1}^{K} \sum_{i=1}^{N} \lambda_i h_i(x, y, \overline{x}_k, \overline{y}_k)\right) , \tag{1.16}$$

on K és el nombre de seqüències de paraules en que es descomposa la frase d'entrada, de forma que cada característica $h_i(x, y, \overline{x}_k, \overline{y}_k)$ depèn del k-èsim segment, mentre que $Z(x)$ es defineix de forma a anàloga a l'Equació (1.15). Si integrem aquest model en el criteri de decisió de l'Equació (1.1),

$$\hat{y} = \underset{y}{\operatorname{argmax}} \left[\frac{1}{Z(x)} \exp\left(\sum_{k=1}^{K} \sum_{i=1}^{N} \lambda_i h_i(x, y, \overline{x}_k, \overline{y}_k)\right)\right] . \tag{1.17}$$

Donat que busquem aquella frase e que maximitze la probabilitat del model mitjançant la funció argmax, podem simplificar aquesta expressió, en primer lloc prescindint de $Z(x)$, ja que és una constant que no afecta al criteri de decisió argmax; i en segon lloc eliminant la funció exp, doncs es tracta d'una funció monòtona creixent que tampoc afecta al resultat de la funció argmax

$$\hat{y} = \underset{y}{\operatorname{argmax}} \sum_{k=1}^{K} \sum_{i=1}^{N} \lambda_i h_i(x, y, \overline{x}_k, \overline{y}_k) . \tag{1.18}$$

Els paràmetres del model log-lineal λ_i deuen de ser ajustats de forma adient. A la Secció 1.4.4 s'explica com es duu a terme aquest ajust en *Moses*.

1.4.2 Característiques del model log-lineal

Fins al moment hem presentat tres models que usualment s'inclouen a un model de traducció log-lineal: model de traducció basat en paraules, model de distorsió i model de llenguatge. Tot seguit descriurem la resta de característiques que implementa *Moses* al seu model log-lineal.

Model de suavitzat lèxic

Per tal de minimitzar l'impacte negatiu que podria provocar una estimació incorrecta del model de seqüències de paraules per a un conjunt de paràmetres determinat, Moses introdueix l'anomenat *model de suavitzat lèxic*[5], el qual modela la probabilitat

[5] *Lexical weighting*, en anglès.

de traducció a nivell de paraula de la seqüència de paraules $\overline{y}$ en la seqüència $\overline{x}$ a partir de l'alineament a definit entre les paraules d'ambdós seqüències.

L'estimació d'aquest model ve donada per la següent equació [Koehn 10]:

$$l(\overline{y} \mid \overline{x}, a) = \prod_{i=1}^{I} \frac{1}{|\{j \mid (i,j) \in a\}|} \sum_{\forall (i,j) \in a} p(y_i \mid x_j)\,, \tag{1.19}$$

on $p(y_i \mid x_j)$ és la probabilitat de traduir la paraula x_j en y_i, i a és un alineament entre les paraules de les seqüències $\overline{x}$ i $\overline{y}$. Donat que una paraula y_i pot estar alineada amb més d'una paraula de $\overline{x}$, la probabilitat de traducció lèxica per a cada y_i és normalitzada pel nombre d'alineaments que presenta la paraula y_i.

Penalització per paraula

Un altre efecte negatiu que pot afectar al procés de traducció és la deficiència que presenten els models de llenguatge d'n-grames. A banda de que, per la seva naturalesa, modelen tant les frases correctes com les incorrectes del llenguatge destí des d'un punt de vista lingüístic, aquests models assumeixen que, per a qualsevol llenguatge, les frases curtes són molt més probables que les frases llargues. A major longitud de frase, major nombre d'n-grames que intervenen en el càlcul de la funció de probabilitat del model, el que equival a molts termes de probabilitat multiplicant-se, donant lloc a un valor de probabilitat cada cop més baix.

L'impacte negatiu que provoca la deficiència del model de llenguatge, des de la perspectiva del procés de traducció, es tradueix en una tendència del sistema a decantar-se per traduccions de menor longitud i possiblement incorrectes des d'un punt de vista lingüístic, en detriment de traduccions de major longitud i possiblement correctes. En termes de decisió de la traducció més probable, aquesta circumstància s'observa en que la baixa puntuació atorgada pel model de llenguatge a les traduccions més llargues i l'alta puntuació conferida a les traduccions més curtes pot alterar aquesta decisió (veure Equació (1.18)).

Per corregir aquest comportament no desitjat del sistema podem considerar dues opcions: bé emprar un model de llenguatge més complex que reduïsca o elimine per complet la deficiència que presenta el model actual, o bé incloure al model log-lineal una característica que compense l'efecte negatiu del model de llenguatge. El sistema *Moses* implementa la segon opció, amb la inclusió d'un factor ω anomenat *penalització per paraula*[6], el qual bonifica les traduccions generades de major longitud per contrarestar la sobrevaloració de les traduccions de menor longitud.

Penalització per seqüències de paraules

De la mateixa forma que resulta interessant controlar la longitud en paraules de la frase sencera, també és convenient regular el nombre de seqüències de paraules K en que es segmenta en parell de frases. Això s'aconsegueix mitjançant la inclusió al model log-lineal d'un factor ρ anomenat *penalització per seqüències de paraules*[7]. La

[6] *Word penalty*, en anglés.
[7] *Phrase penalty*, en anglès.

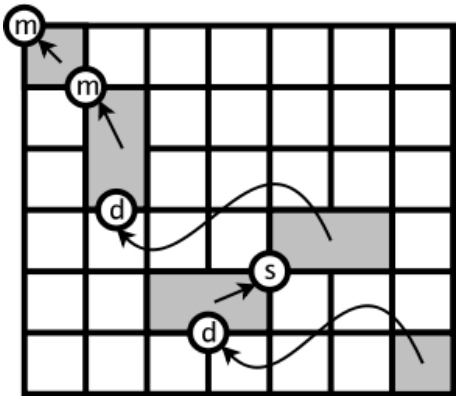


Figura 1.5: Exemples dels tres tipus d'orientacions que poden donar-se lloc a un model de reordenament lexicalitzat.

principal motivació de la inclusió d'aquesta característica és que totes les possibles segmentacions són equiprobables, sent la resta de factors (models de traducció, reordenament, llenguatge) els qui determinen, de forma indirecta, la millor segmentació possible.

Un escenari negatiu que es pot donar si no es realitza aquest control és el següent: si en el procés de traducció d'una frase s'ha emprat una seqüència de paraules molt llarga per generar la frase d'eixida, és molt probable que, per completar la frase, s'empren seqüències de paraules de curta longitud i de mala qualitat, generant-se finalment una traducció dolenta, però que, no obstant, pot tenir assignada una probabilitat alta per part del model de traducció, degut a que les seqüències de paraules molt llargues solen tenir associades grans valors de probabilitat de traducció. Per aquest motiu, aquestes frases poden arribar a erigir-se com les traduccions més probables, fet que no és desitjable.

La inclusió d'aquest factor afavoreix per tant aquelles segmentacions que donen lloc a un major nombre de seqüències de paraules. Per cada seqüència de paraules emprada per construir la frase d'eixida, s'afegeix un factor de bonificació ρ a la puntuació o probabilitat associada a la traducció, atenuant-se l'efecte negatiu provocat pel model de traducció al sobrevalorar traduccions possiblement incorrectes generades mitjançant l'ús de seqüències de paraules molt llargues.

Model de reordenament lexicalitzat

A la Secció 1.3.2 hem presentat un model de distorsió $d(\gamma)$ molt senzill que tant sols pren en compte la distància del moviment realitzat per cada seqüència de paraules de la frase destí al ser reubicada. Moses inclou addicionalment un model de reordenament més complex que pren en compte informació lingüística a nivell de seqüències de paraules, que rep el nom de *model de reordenament lexicalitzat*[8].

[8] *Lexicalized reordering model*, en anglès.

En termes generals, aquest model contempla tres tipus diferents de reordenament d'una seqüència de paraules respecte a la seqüència que la precedeix: monòton (m), intercanvi (s) i discontinu (d)[9], el significat dels quals es pot entendre de forma intuïtiva a la Figura 1.5. Aquest model prediu el tipus d'orientació $o \in \{m, s, d\}$ més probable que seguirà un parell de seqüències de paraules $\overline{x}, \overline{y}$, és a dir, $p(o \mid \overline{x}, \overline{y})$.

Aquest model s'estima a partir de les seqüències de paraules extretes heurísticament del corpus d'entrenament (veure Secció 1.3.2) de la següent forma:

$$p(o \mid \overline{x}, \overline{y}) = \frac{N(o, \overline{x}, \overline{y})}{\sum_{o'} N(o', \overline{x}, \overline{y})}, \tag{1.20}$$

on $N(o, \overline{x}, \overline{y})$ és el nombre d'ocurrències al corpus del parell de seqüències de paraules $(\overline{x}, \overline{y})$ seguint una orientació o respecte al parell de seqüències anterior. En la pràctica aquest model és suavitzat per evitar problemes de sobreentrenament [Koehn 10].

Existeixen altres variants d'aquest model de reordenament: per exemple, podem prendre en compte no només el tipus de reordenament existent respecte a la seqüència de paraules anterior, sinó també respecte a la seqüència de paraules posterior [Koehn 10].

1.4.3 El model logarítmic-lineal de Moses

El model log-lineal de Moses inclou les següents característiques:

- Associades al model de traducció (5):
  - Model de traducció de seqüències de paraules directe i invers: $p(\overline{y}_k \mid \overline{x}_k)$ i $p(\overline{x}_k \mid \overline{y}_k)$.
  - Model de suavitzat lèxic directe i invers: $l(\overline{y}_k \mid \overline{x}_k)$ i $l(\overline{x}_k \mid \overline{y}_k)$.
  - Penalització per seqüències de paraules ρ.
- Penalització per paraula ω.
- Model de distorsió uniforme $d(|\text{inici}_k - \text{fi}_{k-1} - 1|)$.
- Associades al model de reordenament (variable, amb configuració *msd-bidirectional-fe*):
  - Models d'orientació respecte a la seqüència de paraules anterior: $p(m, p \mid \overline{x}_k, \overline{y}_k)$, $p(s, p \mid \overline{x}_k, \overline{y}_k)$, i $p(d, p \mid \overline{x}_k, \overline{y}_k)$
  - Models d'orientació respecte a la seqüència de paraules posterior: $p(m, n \mid \overline{x}_k, \overline{y}_k)$, $p(s, n \mid \overline{x}_k, \overline{y}_k)$, i $p(d, n \mid \overline{x}_k, \overline{y}_k)$
- Model de llenguatge $p(\overline{e}_k)$.

[9] En anglès: *monotone* (m), *switch* (s), i *discontinous* (d).

1.4.4 Ajustament dels paràmetres del model log-lineal

Com hem comentat al final de la Secció 1.4.1, els paràmetres o pesos associats a les característiques del model log-lineal de Moses requereixen ser ajustats i optimitzats per tal de millorar la qualitat del sistema de traducció. El procés d'optimització d'aquests pesos es realitza mitjançant un entrenament per mínima taxa d'error, de l'anglès *Minimum Error Rate Training* [Och 03a], MERT d'ara endavant. Aquest entrenament té per objectiu trobar els valors òptims dels paràmetres del model log-lineal que maximitzen les prestacions del sistema en termes de BLEU, de forma iterativa i a partir d'un conjunt de validació V. Aquest conjunt de validació sol ser d'una grandària molt menor respecte del conjunt d'entrenament, i és aconsellable que continga frases no observades en l'estimació dels models de traducció i de llenguatge.

La convergència d'aquest algorisme es produeix quan el valor dels paràmetres del model log-lineal es modifiquen en un rang inferior a un llindar donat, o bé s'arriba a un nombre màxim d'iteracions. L'exploració de l'espai paramètric és una tasca complexa, degut a l'alta dimensionalitat. Existeixen algorismes com el *Simplex Algorithm* o *Powell Search* que resolen aquest problema [Koehn 10].

1.4.5 Procés de traducció

La tasca de traducció o descodificació és el procés en el que el sistema explora l'espai de cerca de les possibles traduccions de la frase d'entrada, escollint aquella que maximitza la probabilitat o puntuació conferida pel model log-lineal

$$\hat{y} = \underset{y}{\operatorname{argmax}} \sum_{k=1}^{K} \sum_{i=1}^{N} \lambda_i h_i(x, y, \overline{x}_k, \overline{y}_k)\,. \tag{1.21}$$

Malauradament, s'ha demostrat que el problema de la cerca de la traducció més probable és NP-Complet [Knight 99], ja que el nombre de possibles hipòtesis a explorar és exponencial respecte a la longitud de la frase d'entrada.

El procés de traducció consisteix en segmentar de totes les formes possibles la frase d'entrada, per a posteriorment traduir de diferents formes cada seqüència de paraules definida per la segmentació de la frase d'entrada, i per últim es genera la frase d'eixida de forma monòtona (d'esquerra a dreta i de forma incremental), reordenant de tota forma possible les seqüències de paraules traduïdes (un cas concret de traducció l'havem vist a la Figura 1.3).

L'espai de cerca és definit per estats o hipòtesis que, entre altres coses, determinen el nombre de paraules de la frase origen que han estat cobertes (traduïdes) i l'última seqüència de paraules de la frase destí generada. Durant el procés de cerca aquestes hipòtesis són expandides amb l'ús de noves opcions de traducció (traducció de seqüències de paraules de la frase d'entrada que engloben paraules no cobertes) per donar lloc a noves hipòtesis que abasten (tradueixen) paraules de la frase origen no cobertes anteriorment. Una hipòtesi que cobreix totes les paraules de la frase origen és un estat solució, i la traducció associada a tal estat s'obté recorrent el camí que parteix des de la hipòtesi inicial (cap paraula coberta) fins la hipòtesi solució.

El procés de cerca de la traducció més probable es realitza mitjançant un algorisme de cerca heurístic A* no complet, admissible i que segueix una estratègia de cerca per amplària. Per tal d'acotar l'espai de cerca i, en conseqüència, accelerar el procés de traducció, s'empren les següents tècniques:

- **Recombinació d'hipòtesis**: Les hipòtesis semblants, que són aquelles cobreixen les mateixes paraules de la frase d'entrada, són recombinades, conservant únicament aquella que presenta major probabilitat.
- **Limitació del paràmetre de distorsió**: En la pràctica el nombre de posicions que es poden saltar a l'hora de reordenar una seqüència de paraules d'entrada és limitada, tant per aspectes computacionals com per aspectes qualitatius del procés de traducció.
- **Poda explícita d'hipòtesis**: Les hipòtesis generades són distribuïdes en piles de capacitat limitada d'acord amb el nombre de paraules de la frase d'entrada que cobreixen, de forma que, quan una pila excedeix la seva capacitat, les hipòtesis menys probables són descartades. Cal notar que l'ús d'aquesta tècnica de poda provoca la pèrdua de l'admissibilitat de l'algorisme, ja que es corre el risc de podar estats o hipòtesis que poden conduir a la traducció òptima. Un aspecte important a tenir en compte és que, a menor capacitat d'aquestes piles, major acotament de l'espai de cerca i major acceleració de l'algorisme, però a la vegada major risc de podar hipòtesis prometedores.

En definitiva, aplicar aquestes tècniques de poda permet reduir la complexitat temporal del procés de cerca d'un ordre exponencial respecte al nombre de paraules de la frase d'entrada a un ordre quadràtic, tot a costa de perdre l'admissibilitat de l'algorisme. Per obtindre informació més detallada sobre el procés de traducció i l'algorisme de cerca, recomanem consultar [Koehn 10].

1.5 Avaluació de la qualitat de la traducció

L'avaluació de la qualitat dels sistemes de TA es realitza de forma automàtica comparant l'eixida del sistema (hipòtesi) amb la referència associada a la frase d'entrada mitjançant una mètrica d'avaluació. Algunes de les mètriques més emprades són les següents:

- **WER** (*Word Error Rate*) [Och 03b]: Aquesta fou la primera mètrica d'avaluació automàtica, adoptada directament dels mètodes d'avaluació dels sistemes de reconeixement de la parla. Consisteix en el càlcul del nombre mínim d'operacions elementals d'edició (substitució, esborrat i inserció) necessàries per convertir la frase d'eixida del sistema en la referència proporcionada. Les prestacions del sistema són inversament proporcionals a la mesura d'aquesta mètrica: a menor nombre d'operacions d'edició, millor és la qualitat del sistema, i viceversa.
- **TER** (*Translation Edit Rate*) [Snover 06]: És una mesura molt similar a la mètrica WER, a diferència que inclou el moviment / intercanvi de seqüències

de paraules com una operació elemental del mateix cost que les operacions d'inserció, esborrat i substitució.

- **BLEU** (*BiLingual Evaluation Undestudy*) [Papineni 01]: La taxa BLEU, la més comunament emprada en el món de la TA, mesura la precisió a nivell d'unigrames, bigrames, trigrames i quatrigrames de l'eixida del sistema respecte a la referència, i a més inclou una penalització a nivell de longitud de les hipòtesis, de forma que les traduccions curtes obtenen menor puntuació. La puntuació obtinguda per aquesta mètrica és directament proporcional a la qualitat del sistema: a major puntuació BLEU, millors prestacions, i viceversa.

Per a l'avaluació dels nostres sistemes emprarem la mètrica BLEU, donada la gran popularitat que té en aquesta disciplina.

Capítol 2

Modelat Explícit de la Longitud en TAE

Al capítol introductori hem donat una visió general de la disciplina de la traducció automàtica estadística, posant major èmfasi en l'aproximació basada en seqüències de paraules, ja que representa l'estat de l'art actual [Callison-Burch 10]. Com hem esmentat, s'ha definit el sistema de TAE Moses [Koehn 07] com a sistema base per implementar les millores que es proposen a aquest treball.

2.1 Motivació

A la Secció 1.4 s'ha presentat el model log-lineal de Moses, que inclou un conjunt de característiques relacionades amb models de traducció, models de reordenament, o el model de llenguatge, entre d'altres (veure Secció 1.4.3). No obstant, al model no s'inclou cap característica que prenga en compte de forma explícita la longitud de les seqüències de paraules, com per contra sí ocorre als models basats en paraules d'IBM (veure Secció 1.3.1). Existeix una característica que modela implícitament la longitud, el factor de penalització per seqüències de paraules (veure Secció 1.4.2), però s'estima insuficient si es pren en compte la benignitat que pot oferir un modelat explícit.

La idea és, per tant, incorporar una distribució de probabilitat al model log-lineal que prenga en compte les longituds de les seqüències de paraules que intervenen en el procés de traducció; una informació que, a priori, pot ser de gran utilitat per guiar el procés de traducció.

De fet, s'ha demostrat que el modelat de la longitud ha oferit resultats positius quan s'ha pres en consideració en aplicacions com reconeixement d'autors en text manuscrit [Uzuner 05], reconeixement d'escriptura i de veu [Zimmermann 02], o classificació de texts [Giménez 05]. Altres exemples de modelat explícit de la longitud es poden trobar en treballs relacionats amb el modelat del llenguatge [Kneser 96, Matusov 06] o l'alineament de frases bilingües [Brown 91, Gale 91].

2.2 Treball Relacionat

El modelat de la longitud en TAE ha estat una tasca poc explorada en recerca des de l'article més rellevant de Brown [Brown 93] fins els nostres dies. Els treballs més recents relacionats amb el modelat de la longitud s'han dut a terme baix l'aproximació estat de l'art, la basada en seqüències de paraules, a excepció de [Zens 06]. El primer a veure llum fou presentat a [Venugopal 03], on la raó entre la longitud de les seqüències de paraules d'entrada i d'eixida és emprada en l'extracció de seqüències de paraules i en la posterior estimació dels models. Zhao i Vogel [Zhao 95] plantejaren l'estimació d'un model de longitud de seqüències de paraules a partir d'un model de fertilitat de paraules [Brown 93], el qual integraren posteriorment en el seu sistema de TAE. En [Deng 08] és presentat un model de traducció de paraules a seqüències de paraules amb el seu corresponent model de longitud. Finalment, [Andrés-Ferrer 09] descriu la derivació i estimació d'un model de traducció basat en seqüències de paraules que inclou el modelat de les longituds de les seqüències d'entrada i d'eixida.

Ara bé, cap dels treballs anteriors ha proveït resultats mostrant la contribució del modelat explícit de la longitud en un sistema de TAE basat en seqüències de paraules estat de l'art, i és precisament aquest buit experimental el que s'ha pretés cobrir.

El present treball es troba inspirat en el model de longitud proposat en [Andrés-Ferrer 09], però aplicat a un sistema de traducció log-lineal basat en seqüències de paraules com és Moses [Koehn 07], amb l'objectiu d'estudiar els possibles beneficis del modelat explícit de la longitud en TAE.

2.3 Modelat Explícit de la Longitud

Com hem esmentat abans, el nostre treball s'inspira en [Andrés-Ferrer 09], on es presenta un model de traducció monòton de seqüències de paraules basat en Semi-Models Ocults de Markov[1], en el que es modela de forma explícita la forma en que es segmenten les frases d'entrada i d'eixida mitjançant la inclusió al model de traducció de dues variables ocultes, l i m, respectivament

$$p(x \mid y) = \sum_{l} \sum_{m} p(x, l, m \mid y)\,. \tag{2.1}$$

Aplicant la regla de la cadena a l'Equació (2.1), tenim que

$$p(x, l, m \mid y) = p(m \mid y)\, p(l \mid m, y)\, p(x \mid l, m, y)\,, \tag{2.2}$$

on $p(m \mid y)$ i $p(l \mid m, y)$ són models de longitud, mentre que el terme $p(x \mid l, m, y)$ és un model de traducció.

Tot seguit, factoritzem de forma independent cadascun dels termes de l'Equació (2.2) com segueix

[1] *Phrase-Based Semi-Hidden Markov Models* (PBSHMM)

$$p(m \mid y) = \prod_t p(m_t \mid m_1^{t-1}, y)\,, \tag{2.3}$$

$$p(l \mid m, y) = \prod_t p(l_t \mid l_1^{t-1}, m, y)\,, \tag{2.4}$$

$$p(x \mid l, m, y) = \prod_t p(x(t) \mid x(1), \ldots, x(t-1), l, m, y)\,, \tag{2.5}$$

on t abasta totes les possibles posicions de segmentació de la frase destí, l_t i m_t representen la longitud de la t-èsima seqüència de paraules d'entrada i d'eixida, respectivament, i $x(t)$ és la t-èsima seqüència de paraules d'entrada.

En aquest treball hem considerat que el model de longitud de l'Equació (2.3) és aproximat per la característica del model log-lineal penalització per seqüències de paraules, doncs aquesta modela de forma implícita la longitud de la seqüència de paraules d'eixida controlant el nombre de segments que intervenen en el procés de descodificació (veure Secció 1.4.2). Per la seva banda, l'Equació (2.5) és simplificada condicionant únicament la t-èsima seqüència de paraules d'entrada amb la t-èsima seqüència de paraules d'eixida. Aquesta assumpció converteix l'Equació (2.5) en el model de traducció, el qual ja es troba inclòs al model log-lineal de Moses (veure Secció 1.4.3)

$$p(x(t) \mid x(1), \ldots, x(t-1), l, m, y) := p(x(t) \mid y(t))\,, \tag{2.6}$$

amb el conjunt de paràmetres $\theta = \{p(u \mid v)\}$, per a cada seqüència de paraules d'entrada u i d'eixida v.

Per últim, l'Equació (2.4) ens permetrà derivar els models de longitud que seran inclosos com característiques addicionals al model log-lineal.

Atenent a la forma de parametritzar els models de longitud, considerarem dues variants: l'aproximació no paramètrica i l'aproximació paramètrica. En el primer cas, la distribució de probabilitat és aproximada amb una taula de contingències, és a dir, cada esdeveniment és considerat com una entrada (paràmetre) de la taula. En el segon cas, la distribució de probabilitat és aproximada amb un model paramètric, que en el nostre cas és una Poisson convenientment renormalitzada, amb l'objectiu de reduir el conjunt de paràmetres necessaris per estimar les probabilitats dels models, i conseqüentment, d'augmentar la robustesa del model.

2.3.1 Model de Longitud Estàndard

El model de longitud estàndard es deriva de l'Equació (2.4) assumint que la longitud de la seqüència de paraules d'entrada l_t depén únicament de la longitud de la corresponent seqüència de paraules d'eixida m_t

$$p(l_t \mid l_1^{t-1}, m, y) \approx p(l_t \mid m_t)\,. \tag{2.7}$$

L'aproximació paramètrica al model estàndard consisteix en assumir que aquesta distribució de probabilitat segueix una distribució de Poisson

$$p_{\lambda_{m_t}}(l_t \mid m_t) \propto \lambda_{m_t}^{l_t} \exp(\lambda_{m_t}) \,. \tag{2.8}$$

Cal tenir en compte que la funció de massa de probabilitat deu ser renormalitzada per tal de garantir que es tracta d'una distribució de probabilitat ben definida (tota la massa de probabilitat dels esdeveniments observables deu sumar 1), ja que la distribució de Poisson és infinita per a qualsevol enter positiu. Això és possible gràcies a que la longitud de les seqüències de paraules és limitada. El conjunt de paràmetres del model en aquesta aproximació és llavors $\gamma = \{\lambda_m\}$, per a tota longitud de les seqüències de paraules d'eixida m.

Per la seva banda, en l'aproximació no paramètrica cada terme $p(l_t \mid m_t)$ de l'Equació (2.7) representa un paràmetre, i conseqüentment, el conjunt de paràmetres del model és $\gamma = \{p(l \mid m)\}$, per a tota possible longitud de les seqüències de paraules d'entrada, l, i d'eixida, m. Donat que aquesta aproximació és més dispersa que la versió paramètrica (major nombre de paràmetres a estimar), s'aplica un suavitzat al model per evitar problemes de sobreentrenament, que consisteix en una interpolació lineal entre el model de longitud i una distribució uniforme de la longitud de les seqüències de paraules

$$\tilde{p}(l \mid m) := (1 - \varepsilon) \cdot p(l \mid m) + \varepsilon \cdot \frac{1}{M} \,, \tag{2.9}$$

amb un valor apropiat del factor d'interpolació ε. M representa la longitud màxima de les seqüències de paraules.

2.3.2 Model de Longitud Específic

El model de longitud específic s'obté prenent una assumpció més general a l'Equació (2.4), considerant la dependència de la longitud de la seqüència de paraules d'entrada amb la seqüència de paraules d'eixida, en lloc de la seva longitud

$$p(l_t \mid l_1^{t-1}, m, y) \approx p(l_t \mid y(t)) \,. \tag{2.10}$$

De forma similar al model estàndard, l'aproximació paramètrica assumeix que el model de l'Equació (2.10) segueix una distribució de Poisson

$$p_{\lambda_{y(t)}}(l_t \mid y(t)) \propto \lambda_{y(t)}^{l_t} \exp(\lambda_{y(t)}) \,, \tag{2.11}$$

amb la funció de massa de probabilitat convenientment renormalitzada, com s'ha explicat adés. El conjunt de paràmetres del model és $\gamma = \{\lambda_v\}$ per a tota seqüència de paraules d'eixida v. Cal ressaltar que la diferència entre l'Equació (2.8) i l'Equació (2.11) resideix en el nombre de paràmetres del model: en la primera s'assumeix una distribució de Poisson per a tota possible longitud de les seqüències de paraules d'eixida, mentre que en la segon s'assumeix una Poisson per a tota possible seqüència de paraules d'eixida. En el primer cas parlem d'uns pocs paràmetres, tants com la longitud màxima de les seqüències de paraules, mentre que en el segon cas parlem de milers, o inclús milions, depenent de la grandària del corpus.

Pel que fa a l'aproximació no paramètrica, de forma similar al model estàndard, cada terme $p(l_t \mid y(t))$ de l'Equació (2.11) conforma un paràmetre, donant lloc al conjunt de paràmetres $\psi = \{p(l \mid v)\}$.

Tal i com s'ha esmentat anteriorment, ambdós aproximacions del model de longitud específic són molt més disperses en comparació amb al model estàndard, ja que la probabilitat de la longitud de la seqüència d'eixida està condicionada no a un conjunt reduït de valors enters (longituds), sinó al conjunt de les seqüències de paraules d'eixida observades al corpus d'entrenament. Llavors, per minimitzar els problemes relacionats amb el sobreentrenament dels models específics, els seus paràmetres són suavitzats interpolant linealment el model específic amb el model estàndard

$$\tilde{p}(l \mid v) := (1 - \varepsilon) \cdot p(l \mid v) + \varepsilon \cdot p(l \mid |v|)\,, \tag{2.12}$$

sent $|\cdot|$ la longitud de la seqüència de paraules corresponent. El factor d'interpolació ε és ajustat empíricament en termes d'optimització de la taxa BLEU avaluada sobre un conjunt de validació.

2.3.3 Estimació dels Models

Els paràmetres dels models introduïts a les seccions anteriors es poden ser estimats per màxima versemblança mitjançant un entrenament EM [Dempster 77], atés que estem treballant amb un problema d'optimització de paràmetres amb variables ocultes, l i m. Tal i com es mostra a [Andrés-Ferrer 09], el model de traducció basat en seqüències de paraules és estimat com segueix

$$p(u \mid v) = \frac{N(u, v)}{\sum_{u'} N(u', v)}\,, \tag{2.13}$$

on $N(u, v)$ són els comptes esperats del parell de seqüències de paraules (u, v). Per la seva banda, l'estimació dels paràmetres del model estàndard es realitza de la següent forma

$$p(l \mid m) = \frac{N(l, m)}{\sum_{l'} N(l', m)}\,, \tag{2.14}$$

on

$$N(l, m) = \sum_{u,v} \delta(l, |u|)\, \delta(m, |v|)\, N(u, v)\,, \tag{2.15}$$

sent δ la delta de Kronecker.

Pel que respecta a l'aproximació paramètrica del model estàndard, l'estimació del paràmetre λ_m es realitza de forma semblant a l'Equació (2.14)

$$\lambda_m = \frac{\sum_l l \cdot N(l, m)}{\sum_l N(l, m)}\,. \tag{2.16}$$

Els paràmetres tant de l'aproximació paramètrica com de l'aproximació no paramètrica del model específic $p(l\,|\,v)$, són estimats de forma anàloga.

No obstant això, en els sistemes convencionals de TAE basats en seqüències de paraules, com és el cas de Moses, els comptes esperats $N(u,v)$ són aproximats amb uns comptes aproximats $N^*(u,v)$ obtesos mitjançant algorismes heurístics d'extracció de seqüències de paraules (veure Secció 1.3.2). En aquest sentit, proposem el nostre primer mètode d'estimació, que consisteix en aproximar $N(l,m)$ de l'Equació (2.14) amb el comptes obtinguts de forma heurística

$$N^*(l,m) = \sum_{u,v} \delta(l,|u|)\,\delta(m,|v|)\,N^*(u,v)\,. \tag{2.17}$$

Aquest mètode s'anomena estimació a partir de seqüències de paraules extretes en l'entrenament, o més breument emprant terminologia anglesa, estimació phrase-extract.

En aquest treball es proposa un segon mètode alternatiu d'estimació basat en la idea d'una aproximació per Viterbi a l'Equació (2.1). Aquest mètode únicament considera aquelles segmentacions de les frases d'entrada i d'eixida, l i m, respectivament, que maximitzen l'Equació (2.1)

$$\hat{l},\hat{m} = \underset{l,m}{\operatorname{argmax}}\ Pr(x,l,m \mid y)\}. \tag{2.18}$$

D'aquesta manera, els comptes de l'Equació (2.13) passen a ser comptes exactes en lloc de comptes esperats, motiu pel qual aquest mètode d'estimació és més realista.

El procés de cerca denotat per l'Equació (2.18) és dut a terme mitjançant l'ús d'un descodificador basat en un algorisme de cerca heurístic A^* d'un sistema de TAE basat en seqüències de paraules, que en el nostre cas és el descodificador de Moses (veure Secció 1.4.5). Es tracta d'un procés de cerca guiada en el que es coneixen tant la frase d'origen com la corresponent traducció, doncs ens interessa obtenir les segmentacions d'ambdós frases que donen lloc a la traducció proveïda, que és precisament la més probable segons el sistema. Malauradament, el model log-lineal de Moses presenta una sèrie de deficiències provocades per l'ús de tècniques heurístiques que impedeixen, en un nombre de casos significatiu, obtindre forçadament la referència d'una frase d'entrada, motiu que redueix la robustesa d'aquest mètode d'estimació al minvar-se el nombre d'esdeveniments observables en l'entrenament dels models de longitud.

Capítol 3

Corpora i Experimentació

En aquest capítol s'avaluen els possibles beneficis del modelat explícit de la longitud en TAE basada en seqüències de paraules. En primer lloc, presentarem el corpus que s'ha emprat per dur a terme aquestes proves, i en segon lloc, descriurem els experiments i mostrarem els seus resultats.

3.1 Corpora

Els experiments d'aquest treball s'han dut a terme amb el corpus paral·lel anomenat Europarl [Koehn 05]. Aquest corpus, que constitueix una tasca de referència al camp de la TAE, arreplega les transcripcions de les sessions plenàries de l'Europarlament i les seves corresponents traduccions a un total d'11 llenguatges: anglès, alemany,

Taula 3.1: Estadístiques bàsiques del corpus Europarl-v3 per a la partició Anglès-Espanyol (M = Mega = 1.000.000, K = Kilo = 1.000).

Conjunt Entrenament	Monolingüe		Bilingüe	
Llenguatges	En	Es	En	Es
Total Frases	1.4M		965K	
Grandària Vocabulari	115.7K	167.6K	81.8K	113.0K
Total Paraules	38.3M	40.3M	20.3M	20.9M

	Validació				Test	
	dev2006		devtest2006			
Llenguatges	En	Es	En	Es	En	Es
Total Frases	2K		2K		2K	
Grandària Vocabulari	6.1K	7.7K	6.1K	7.8K	6.0K	7.8K
Total Paraules	58.8K	60.5K	58.1K	60.2K	59.2K	61.3K
Perplexitat (5-grames)	74	75	73	76	71	76

Taula 3.2: Estadístiques bàsiques del corpus Europarl-v3 per a la partició Anglès-Alemà (M = Mega = 1.000.000, K = Kilo = 1.000).

Conjunt Entrenament	Monolingüe		Bilingüe	
Llenguatges	En	De	En	De
Total Frases	1.4M		995K	
Grandària Vocabulari	115.7K	327.2K	74.6K	226.9K
Total Paraules	38.3M	36.7M	21.5M	20.4M

	Validació				Test	
	dev2006		devtest2006			
Llenguatges	En	De	En	De	En	De
Total Frases	2K		2K		2K	
Grandària Vocabulari	6.1K	8.8K	6.1K	8.7K	6.0K	8.8K
Total Paraules	58.8K	55.1K	58.1K	54.2K	59.2K	55.6K
Perplexitat (5-grames)	74	119	73	118	71	121

francès, italià, holandès, portuguès, danès, suec, finès, grec i espanyol. Concretament, s'ha avaluat la contribució dels models de longitud en les particions del corpus Anglès-Espanyol i Anglès-Alemà en la versió 3 de l'Europarl, que arreplega totes les transcripcions i traduccions entre els mencionats parells d'idiomes generades en les sessions del parlament esdevingudes entre abril de 1996 i octubre de 2006. La finalitat d'escollir dos parells diferents és determinar si la contribució dels models de longitud presenta dependències en els llenguatges implicats en la traducció.

El corpus es troba dividit en tres conjunts diferents: entrenament, validació i test. A la Taula 3.1 podem observar les estadístiques bàsiques de la partició Anglès-Espanyol del corpus, mentre que a la Taula 3.2 es mostra la mateixa informació per a la partició Anglès-Alemà. El conjunt d'entrenament es troba dividit en dos conjunts de dades. El primer és un conjunt monolingüe, emprat per entrenar els models de llenguatge, mentre que el segon és un conjunt de frases bilingües utilitzat per entrenar la resta de models probabilístics integrats com a característiques al model log-lineal: models de traducció, models de suavitzat lèxic i models de reordenament lexicalitzats. De forma similar, la partició de validació es troba dividida en dos conjunts: dev2006, emprat per optimitzar el pesos de les característiques del model log-lineal amb MERT (veure Secció 1.4.4); i devtest2006, utilitzat per ajustar el factor d'interpolació al suavitzat del model específic (veure Equació (2.12)).

Per últim, cal afegir que, per simplificar i millorar la qualitat del procés d'experimentació, el corpus ha requerit d'un preprocés consistent en tres passos: tokenització, eliminació de frases de longitud major que 40 paraules, i conversió de majúscules a minúscules.

3.2 Experimentació

En la present secció detallarem cadascun dels experiments realitzats, destinats a mesurar la contribució dels models de longitud proposats al capítol anterior. El conjunt de característiques exposades a la Secció 1.4.3 defineix el sistema base, el qual compararem amb els nostres sistemes ampliats que inclouen informació sobre el modelat de la longitud. Aquests sistemes afegeixen dues característiques addicionals al model log-lineal: un model de longitud directe i un model de longitud invers. Així, comparar els sistemes augmentats amb el sistema base ens permetrà comprovar si efectivament els models de longitud tenen un impacte positiu en la qualitat de les traduccions, i en conseqüència, si representen una millora de l'estat de l'art. Addicionalment, realitzarem un estudi comparatiu dels dos models de longitud proposats, el model estàndard i el model específic, junt amb les dues aproximacions paramètriques, l'aproximació per Poisson i l'aproximació no parametritzada, i els dos mètodes d'estimació de paràmetres, l'estimació per phrase-extract i l'estimació per Viterbi.

Els sistemes han estat entrenats per a les dues direccions de traducció possibles per a cada parell d'idiomes, és a dir: Anglès-Espanyol (En-Es), Espanyol-Anglès (Es-En), Anglès-Alemà (En-De) i Alemà-Anglès (De-En). Per a cada direcció de traducció s'ha limitat la longitud màxima de les seqüències de paraules, amb valors que oscil·len entre 3 i 7, per tal de comparar la tendència de cadascun dels sistemes. Per mesurar la qualitat o prestacions dels sistemes, s'ha emprat la mètrica BLEU (presentada a la Secció 1.5) avaluada sobre el conjunt de test. En tots els casos s'ha aplicat la tècnica de bootstrapping [Koehn 04, Bisani 04], que ens permet obtindre la taxa BLEU d'un conjunt d'avaluació amb un interval del 95% de confiança. Com a resultats dels experiments, prescindirem de mostrar els intervals per donar una major claredat a les gràfiques, i llavors proporcionarem únicament la mitjana dels seus extrems.

Cal destacar que en els primers experiments realitzats vam detectar un comportament inestable dels sistemes conforme variava la longitud màxima de les seqüències de paraules. Aquesta inestabilitat es manifestava a les gràfiques amb constants encreuaments dels perfils de les corbes, quan el que s'esperava obtindre eren perfils aproximadament paral·lels. Creiem que això estava provocat pel procés d'optimització dels pesos del model log-lineal, MERT. Cada punt dibuixat a les gràfiques representava un sistema al qual s'havien optimitzat els pesos de forma independent mitjançant un entrenament MERT. El problema venia motivat perquè, depenent de les condicions en que es realitze, un entrenament MERT pot arribar a provocar fluctuacions de fins a 5 dècimes de BLEU en l'avaluació del conjunt de test. Aquestes fluctuacions són les que provocaven el constant encreuament de les corbes, i que per tant, dificultaven l'extracció de conclusions fiables a partir dels resultats.

Per minvar aquest soroll, es va optar per optimitzar únicament els pesos del model log-lineal dels sistemes entrenats amb màxima longitud de seqüències de paraules igual a 7, bàsicament per dos motius: en primer lloc, perquè les millors prestacions en termes de BLEU generalment s'obtenen limitant la longitud a 7, i en segon lloc, perquè és en aquest context en el que creiem que els models de longitud poden aportar informació més valuosa, i, llavors, en el que deurien estar millor valorats els seus corre-

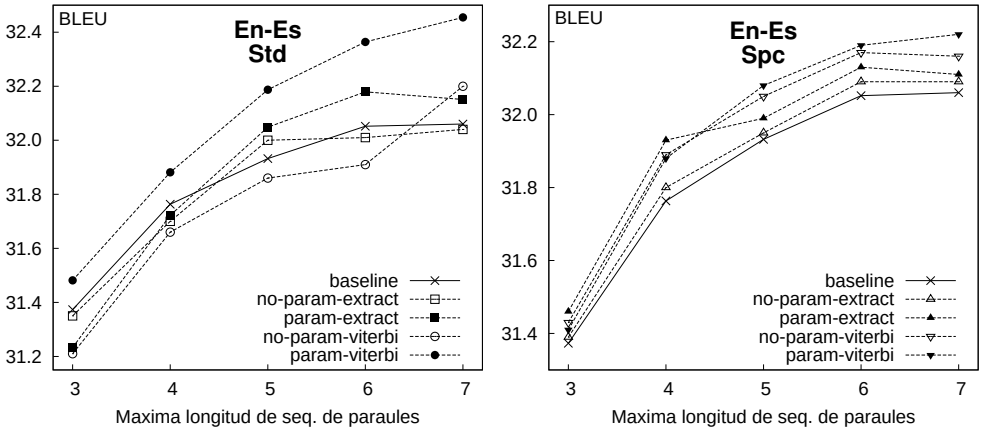


Figura 3.1: Comparació, en termes de BLEU, de les dues tècniques d'estimació (extract i viterbi) i de les dues parametritzacions (paramètrica i no paramètrica) per als models de longitud estàndard (Std) i específic (Spc) en la direcció de traducció Anglès-Espanyol (En-Es).

sponents pesos optimitzats. Així, els pesos per a màxima longitud 7 foren traslladats a la resta de sistemes entrenats amb una longitud màxima menor. El avantatges de procedir d'aquesta forma foren, en primer lloc, minvar la inestabilitat dels resultats, tal i com ja hem esmentat, i en segon lloc, reduir de forma substancial el temps d'experimentació, ja que cada entrenament MERT podia arribar a abastar 2 dies de còmput en paral·lel. Ara bé, cal tenir en compte que no optimitzar cadascun dels sistemes ha impedit obtindre els millors resultats possibles per a màximes longituds menors que 7, però això no ha estat cap problema donat que els sistemes ampliats s'han comparat amb els sistemes base que també han estat entrenats baix les condicions exposades.

3.2.1 Comparació dels mètodes d'estimació

La Figura 3.1 mostra l'evolució de la taxa BLEU (eix vertical) en funció de la màxima longitud de les seqüències de paraules en que ha estat limitat l'entrenament dels models (eix horitzontal), per tal d'estudiar el comportament dels dos mètodes d'estimació (extract i viterbi), així com les aproximacions paramètrica i no paramètrica per al model de longitud estàndard (Std) i específic (Spc) en la tasca Anglès-Espanyol (En-Es).

En cap dels dos casos existeixen diferències estadísticament significatives entre l'estimació phrase-extract i la estimació per Viterbi, com tampoc entre la versió paramètrica i la versió no paramètrica, i inclús quan els sistemes augmentats es comparen amb el sistema base, ja que els intervals de confiança (no mostrats) es solapen en tots els casos. No obstant això, el model de longitud estàndard millora les prestacions oferides pel sistema base, excepte quan es combina la versió no parametritzada amb l'estimació phrase-extract. En el cas del model específic, totes les combinacions pos-

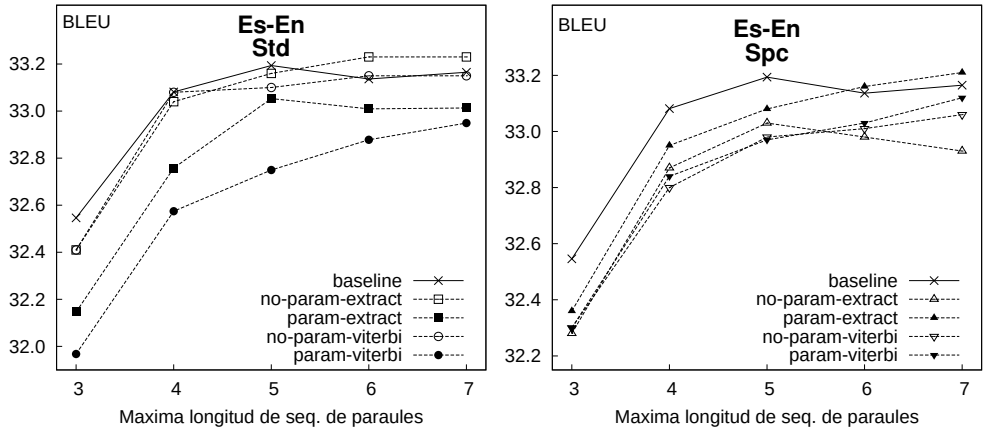


Figura 3.2: Comparació, en termes de BLEU, de les dues tècniques d'estimació (extract i viterbi) i de les dues parametritzacions (paramètrica i no paramètrica) per als models de longitud estàndard (Std) i específic (Spc) en la direcció de traducció Espanyol-Anglès (Es-En).

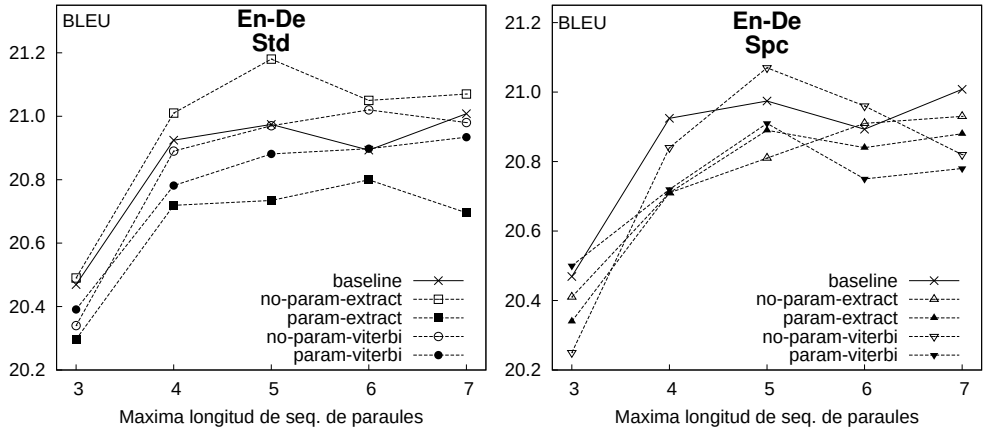


Figura 3.3: Comparació, en termes de BLEU, de les dues tècniques d'estimació (extract i viterbi) i de les dues parametritzacions (paramètrica i no paramètrica) per als models de longitud estàndard (Std) i específic (Spc) en la direcció de traducció Anglès-Alemà (En-De).

sibles abasten millors resultats que el sistema base. La millora més significativa sobre el sistema base ha estat de 0.4 punts de BLEU per al model estàndard parametritzat per Poisson i estimat per Viterbi.

Les conclusions que podem extraure d'aquests resultats són, d'una banda, que l'estimació per Viterbi proveeix millors resultats en BLEU, especialment per a valors

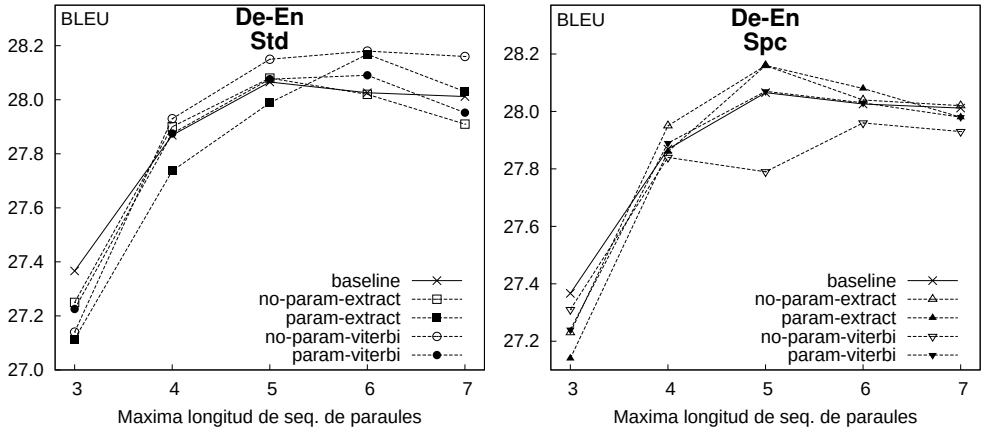


Figura 3.4: Comparació, en termes de BLEU, de les dues tècniques d'estimació (extract i viterbi) i de les dues parametritzacions (paramètrica i no paramètrica) per als models de longitud estàndard (Std) i específic (Spc) en la direcció de traducció Alemà-Anglès (De-En).

grans de la longitud màxima de les seqüències de paraules, on la contribució de la informació de la longitud sembla ser més important. Aquesta circumstància es dona a causa de les característiques de l'algorisme d'extracció de seqüències de paraules, el qual genera de forma heurística la població de les seqüències de paraules amb independència de quines són les segmentacions òptimes. D'altra banda, la parametrització per Poisson sembla ser més robusta que l'aproximació no paramètrica, oferint millors resultats en la majoria dels casos. Creiem que això es degut a causa de que la distribució de probabilitat que ofereix la parametrització per Poisson és més suavitzada que la versió no parametritzada, i que per eixe motiu s'aproxima més a la distribució de probabilitat real.

La Figura 3.2 mostra els mateixos resultats que la Figura 3.1 sols que per a la direcció de traducció Espanyol-Anglès. Tot i que no continuen havent-hi diferències estadísticament significatives entre els sistemes, en aquest cas s'observa com cap dels models de longitud aporten informació favorable al procés de traducció, obtenint-se resultats molt similars al sistema base, i en ocasions pitjors, especialment en les versions parametritzades del model de longitud estàndard, que casualment són les que millor funcionaven a la direcció de traducció inversa.

Les Figures 3.3 i 3.4 mostren gràfiques similars a les anteriors però per a les tasques Anglès-Alemà i Alemà-Anglès, respectivament. Els resultats segueixen un tònica semblant a l'observada a la tasca Espanyol-Anglès: no existeixen diferències estadísticament significatives, i s'obtenen resultats molt similars al sistema base, especialment en la direcció Alemà-Anglès. L'única excepció notable la trobem en la tasca Anglès-Alemà (Figura 3.3), en la que el model estàndard, en la seva versió no paramètrica estimada per phrase-extract millora de forma sistemàtica el sistema base, sent la màxima diferència de 0.2 punts de BLEU per a màxima longitud igual a 5.

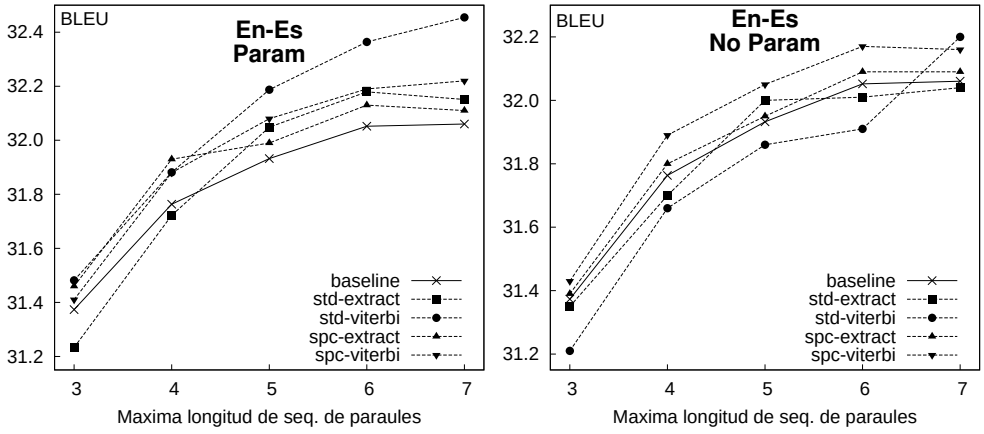


Figura 3.5: Valor de BLEU en funció de la longitud màxima de seqüències de paraules per a la parametrització de Poisson (Param) i per a versió no parametritzada (No Param), per tal de comparar els dos models de longitud juntament amb els dos mètodes d'estimació en la tasca Anglès-Espanyol (En-Es).

3.2.2 Comparació dels models de longitud

Com havem vist a les figures anteriors, els models de longitud sols aporten millores substancials al procés de traducció quan es tradueix de l'anglès a l'espanyol, mentre que per a la resta de direccions de traducció les millores són molt puntuals, i en molts casos es produeix una davallada de prestacions. Amb l'objectiu de comparar els dos models de longitud, ens centrarem únicament en la direcció Anglès-Espanyol, ja que careix de sentit comparar-los en escenaris en el que no aporta un benefici clar.

Així, en la Figura 3.5 es comparen directament les prestacions entre els models estàndard i específic, juntament amb els dos mètodes d'estimació, tant per a la parametrització per Poisson (Param) com per a la versió no parametritzada (No Param). Com ja s'havia observat adés, l'estimació per Viterbi millora les prestacions oferides pel mètode phrase-extract tant en el model estàndard com en el model específic. Quant a les diferències entre els models, si els comparem fixant el mètode d'estimació es poden observar dos casos diferents: d'una banda, si considerem l'aproximació parametritzada, el model de longitud estàndard supera de forma sistemàtica el model de longitud específic, especialment per a longituds màximes grans; mentre que en la versió no parametritzada és el model específic el que millora de forma sostinguda al model estàndard. Creiem que això és degut a que la distribució que segueixen les longituds de les seqüències de paraules condicionades a una seqüència concreta (model específic) no s'ajusta tant bé a una distribució de Poisson com ocorre al model estàndard, en el que es demostra que l'efecte suavitzador de la distribució de Poisson fa aproximar més la distribució del model a la distribució real.

Capítol 4

Conclusions i treball futur

4.1 Conclusions

En aquest treball hem presentat una forma de modelar de forma explícita la longitud en sistemes estat de l'art de TAE basats en seqüències de paraules. Hem plantejat dos models de longitud de seqüències de paraules juntament amb dos mètodes alternatius d'estimació i dues parametritzacions possibles. Els models proposats s'han inclòs com a característiques al model log-lineal del sistema de TAE estat de l'art anomenat Moses. Hem avaluat les prestacions dels models en les tasques de traducció Anglès-Espanyol, Anglès-Alemà, i les seves respectives inverses en el corpus Europarl-v3, amb l'objectiu de determinar si la contribució dels models de longitud és independent de la tasca considerada. Els resultats han provat que no és així: només a la tasca Anglès-Espanyol els models de longitud han aconseguit millorar de forma sistemàtica el sistema base de referència. En la resta de tasques s'han obtés resultats similars al sistema base, amb millores molt puntuals. Cal dir que, en cap cas, les diferències entre els sistemes comparats han estat estadísticament significatives, encara que en la tasca Anglès-Espanyol el model de longitud estàndard ha oferit una millora sostinguda respecte al sistema base de fins a 4 dècimes de BLEU.

Dels resultats obtesos en la tasca Anglès-Espanyol es poden extreure tres conclusions. En primer lloc, l'estimació per Viterbi és sistemàticament millor que l'estimació phrase-extract, sobretot quan els models tracten amb seqüències de paraules de major longitud. És un fet que calia esperar, degut a que l'estimació per Viterbi pren en compte únicament les seqüències de paraules provinents de les segmentacions òptimes. En segon lloc, l'aproximació paramètrica dels models mostra millors resultats que l'aproximació no paramètrica, gràcies a l'efecte suavitzador de la Poisson renormalitzada. Per últim, el model de longitud estàndard es mostra en general superior al model específic. A pesar de que el model específic deuria d'aportar més informació que l'estàndard, el seu elevat nombre de paràmetres ha derivat en un sobreentrenament del model que no ha estat possible de pal·liar aplicant les tècniques de suavitzat proposades.

4.2 Contribucions científiques

La realització d'aquest treball ha generat les següents publicacions:

- *Joan Silvestre-Cerdà and Jesús Andrés-Ferrer and Jorge Civera. Explicit Length Modelling for Statistical Machine Translation. In Proc. of the 5th Iberian Conference on Pattern Recognition and Image Analysis. Las Palmas de Gran Canaria (Spain), June 2011.*
  - Tipus: Conferència.
  - Rànquing: Core C.
  - Estat: Publicat en *Lecture Notes in Computer Science, Springer Link*, Volum 6669/2011, pàgines 273-280.
- *Joan Silvestre-Cerdà and Jesús Andrés-Ferrer and Jorge Civera. Explicit Length Modelling for Statistical Machine Translation. Pattern Recognition Journal.*
  - Tipus: Revista.
  - Rànquing: Factor d'impacte 2,607 JCR
  - Estat: Enviat a espera de revisió.

4.3 Treball futur

Com hem esmentat al capítol anterior, als experiments inicials vam observar un comportament extremadament inestable dels sistemes en termes de BLEU conforme variava la longitud màxima de les seqüències de paraules. Aquesta inestabilitat venia donada pel procés d'optimització de pesos del model log-lineal, MERT, i llavors, per minvar dits problemes, vam emprar els pesos optimitzats en màxima longitud 7 per a la resta de sistemes entrenats amb menor longitud màxima. A pesar d'això, no s'ha aconseguit eliminar per complet el soroll de les gràfiques. Per aquest motiu, com a treball futur pretenem emprar tècniques més robustes d'optimització com és l'Adaptació Baiesiana proposada a [Sanchis-Trilles 10].

Una altra possibilitat d'ampliar aquest treball seria estudiar altres aproximacions paramètriques dels models de longitud, com per exemple la distribució Gamma [Giménez 05], atés que la parametrització per Poisson ha donat millors resultats que les versions no parametritzades.

Per últim, tenim la intenció de realitzar un entrenament iteratiu per Viterbi dels models de longitud, tal com ocorre en un entrenament EM [Dempster 77]. Aquest procediment consistiria en entrenar els models de longitud i incorporar-los al model log-lineal, i posteriorment, amb el sistema ampliat, tornar a generar les segmentacions òptimes i reentrenar els models de longitud en successives iteracions. Creiem que iterar l'entrenament millorà de forma notable les prestacions del mètode d'estimació per Viterbi.

Bibliografia

[Andrés-Ferrer 09] Jesús Andrés-Ferrer & Alfons Juan. *A phrase-based hidden semi-Markov approach to machine translation.* In Procedings of European Association for Machine Translation (EAMT), pages 168–175, Barcelona, Spain, May 2009. European Association for Machine Translation.

[Bisani 04] M. Bisani & H. Ney. *Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation.* In IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 1, pages 409–412, Montreal, may 2004.

[Brown 90] P. F. Brown*et al. A Statistical Approach to Machine Translation.* Computational Linguistics, vol. 16, no. 2, pages 79–85, 1990.

[Brown 91] Peter F. Brown*et al. Aligning sentences in parallel corpora.* In Proc. of ACL, pages 169–176, 1991.

[Brown 93] P. F. Brown*et al. The Mathematics of Statistical Machine Translation: Parameter Estimation.* Computational Linguistics, vol. 19, no. 2, pages 263–311, 1993.

[Callison-Burch 10] Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki & Omar Zaidan. *Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation.* In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 17–53, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[Dempster 77] A. P. Dempster, N. M. Laird & D. B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm.* JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, vol. 39, no. 1, pages 1–38, 1977.

[Deng 08] Yonggang Deng & W. Byrne. HMM *Word and Phrase Alignment for Statistical Machine Translation.* IEEE Trans. Audio, Speech, and Lang. Proc., vol. 16, no. 3, pages 494–507, 2008.

[Gale 91] William A. Gale & Kenneth W. Church. *A program for aligning sentences in bilingual corpora.* In Proc. ACL, pages 177–184, 1991.

[Giménez 05] Adriá Giménez*et al.* *Modelizado de la longitud para la clasificación de textos.* In Actas del I Workshop de Rec. de Formas y Análisis de Imágenes, pages 21–28, 2005.

[Kneser 96] R. Kneser. *Statistical language modeling using a variable context length.* In Proc. of ICSLP, 1996.

[Knight 99] K. Knight. *Decoding complexity in word-replacement translation models.* Computional Linguistics, vol. 25, no. 4, pages 607–615, 1999.

[Koehn 04] Philipp Koehn. *Statistical Significance Tests for Machine Translation Evaluation*, 2004.

[Koehn 05] P. Koehn. *Europarl: A Parallel Corpus for Statistical Machine Translation.* In Proc. of the MT Summit X, pages 79–86, September 2005.

[Koehn 07] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin & Evan Herbst. *Moses: Open Source Toolkit for Statistical Machine Translation.* In ACL. The Association for Computer Linguistics, 2007.

[Koehn 10] Philipp Koehn. Stadistical machine translation. Cambridge University Press, Edinburgh (United Kingdom), 2010.

[Manning 99] Christopher D. Manning & Hinrich Schütze. Foundations of statistical natural language processing. The MIT Press, Cambridge, Massachusetts, 1999.

[Matusov 06] Evgeny Matusov*et al.* *Automatic Sentence Segmentation and Punctuation Prediction for Spoken Language Translation.* In Proc. of IWSL, pages 158–165, 2006.

[Ney 95] Hermann Ney, M. Generet & F. Wessel. *Extensions of Absolute Discounting for Language Modeling.* In Proc. of the Fourth European Conference on Speech Communication and Technology, pages 1245–1248, Madrid, Spain, September 1995.

[Och 03a] F. J. Och. *Minimum error rate training in statistical machine translation.* In Proc. of ACL'03, pages 160–167, Morristown, NJ, USA, July 2003. Association for Computational Linguistics.

[Och 03b] Franz Josef Och & Hermann Ney. *A Systematic Comparison of Various Statistical Alignment Models.* Computational Linguistics, vol. 29, 2003.

[Papineni 01] K. Papineni, S. Roukos, T. Ward & W. Zhu. *BLEU: a Method for Automatic Evaluation of Machine Translation.* Rapport technique RC22176, Thomas J. Watson Research Center, 2001.

[Sanchis-Trilles 10] G. Sanchis-Trilles & F. Casacuberta. *Log-linear weight optimisation via Bayesian Adaptation in Statistical Machine Translation.* In COLING, pages 1077–1085, 2010.

[Snover 06] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla & John Makhoul. *A study of translation edit rate with targeted human annotation.* In In Proceedings of Association for Machine Translation in the Americas, pages 223–231, 2006.

[Uzuner 05] Özlem Uzuner & Boris Katz. *A Comparative Study of Language Models for Book and Author Recognition.* In IJCNLP, pages 969–980. 2005.

[Venugopal 03] Ashish Venugopal*et al.* *Effective phrase translation extraction from alignment models.* In Proc. of ACL, pages 319–326, 2003.

[Zens 06] Richard Zens & Hermann Ney. *N-gram posterior probabilities for statistical machine translation.* In Proceedings of WSMT, pages 72–77, 2006.

[Zhao 95] Bing Zhao & Stephan Vogel. *A generalized alignment-free phrase extraction.* In Proc. of ACL Workshop on Building and Using Parallel Texts, pages 141–144, 1995.

[Zimmermann 02] M. Zimmermann & H. Bunke. *Hidden Markov model length optimization for handwriting recognition systems.* In Proc. of IWFHR, pages 369–374, 2002.

Índex de figures

Índex de taules