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

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Aportacions a la millora d'un sistema interactiu d'ajuda a la traducció basat en mètodes estadístics.

Projecte Final de Carrera - Enginyeria Informàtica

Joan Albert Silvestre Cerdà

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

17 d'abril de 2012

A la memòria del meu pare, Luis.

Agraïments

No és molt habitual escriure aquestes línies en un projecte de final de carrera, i realment no entenc el per què. Aquest treball representa la culminació dels estudis universitaris, una etapa molt important de la nostra vida, clarament marcada per la influència de les persones del nostre entorn: familiars, amics, enemics, coneguts, professors, etc. Totes aquestes persones són el nostre referent social i la principal font de coneixements de la que es nodrim en el dia a dia (la universitat de la vida), sent la seva interacció amb nosaltres un factor determinant en la nostra educació, especialment en la universitària, on els estudiants comencem a mostrar-nos al món com persones formades, madures i independents. Per aquest motiu trobe que tota la gent que ha influït de forma positiva en la meva persona durant aquest període de temps es mereix, per la meva part, una menció especial, una espècie de reconeixement personal. I són precisament aquestes línies les més idònies per expressar el més sincer i profund agraïment a totes aquelles persones que m'han acompanyat en aquest recorregut.

He de confessar que aquestes són les línies més difícils d'escriure de tot el document, doncs les paraules que segueixen no són més que meres aproximacions a l'afecte i estima que tinc a les persones que, sense lloc a dubte, formen part d'aquest projecte de final de carrera, i més encara, de la meva formació com a enginyer informàtic i com a persona.

En primer lloc, i com a principal impulsor d'aquest treball, vull donar les gràcies a Alfons Juan per haver-me brindat l'oportunitat de realitzar aquest apassionant projecte, preocupar-se pel meu futur immediat com a enginyer informàtic i acollir-me al seu grup. Un tren així no passa tots els dies per la vida d'una persona, i, francament, em sent molt afortunat de no haver-lo deixat escapar. En la mateixa línia, vull fer especial menció als directors d'aquest projecte, Jorge Civera i Jesús Andrés, per la gran implicació que han mostrat en tot moment durant l'elaboració d'aquest treball, per la comunicació constant i fluïda que hem mantingut al llarg d'aquests quasi cinc mesos, pels seus grans coneixements i passió per la traducció automàtica i el reconeixement de formes, pel seu altruisme, per preocupar-se pel meu futur, per la seva enorme simpatia, i sobretot, per la paciència infinita que han tingut amb mi. He de destacar, a més, el gran llegat que han deixat en mi en forma de nous coneixements i habilitats, que no són pocs. Moltes gràcies a tots dos.

Per proximitat també voldria donar les gràcies al meus companys de laboratori del DSIC: Adrià, Ihab, Miguel, Nico, Isaías i Ricardo (la comunitat del clúster), el quals m'han acollit molt amablement i m'han ajudat en moltes coses durant la realització d'aquest treball, especialment Miguel, per prestar-me la plantilla Latex d'aquest pro-

jecte i pels seus consells, i Adrià per les seves aportacions de Latex, C++, i els seus xiclets de després de dinar. Tots ells m'han fet passar una estança al laboratori molt agradable. No cal oblidar tampoc als meus companys de carrera, tant de l'Enginyeria Tècnica com de l'Enginyeria Superior, on he trobat amics i amigues que, encara que les nostres vides ara segueixen camins diferents, almenys tenim el grat record d'haver-nos creuat i compartir uns bons ratets junts. De tots ells vull fer especial menció de Jordi, Kike, Miguel i Berni, els meus primers companys de classe amb els que he passat molt bons moments; Cèsar, pels dos magnífics anys que hem compartit al segon cicle d'informàtica; a Raül (Rulo), amb qui mantinc una estreta amistat a pesar dels anys que fa que des que es deixà la carrera; i sobretot, a un gran amic meu de tota la vida i company de pis durant aquests 5 anys: Àngel, una bellíssima persona a la qual podria fins i tot atribuir-li l'obtenció del meu títol d'enginyer tècnic (ell sap que vull dir amb això).

No m'agradaria deixar en el tinter a la gent que he conegut en la meva vessant musical com a trompista, la major de les meves passions que, malauradament, per motius de salut, entre d'altres, no m'he pogut dedicar de forma més intensa o inclús de forma professional com m'hagués agradat. Recorde amb nostàlgia a professors i companys de classe de l'escola de música de Bocairent, del conservatori d'Ontinyent i del conservatori de València. Vull fer menció especial, d'una banda, als companys que fundarem el Grup Instrumental Simfonies, un gran grup d'amics apassionats de la música simfònica, així com a Àngela i Cèsar, dos companys del conservatori d'Ontinyent amb els que encara mantinc una bona amistat; i d'altra banda, als professors que més m'han marcat en la meva formació musical: Rosell, gran trompista i millor persona, però sobretot, Titín, un trompista excepcional i un professor excel·lent que em va contagiar la seva passió per la Música i que va aconseguir produir la meva millor versió com a trompista, al fer-me recuperar la confiança amb mi mateix, fins al punt que l'únic que desitjava era menjar-me l'escenari. No vull oblidar tampoc a Luís Serrano, un gran docent i magnífic compositor de música simfònica del que vaig aprendre molt, i que em va donar la possibilitat d'explorar nous horitzons musicals. També és molt important per a mi el meu ex-alumne i amic Asensio, a qui vaig poder transferir gran part dels meus coneixements musicals ensenyant-li des de zero solfeig musical i Trompa durant 3 anys. I com no, a l'Associació Unió Musical Bocairent, la banda dels meus amors i bressol de la meva educació musical, en la que he viscut vivències inoblidables i on trobe bons amics i coneguts, entre ells amics de tota la vida. Estic molt trist per no poder participar activament en ella, però promet que prompte o tard tornaré. El meu cos ho demana a crits.

També vull aprofitar per mencionar, d'una banda, els meus companys de la delegació d'àrbitres d'Alcoi del Comité Tècnic d'Àrbitres de la FFCV, estament arbitral al que pertany des de fa 4 anys, i amb els que he passat molts bons moments en la pràctica d'una feina molt gratificant i útil per al desenvolupament de la meva persona, encara que de vegades perillosa, tot s'ha de dir. D'altra banda, també vull recordar als membres de la que fou l'Associació d'Informàtica Bocairent.net ja fa bastants anys, als quals dec, en certa manera, la meva afició al món de la informàtica, i

d'entre els quals destaque a Manel Espinós i a Pere Crespo, dos grandíssimes persones amb les que encara mantinc una bona relació, i amb les que en el seu dia vaig passar molts bons ratets junts, gestionant i muntant les successives bekiparty's, entre d'altres activitats. Vull mencionar especialment a Pere, qui ha marcat molt la meva vida com a informàtic, doncs la seva persona ha estat el meu referent com a futur enginyer.

Els meus amics i amigues han estat la base fonamental no sols d'aquesta titulació que estic apunt d'assolir, sinó també de la meva vida: amb ells he viscut experiències genials que m'han servit per desconnectar dels estudis, i sobretot, per gaudir de la vida. N'estic molt agraït pel recolzament que he trobat en ells i per tot allò que han aguantat, sobretot en aquests 4 mesos en els que he realitzat el projecte: crec que tots ells tenen més ganes d'acabar-lo que jo! Podria començar a parlar, un per un, de tots ells, però en son tants i tant bons que podria estendre'm pàgines i pàgines parlant meravelles d'ells, així que espere que em donen la llicència d'evitar aquest despropòsit, que ja trobe que me n'estic excedint bastant en quant a extensió. Tots ells saben quant me'ls estime, així que estic tranquil. Sobren les paraules. Gràcies per estar ahí, genteta.

He de destacar, a més, aquelles persones que més han estat en contacte amb mi durant aquests 5 anys, que són els meus companys de pis a València. Tots ells, Adrià, Àngel, Rubén, i Víctor, companys i amics de tota la vida, són les persones amb les que més vivències, de lluny, he compartit, molt divertides i gratificants, per cert. Sobretot és de lloar com de bé m'han tractat i tot el que han tingut que suportar. Que em feren el sopar els dies que tornava tardíssim a casa perquè em trobava a classes del conservatori, o el gran esforç que realitzaven aguantant les meves dues hores d'estudi diàries de Trompa a casa, són dues mostres de com m'han "mimat". Inclús en circumstàncies extremes, com per exemple aquests els últims 5 mesos, quan em passava hores i hores al laboratori elaborant aquest treball: sempre que tornava a casa em trobava amb el dinar i el sopar fet (recentment amb un nou companyer de pis, Raül). Això no té preu. Sols l'amistat de debò pot fer eixes coses. A tots ells l'únic que puc dir-los és que gràcies, moltes gràcies per ser tant bons companys i amics, de tot cor, aquest projecte també és vostre.

També tinc paraules de gratitud cap a Teresa Llavador i als seus dos fills, Pepe i Paqui Tomàs, amics del meu pare i meus per afortunada herència, per acollir-nos tant bé al seu pis de València durant els últims 3 anys, però sobretot per ser tant bones persones, i per la gran amistat i cordialitat que mantinc amb ells com amb la resta de la seva família.

Per damunt de tot està el paper fonamental que ha jugat la meva família en la meva educació. Vull agrair el recolzament incondicional que he rebut per part dels meus familiars, tant en aquests 5 anys com en tota la meva vida. Especialment, he de donar les gràcies al meu germà Luís, per la seva inestimable companyia, el seu afecte, per iniciar-me al món de la informàtica i de la música, i sobretot, pels seus consells i per la seva saviesa com a germà major que tant aprecie com a germà xicotet; i com no, a la meva mare, Mari Carmen, per dur-me al món, cuidar-me, estimar-me i educar-me

de la forma apropiada per fer-me arribar fins on estic ara: ella és la principal culpable de la redacció d'aquestes línies, no s'equivoquem.

Estaré eternament agraït amb la meva núvia Lídia, per tot el que ha fet per mi i la paciència que ha tingut en tot el temps que estem junts, tot a pesar de la distància, que malauradament se'ns fa eterna. Per estar sempre al meu costat, per cuidar-me i mimar-me, per ajudar-me en els moments més difícils, per fer-me gaudir de tants bons moments... per moltes raons, però sobretot, pel seu amor incondicional que he compartit i que espere compartir amb ella fins que s'apague la llum del sol, i de la lluna.

Per últim, m'agradaria recordar amb aquestes paraules la persona que més mereix la meva admiració: el meu pare Luís, un gran home, persona patidora i lluitadora, que es va desviure pel benestar de la seva família i per poder assegurar un futur millor als seus fills, cosa que sense dubte ha aconseguit. De segur que estaria molt orgullós de veure com els seus fills han respost al seu sacrifici, als valors que ens ha inculcat, al seu esperit lluitador. Malauradament, la seva vida es va esvair de forma prematura, però no així el seu record. És per això que aquest treball vull dedicar-lo íntegrament a la memòria del meu pare, sobretot perquè m'agradaria que sabera, allà on estiga, que l'objectiu que va perseguir fins l'últim dia de la seva vida s'ha assolit. I que per això, i per moltes més raons, n'estic molt agraït i orgullós d'ell.

Gràcies a tots i totes.

Joan Albert Silvestre Cerdà
València, 17 d'abril de 2012

Índex

Capítol 1

Introducció

A l'espai europeu actual, caracteritzat per la convivència de diverses comunitats que escriuen i parlen una gran varietat de llengües, la traducció automatitzada ha adquirit una gran importància, doncs la cooficialitat de diverses llengües implica, entre d'altres coses, la generació de diferents versions d'informació pública. No obstant, aquest problema no és endèmic al continent europeu, sinó que es tracta d'un problema a nivell global. En una societat globalitzada com en la que vivim, la traducció automatitzada cobra especial importància, doncs cada dia es generen al món quantitats ingents d'informació que necessiten ser traduïdes en el menor temps possible a desenes o inclús centenars d'altres idiomes. Ara bé, hom podria aventurar-se a pensar que el constant i progressiu procés de globalització de la nostra societat, cada cop més accentuat a causa de la integració en les nostres vides de les Tecnologies de la Informació i de la Comunicació, podria acabar algun dia amb la multiculturalitat, i per extensió, amb el plurilingüisme. Francament, creure amb eixa possibilitat seria un greu error. Cap procés de globalització és capaç de fulminar els trets més característics d'una societat: folklore, forma de pensar i d'ésser, i menys encara, llenguatge, una de les senyes d'indentitat més fortes d'un poble. La societat globalitzada ha estat, és i serà plurilingüe, per més que passen els anys. Però aquesta demana entendre's, per poder salvar les diferències culturals existents. I és en aquest context on la traducció de la informació juga un paper molt important.

El problema que presenta la Societat de la Informació no és solament el plurilingüisme en sí, que no és ni molt menys menyspreable, doncs s'estima que existeixen aproximadament uns 6000 idiomes diferents (tot i que quasi la totalitat d'aquests llenguatges són minoritaris i la major part de la població mundial té com a llengua materna un dels 25 idiomes més parlats). El principal problema al que s'enfronta la globalització de la informació és la immensa quantitat de dades i texts que es generen diàriament a tot el món i que demanen ser traduïts a múltiples idiomes en el menor temps possible. Un traductor manual pot arribar a transcriure un màxim de 20 pàgines diàries, però és que al dia es produeixen aproximadament 20 milions de paraules d'informació tècnica! [Mur66] És impossible assimilar tal cabdal d'informació sense introduir un retard en el procés de traducció, i menys encara tenint en compte

el constant increment de la demanda de traduccions. Per tant, ens trobem enfront d'un gran coll de botella: el fluix d'entrada d'informació a traduir és molt superior a les capacitats dels traductors professionals, amb la qual cosa l'automatització de la traducció mitjançant computadors és converteix quasi indispensable.

Malauradament, i com veurem més endavant, la traducció completament automatitzada no és factible (encara que amb reserves), doncs la qualitat de les traduccions que ofereixen els sistemes actuals és força qüestionable. Aquestes deficiències han propiciat l'aparició de sistemes interactius d'ajuda a la traducció, que permeten a l'usuari corregir *on-line*[1] les traduccions que se li presenten errònies o de mala qualitat. Així doncs, aquests sistemes són de gran utilitat per a traductors experts i públic en general, ja que faciliten de forma notable el procés de traducció, així com augmenten considerablement la productivitat dels traductors professionals. El funcionament dels sistemes interactius és ben simple: el text origen és processat frase per frase, proporcionant-se propostes de traducció a l'usuari, el qual efectua les correccions necessàries (si s'escau) en la proposta inicial proveïda pel sistema, desant-se la traducció corregida. Les traduccions proporcionades es generen gràcies als avanços de la disciplina de la traducció automàtica[2] (TA), que s'ocupa del complex problema de la traducció de texts a través de computador i de forma automàtica. Cal remarcar que en l'actualitat els sistemes interactius d'ajuda a la traducció estan sent un dels principals objectes de recerca.

El propòsit d'aquest treball és, doncs, millorar les prestacions d'aquests sistemes. S'han plantejat una sèrie de millores que afecten directament al sistema de traducció automàtica subjacent, ja que és aquest el que s'encarrega de generar les propostes de traducció, i és on resideix el punt clau de millora: a major qualitat de la traducció, major qualitat del sistema. Per tant, aquest Projecte de Final de Carrera es centrarà fonamentalment en la disciplina de la traducció automàtica.

En aquest capítol introductori tractarem de proporcionar una visió general de la disciplina de la traducció automàtica, repassant breument la seva història per tal de conèixer els precedents que han donat lloc a l'estat d'art actual, així com les diferents aproximacions[3] que han sorgit al llarg de la història recent de la disciplina. Per descomptat, no deixarem de costat els sistemes interactius d'ajuda a la traducció: explicarem de forma detallada els tipus de sistemes existents i el seu funcionament. Per últim, centrarem la nostra atenció en l'aproximació estadística a la TA, i més concretament, en l'aproximació estadística basada en en segments o seqüències de paraules, sobre la qual s'ha realitzat aquest treball.

[1] Durant el procés.
[2] En anglès, *Machine Translation* (MT).
[3] Entenent per aproximació la forma d'enfocar i resoldre el problema de la traducció automàtica.

1.1 Visió general de la traducció automàtica

1.1.1 Un breu repàs a la història

La traducció automàtica no és ni molt menys un concepte recent. Les primeres idees sorgiren aproximadament al segle XVII de la ment de matemàtics i lingüistes, el quals ja s'adonaren de la necessitat d'automatitzar el procés de traducció en la parla i en l'escriptura. Òbviament aquestes idees no es van poder posar en pràctica fins que no aparegueren els primers computadors, cap a l'any 1940. De fet, les primeres traduccions automatitzades es van realitzar durant la 2ª Guerra Mundial (1939-1945), quan s'empraren els computadors per tractar de desencriptar codis secrets d'exèrcits enemics. No es tractava de traducció entre idiomes, però això va asseure un precedent de posteriors estudis.

Els primers estudis seriosos del que entenem per traducció automàtica els trobem a una sèrie de discussions entre Warren Weaver i A. Donald Booth a l'any 1946 [BF81]. Ambdós, familiaritzats en tasques de descodificació mitjançant computadors, van estimar que els mètodes emprats en aquestes tasques podien ser aplicables a la traducció entre idiomes, i a més van preveure quins serien els principals problemes als quals s'enfrontaria (i encara s'enfronta) la TA: confeccionar un diccionari complet i específic per a cada parell d'idiomes, considerar la semàntica diferent de les paraules depenent del context en que es troben, l'alteració de l'ordre de les categories gramaticals en els diferents llenguatges, impossibilitat de traduir frases fetes i expressions paraula per paraula, etc. En aquella època el problema més crític era la confecció d'un vocabulari, amb la qual cosa semblava que una part important del procés de la traducció automàtica podia ser factible: la traducció paraula a paraula.

D'aquestes discussions sorgeix la primera aproximació a la TA, en un memoràndum anomenat *Translation* presentat per Warren Weaver [Wea55], juntament amb altres investigadors d'IBM a l'any 1949. D'aquest text s'extrauen tres idees molt interessants: en primer lloc el concepte de finestra, que limitava el procés de traducció a un conjunt limitat de paraules del text; en segon lloc, que el text origen i la seva traducció comparteixen la mateixa semàntica; i en darrer lloc, el concepte d'un llenguatge universal intermedi anomenat *Interlingua*, de forma que la traducció del llenguatge origen al llenguatge destí passa per un llenguatge intermedi que suposadament tots els éssers humans comparteixen. Com a curiositat, la segon idea adés mencionada es desprén d'una frase famosa escrita pel mateix Weaver: *Quan llig un article en Rus, em dic a mi mateix: 'Açò es troba realment escrit en anglès, però ha estat codificat amb símbols estranys. Procediré a descodificar-lo'.*

A partir de la publicació del treball de Weaver, la TA va rebre un fort impuls als Estats Units. A l'any 1954 la universitat de Georgtown va presentar en societat el primer sistema de TA entre el rus i l'anglès, però amb un domini molt restringit, ja que disposava d'un vocabulari de tot just 250 paraules i tant sols era capaç de traduir 49 frases. A pesar de les pobres prestacions oferides per aquest sistema, el fet que en aquella època estigués en ple apogeu la guerra freda i per tant resultara interessant el poder realitzar traduccions ràpides del rus a l'anglès (i viceversa) va propiciar que el govern dels EUA invertira fortament en aquest camp d'investigació.

Aquestes inversions es perllongaren al llarg de la dècada dels 50, però als anys 60 va començar una època de decadència de la TA, motivada per la cada vegada més evident complexitat que comportava aquesta disciplina, trencant d'aquesta forma totes les previsions altament optimistes. Foren els informes de Bar-Hillel a l'any 1960 [BH60], i sobretot l'informe ALPAC a l'any 1966 [PC66], els que truncaren totes les esperances dipositades en aquest camp. Es va arribar a la conclusió, entre d'altres coses, de que crear un sistema de TA de qualitat era extremadament ambiciós, i que la TA era lenta, de molt baixa qualitat i dues vegades més costosa que la traducció manual. Com a conseqüència, durant més d'una dècada es va paralitzar la investigació d'aquest camp als EUA.

Aleshores, els anys immediatament posteriors van estar caracteritzats per un estancament en la investigació de la TA. A la dècada dels 70 la major part dels treballs es realitzaren a Europa i Canadà, degut al plurilingüisme present en aquestes regions. Prova d'això és que a l'any 1976 un grup d'investigadors canadencs van construir un sistema de TA anomenat *Météo*, orientat a traduir parts meteorològics [Tih82], i que va oferir molt bones prestacions; al mateix temps que la Comissió de les Comunitats Europees (CE) va elaborar un sistema per traduir del francès a l'anglès anomenat *Systran* [Bil82]. En anys successius es desenvoluparen nous sistemes per traduir entre altres parells de llengües europees. Posteriorment la CE va decidir patrocinar un ambiciós projecte anomenat *Eurotra* [AdT82], un sistema de TA que donaria suport a totes les llengües de la CE. Els investigadors d'aquest projecte van arribar a la conclusió de que, per aconseguir bones prestacions, els sistemes de TA haurien de seguir una aproximació basada en regles gramaticals i/o lògiques (veure Secció 1.1.2).

Durant la dècada dels 80 els esforços es van centrar en la construcció de sistemes de TA basats en regles i en *Interlingua* (veure Secció 1.1.2), però fou a finals de la dècada dels 80 i a principi de la dels 90 quan els grans avanços en els sistemes estadístics del reconeixement de la parla van motivar l'aparició d'una nova aproximació a la TA, la traducció automàtica estadística[4] (veure Secció 1.3). Fou a l'any 1988, durant el segon congrés del TMI-MT[5] quan Peter F. Brown, investigador d'IBM, va presentar i formalitzar l'aproximació purament estadística a la TA [B$^+$90], basada en l'ús d'un corpus d'exemples de traducció per construir el sistema, definint així les bases de gran part dels conceptes que tractarem en aquest treball.

1.1.2 Aproximacions a la traducció automàtica

Entenem com aproximacions a la traducció automàtica aquells fonaments teòrics i metodològics que tracten d'abordar i resoldre el problema de la traducció automatitzada. En aquesta disciplina trobem, principalment, dues aproximacions completament contraposades: l'aproximació basada en regles i l'aproximació empírica o basada en corpus [GV03].

[4] En anglès, *Stadistical Machine Translation* (SMT).
[5] *Theoretical and Methodologies Issues in Machine Translation.*

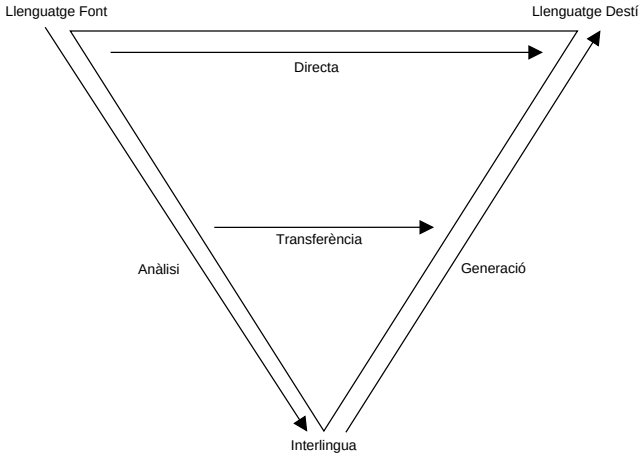


Figura 1.1: Triangle de Vauquois

Aproximació basada en regles

D'una banda, l'aproximació basada en regles, com el seu propi nom indica, es basa en definir un conjunt de regles que estableixen la manera en la que s'ha de realitzar la traducció. El punt clau d'aquesta aproximació resideix en el disseny d'aquestes regles lingüístiques, doncs requereix un coneixement profund i especialitzat dels llenguatges que participen en la traducció. En general, aquests sistemes segueixen un procés dividit en dues fases: la fase d'anàlisi, en la que es processa el text d'entrada, i la fase de generació, en la qual es realitza la traducció a l'idioma destí. Hi han aproximacions que introdueixen una tercera fase anomenada fase de transferència, a cavall entre les dues anteriors, i que tracta d'adaptar la fase d'anàlisi amb la fase de generació. A la Figura 1.1 podem observar l'anomenat *Triangle de Vauquois* [VB85], en el que s'aprecia l'esquema general de les aproximacions basades en regles: traducció directa, traducció per transferència, i *Interlingua*. Com es pot observar, conforme més baix és el nivell de la piràmide, major èmfasi en la fase d'anàlisi i menor en la fase de transferència, i viceversa. Tot seguit expliquem breument aquests tres enfocaments [GV03].

- **Traducció Directa**: Aquesta fou l'aproximació adoptada en els primers sistemes de TA, com és el sistema *Systran*. Es basa en la traducció paraula per paraula mitjançant un anàlisi morfosintàctic, en el que s'identifiquen categories gramaticals i altres característiques lingüístiques, com són el gènere, temps verbal, etc.

- **Traducció per Transferència**: En aquesta aproximació es realitza un anàlisi més detallat del text origen, amb l'objectiu d'obtindre una representació lògico-sintàctica en forma d'arbre de les frases del text d'entrada. Posteriorment, en la fase de transferència es realitza la conversió de la representació lògica

de les frases origen a una representació equivalent en el llenguatge destí, tot mitjançant l'aplicació de regles que indiquen la correspondència dels fragments de l'arbre sintàctic origen als de l'arbre sintàctic destí. Per últim, l'arbre destí es transforma en un text llegible per a l'usuari.

- **Interlingua**: Aquesta aproximació és el cas extrem de la traducció per transferència, doncs es realitza un minuciós i exhaustiu procés d'anàlisi del text origen fins aconseguir una representació conceptual completament independent dels llenguatges origen i destí anomenada *Interlingua*. Aquesta aproximació presenta el gran avantatge de que, en lloc de definir correspondències entre cada parell de possibles llenguatges a traduir, és suficient amb definir la correspondència entre cada llenguatge i l'*Interlingua*. Ara bé, la complexitat d'aquesta aproximació resideix en dissenyar un *Interlingua* adequat, que permeta tant representar conceptes de forma única com establir correspondències úniques (cada paraula de l'*Interlingua* estarà relacionada amb una i sols una paraula d'un llenguatge concret).

Aproximació basada en corpus

D'altra banda, l'aproximació basada en corpus es concep com una aproximació empírica, ja que la informació emprada per construir el sistema s'obté a partir d'un corpus d'exemples de traduccions del llenguatge font al llenguatge destí. El gran avantatge que presenten aquests sistemes és que la tecnologia que els implementa pot ser fàcilment reutilitzada per a altres dominis d'aplicació o inclús altres parells d'idiomes, però per contra requereixen recopilar grans quantitats d'informació per tal de capturar una representació significativa del domini en el que es pretén realitzar el procés de traducció. A trets generals, aquests sistemes parteixen d'un corpus d'entrenament conformat per parells de frases traduïdes per un grup d'experts, i que en la majoria del casos requereixen una fase preprocés (separar signes de puntuació de les paraules, transformar majúscules en minúscules, etc.) per tal de facilitar i garantir l'adquisició de fonts de coneixement robustes i de qualitat durant la fase d'entrenament. Amb el sistema entrenat es pot dur a terme la tasca de traducció d'una frase d'entrada, construint de la forma més òptima possible la frase en l'idioma destí, en base a tota la informació que el sistema té al seu abast. Si en la fase d'entrenament s'ha realitzat un preprocés de les dades, aleshores és imprescindible que la frase a traduir siga sotmesa al mateix preprocés, i posteriorment, a una fase de postprocés, en la que es desfan els canvis realitzats al preprocés. Dins d'aquesta categoria existeixen dues grans aproximacions: els sistemes basats en exemples[6] i els sistemes estadístics[7] [GV03].

- **Sistemes Basats en Exemples**: L'idea principal d'aquesta aproximació és que durant el procés de traducció la frase d'entrada es compara en una base de dades construïda a partir del corpus d'entrenament, generant-se hipòtesis i combinant-les de forma apropiada fins obtindre la traducció corresponent. Per

[6] EBMT, *Example-Based Machine Translation.*
[7] SMT, *Stadistical Machine Translation.*

ser un poc més concrets, en primer lloc es descomposa la frase d'entrada en fragments, donant lloc a un gran nombre de possibles hipòtesis o camins de traducció (ja que la descomposició es pot realitzar de moltes maneres diferents); en segon lloc, els fragments es comparen amb la base de dades d'exemples i s'identifiquen les seves traduccions corresponents; i en tercer lloc, es combinen de forma adequada els fragments traduïts, originant la frase final.

- **Sistemes Estadístics**: L'aproximació estadística a la TA és molt similar a la seguida pels sistemes basats en exemples, però es desmarca clarament en el procés de comparació-combinació emprat per aquests, ja que en el seu lloc s'empren tècniques estadístiques tant en l'entrenament del sistema com en el procés de traducció.

Aquest treball es centra en el marc estadístic de la TA, amb la qual cosa en la Secció 1.2 s'explicarà en major detall els aspectes més importants que concerneixen a la traducció automàtica estadística.

1.1.3 Sistemes d'ajuda a la traducció

Com ja sabem, el principal propòsit de la traducció automàtica és dissenyar sistemes capaços de traduir texts sense la participació humana, és a dir, de forma completament automàtica. No obstant, la tecnologia actual no permet assolir tal propòsit [Kay97], amb la qual cosa es requereix una post-edició manual de les traduccions proveïdes per aquests sistemes. Aquests dos processos aïllats impedeixen, d'una banda, que el sistema aprenga del coneixement del traductor humà, i d'altra banda, que el traductor humà obtinga beneficis de les capacitats adaptatives i de l'eficiència que presenta d'un sistema de TA.

Una forma de resoldre aquesta manca de *feedback* és, precisament, donar lloc a un espai cooperatiu on màquina i humà participen interactivament en el procés de traducció [IC97], en entorns coneguts com sistemes d'ajuda a la traducció[8]. Històricament, els sistemes d'ajuda a la traducció i els sistemes de traducció automàtica han estat considerats sistemes diferents, encara que tecnològicament semblants, doncs podem considerar que els sistemes d'ajuda a la traducció són un superconjunt dels sistemes de traducció automàtica.

Històricament, la interactivitat entre sistemes de traducció i humans s'ha materialitzat en diferents enfocaments. Un dels més exitosos ha estat el basat en memòries de traducció, en el que l'usuari fa ús d'un sistema que compta amb una base de dades immensa d'exemples de traducció. Un altre enfocament fou requerir la intervenció dels humans per tal de resoldre ambigüitats de caire lèxic, sintàctic o semàntic del text d'entrada, o inclús mantindre i actualitzar diccionaris d'usuari o buscar a través d'ells [Slo85, WWC+86], amb més o menys èxit. Posteriorment, el projecte *TransType* [LFL00, LLL02, FLL02, Fos02] va aportar un nou enfocament que consistia en centrar la interactivitat directament al text traduït, integrant sistemes autònoms de TA en l'entorn interactiu. Aquest darrer enfocament va donar origen al que coneixem per

[8] En anglès, *Computer-Assisted Translation Systems*.

sistemes de traducció interactius-predictius, el quals estan rebent en l'actualitat un ampli suport de la comunitat investigadora [Civ08, BBC+09].

Tot seguit detallem el funcionament de les dues aproximacions més importants de la traducció assistida: els sistemes de memòria de traducció, i els sistemes interactius-predictius.

Sistemes de memòria de traducció

Aquesta aproximació es basa en la recopilació d'exemples de traducció en una (presumiblement) extensa base de dades [UoG95]. Durant el procés de traducció, la frase origen és segmentada, de forma que per a cada segment es realitza una cerca de l'exemple de traducció més semblant emmagatzemat a la base de dades. Si aquesta conté un segment en l'idioma origen que coincideix exactament en el segment buscat, aleshores la traducció corresponent és mostrada per pantalla, amb un grau de similitud del 100%. Per contra, si no s'ha trobat un segment idèntic, aleshores es proporciona la traducció més similar existent a la base de dades. Siga com siga, el sistema presenta a l'usuari la proposta de traducció, de forma que aquest pot, bé acceptar la proposta, o bé modificar-la / corregir-la. En aquest darrer cas, la traducció modificada és emmagatzemada a la base de dades. Pot haver-hi casos en els que no es proporcione cap proposta de traducció (p.e. no existeix a la base de dades cap segment amb un grau de similitud superior a un llindar mínim preestablert), amb la qual cosa l'usuari ha de realitzar manualment la traducció que posteriorment és emmagatzemada a la base de dades. Notar la gran similitud tecnològica existent entre aquest tipus de sistemes i els sistemes de TA basats en exemples de traducció (Capítol 1.1.2): podem considerar aquests com una extensió dels sistemes de TA originals.

Amb tot, aquests sistemes milloren les seves prestacions amb el pas del temps, ja que la realimentació existent entre el sistema i l'usuari permet incrementar el nombre d'exemples existents a la base de dades, a la vegada que el sistema s'especialitza si es treballa amb dominis molt concrets. És per això que aquests sistemes funcionen especialment bé amb texts que presenten moltes repeticions, o en traduccions incrementals realitzades sobre documents prèviament traduïts.

Cal remarcar que en aquests sistemes el motor de cerca és el punt crític, ja que per trobar coincidències s'ha d'explorar grans quantitats d'informació, així com identificar patrons i similituds en el cas en que no es troben coincidències. Per tant, la rapidesa del motor de cerca defineix en gran mesura les prestacions del sistema, donat que la latència dels resultats de la cerca afecta directament a l'usuari que es troba interactuant amb el sistema. D'altra banda, cal notar que la qualitat de les traduccions depèn de la grandària de la base de dades, però al mateix temps una grandària excessiva de la base de dades pot induir a latències de cerca elevades (degut a la major complexitat del procés de cerca).

Existeixen diferent sistemes comercials de memòria de traducció com són *Trados* (SLD Trados Studio) [HK], *Dejavú* [Dej] o *Google Translate Toolkit* [Goo].

Sistemes interactius-predictius

Aquesta aproximació pretén fusionar els paradigmes de la traducció assistida i la traducció automàtica. Davant d'un text d'entrada, un sistema complet de TA produeix hipòtesis de traducció de frases completes o de porcions d'aquestes, que poden ser acceptades o corregides per l'usuari. Però la característica més interessant d'aquests sistemes és que són capaços de proporcionar prediccions de traducció [Civ08]: davant d'una traducció parcial introduïda per l'usuari d'una frase origen (prefix), el sistema prediu les traduccions més probables que completen la frase (sufixos), de les quals l'usuari podrà acceptar o modificar una d'elles. Siga com siga, cada segment corregit és processat pel sistema de TA subjacent per tal de millorar la qualitat tant de les traduccions com de les prediccions. Així doncs, aquests sistemes són requerits per generar prediccions adequades i de forma eficient.

1.2 Traducció automàtica estadística

En aquesta secció introduirem alguns conceptes bàsics per entendre de forma intuïtiva l'enfocament estadístic de la TA. D'ara endavant considerarem el problema de traduir una frase f d'un vocabulari[9] d'entrada $\mathcal{F}$ a una frase e d'un vocabulari d'eixida $\mathcal{E}$[10].

1.2.1 Conceptes bàsics de probabilitat

En primer lloc introduirem una sèrie de nocions bàsiques estadístiques que seran àmpliament emprades en aquest document, especialment el càlcul de probabilitats, explicant intuïtivament el seu significat dins del context de la TA.

El càlcul de probabilitats, com bé sabem, és una forma de quantificar la incertesa que tenim sobre les coses, com per exemple la possibilitat de si demà plourà o no, o la possibilitat de que la nostra participació en un sorteig ens convertisca en milionaris. Aquesta incertesa es pot capturar fàcilment amb l'ús de variables aleatòries: una variable aleatòria X representa el resultat d'un esdeveniment en concret i pren una sèrie de valors dependents del resultat d'eixe esdeveniment. Per exemple, si la variable aleatòria X representa l'esdeveniment "resultat del llançament d'una moneda", i considerem com a possible resultat de l'esdeveniment "cara" ($X = \textit{cara}$), aleshores la probabilitat de que el resultat de llançar una moneda siga cara és $P(X = \textit{cara}) = 1/2$ (per simplicitat escriurem $P(\textit{cara}) = 1/2$).

Tota variable aleatòria segueix una distribució de probabilitat, que és la forma en que es reparteix la totalitat de la massa de probabilitat entre tots els successos possibles. Hi ha casos en els que és factible estimar la distribució de probabilitat que segueixen els valors que pot prendre d'una variable aleatòria mitjançant un anàlisi freqüencial (recopilar dades de possibles successos i estimar la probabilitat d'ocurrència de cadascun d'ells), però hi ha d'altres en els que la mesura d'incertesa dels

[9] Anomenarem vocabulari al conjunt de totes les paraules que pertanyen un llenguatge.

[10] Per conveni, s'utilitza f (*french, foreign*) i e (*english*) per denotar les frases d'entrada i d'eixida respectivament.

successos es pot ajustar a alguna distribució de probabilitat ja coneguda que es dona lloc en altres escenaris[11].

Per exemple, el llançament d'una moneda segueix una distribució de probabilitat uniforme, en la que tots els possibles esdeveniments són equiprobables: $P(cara) = P(creu) = 1/2$. Imaginem que per uns moments no se'n refiem i decidim llançar, per exemple, un milió de vegades una moneda a l'aire, a veure que passa. Comprovaríem com aproximadament la meitat de vegades apareixeria cara, i l'altra meitat apareixeria creu: hauríem estimat la distribució de probabilitat mitjançant un anàlisi freqüencial, que no és més que comptar freqüències d'aparició d'un esdeveniment i normalitzar amb total d'esdeveniments. Aquest tipus d'estimació, en la qual hem tractat d'ajustar la nostra distribució de probabilitat el màxim possible a les dades que hem recaptat, s'anomena estimació per màxima versemblança.

Per exemple, suposem que al llançar un milió de vegades una moneda a l'aire hem obtingut cara 500.427 vegades:

$$P(\text{cara}) = \frac{N(\text{cara})}{N(\text{llançaments})} = \frac{500427}{1000000} = 0.500427 \approx 1/2 \tag{1.1}$$

Com podem comprovar, no és necessari invertir tant de temps recopilant dades sobre aquesta variable aleatòria: sabem que el resultat del llançament d'una moneda ve modelat per una distribució uniforme.

En definitiva, podem veure una distribució de probabilitat com una funció que assigna, a cada succés definit sobre una variable aleatòria, la probabilitat de que aquest succés s'esdevinga. Tota distribució de probabilitat compleix dues propietats: en primer lloc, la probabilitat d'un succés en concret pren valors entre 0 i 1 (veure Equació (1.2)), i en segon lloc, la suma de les probabilitats de tots els possibles successos deu sumar 1 (veure Equació (1.3)).

$$\forall x : 0 \leq p(x) \leq 1 \tag{1.2}$$

$$\sum_x p(x) = 1 \tag{1.3}$$

Amb aquestes nocions bàsiques, es centrarem ara en el context de la TA. En aquest document tractarem principalment amb dues variables aleatòries: F (frase d'entrada) i E (frase d'eixida), amb valors e i f, respectivament. Tot seguit detallem les expressions de probabilitat que emprarem a aquest document:

$p(e)$ És la probabilitat *a priori*. Representa la possibilitat de que e s'esdevinga. Per exemple, si e representa a la frase en anglès "I like bees", aleshores $p(e)$ expressa la probabilitat de que una persona concreta pronuncie en un moment donat la frase "I like bees". Notar que també ens trobarem $p(f)$, amb un significat anàleg.

[11] Hi existeixen nombroses distribucions de probabilitat com són la uniforme, la normal (o gaussiana), la binomial o la multinomial, entre d'altres.

$p(f \mid e)$ És la probabilitat condicional. Representa la possibilitat que s'esdevinga f donat que e s'ha esdevingut. Per exemple, si e representa la frase en anglès "I like bees", i f representa la frase en català "Demà plourà", aleshores $p(f \mid e)$ expressa la probabilitat de que un traductor expert, després de llegir la frase e, la traduïsca en f (no sembla molt probable). Ara bé, si en canvi considerem que f representa la frase en català "M'agraden les abelles", aleshores la cosa canvia. Cal notar que també ens trobarem $p(e \mid f)$, amb un significat anàleg.

$p(e, f)$ És la probabilitat conjunta. Representa la possibilitat de que s'esdevinguen simultàniament e i f. Si e i f són independents (no existeix cap influencia entre una i altra), aleshores $p(e, f) = p(e)\,p(f)$. Per exemple, si e representa l'event "Llançar una moneda i treure cara", i e "Llançar una moneda i treure creu", òbviament aquests esdeveniments no interfereixen l'un en l'altre, amb la qual cosa la possibilitat de que a l'efectuar dos llançaments de moneda, en primer lloc s'obtinga cara i després creu és $p(e, f) = p(e)\,p(f) = 1/2 \times 1/2 = 1/4$. D'altra banda, si e i f estan relacionades, aleshores s'aplica l'anomenada regla de la cadena, de forma que $p(e, f) = p(e)\,p(f \mid e)$ (o de forma equivalent, $p(e, f) = p(f)\,p(e \mid f)$). Això significa que la probabilitat de que e i f s'esdevinguen simultàniament ve donada per la probabilitat de que "e s'esdevinga" multiplicat per la probabilitat de que "si e s'esdevé, aleshores f també s'esdevé". En el cas de la TA, si e i f són traduccions mútues, aleshores existeix una clara relació d'influència entre elles. Cal notar que $p(e, f)$ i $p(f, e)$ representen el mateix concepte.

Per acabar, introduïm la regla de Bayes, que és el pilar estadístic sobre el qual es fonamenta la TAE:

$$p(e \mid f) = \frac{p(f, e)}{p(f)} = \frac{p(e)\,p(f \mid e)}{p(f)} \tag{1.4}$$

La regla de Bayes ens permetrà reformular la forma de calcular com de bona o dolenta és una frase e com a traducció d'una frase f. Més endavant raonarem sobre açò.

1.2.2 Traducció estadística

Podem definir més formalment el problema al que s'afronta la TAE de la següent forma: donada una frase f d'un vocabulari origen $\mathcal{F}$, desitgem trobar aquella frase e d'un vocabulari destí $\mathcal{E}$ que maximitza $p(e|f)$, és a dir, busquem la traducció més probable $\hat{e}$:

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, p(e \mid f) \tag{1.5}$$

La funció argmax la podem llegir de la següent forma: de totes les possibles frases (traduccions) e, busquem aquella que maximitze el valor de $p(e \mid f)$, és a dir, la traducció més probable $\hat{e}$. Ara bé, com obtenim el valor de $p(e|f)$ per a cada possible

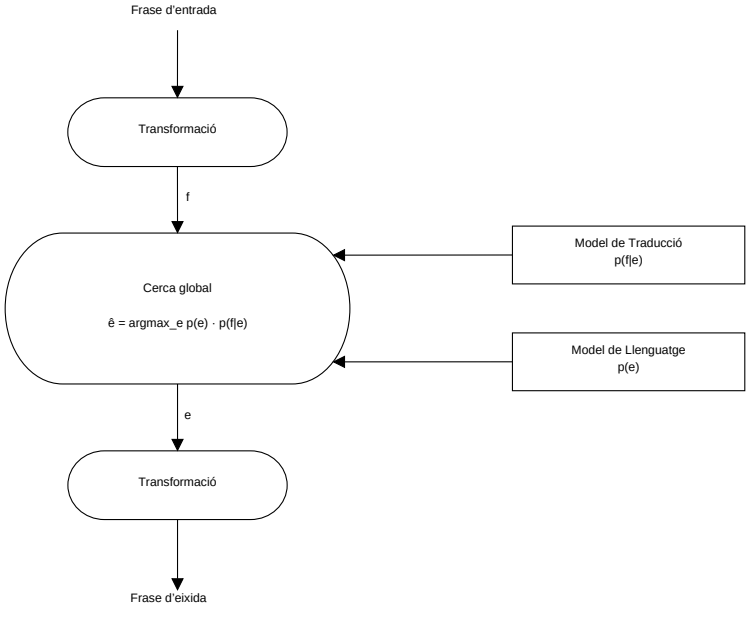


Figura 1.2: Arquitectura general del procés de traducció basat en el raonament sobre la regla de Bayes.

cadena e? La regla de Bayes ens permet raonar sobre el càlcul d'aquesta distribució de probabilitat:

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, p(e \mid f) = \underset{e}{\operatorname{argmax}}\, \frac{p(e)\, p(f \mid e)}{p(f)} = \underset{e}{\operatorname{argmax}}\, p(e)\, p(f \mid e) \tag{1.6}$$

Cal notar que, com busquem la traducció més probable, podem estalviar-nos modelar la distribució de probabilitat $p(f)$, ja que no depèn d'e i per tant no afecta a la funció *argmax*.

D'aquesta manera, la traducció $\hat{e}$ més probable és aquella que maximitza el producte de dos termes: d'una banda, la probabilitat que s'esdevinga la frase e (p.e. la possibilitat de que algú pronuncie en un moment donat la frase e), i d'altra banda, la probabilitat de generar f havent observat e (p.e. la possibilitat de que algú traduïsca e en f). Aquesta aproximació es coneix com el model de la font i del canal [B$^+$90], el qual defineix l'equació fonamental de la traducció automàtica estadística [B$^+$93].

Com veurem més endavant, $p(e)$ es calcula mitjançant models de llenguatge (veure Secció 1.2.3), i $p(f \mid e)$ mitjançant models de traducció (veure Secció 1.2.4). A la Figura 1.2 podem observar l'arquitectura general de la traducció automàtica estadística basada en la inferència de Bayes [GV03].

Aplicació de la regla de Bayes

Si analitzem un poc més el significat de l'Equació (1.6), podem comprovar com se'ns insta a generar una frase e en l'idioma destí per posteriorment convertir-la en una frase f en l'idioma origen, quan en realitat el que ens interessa és justament el contrari, convertir una frase f en una frase e en un idioma destí. Per que això, i no modelar directament $p(e \mid f)$? Tot seguit entendrem per què amb un senzill exemple [Kni99b].

Anem a raonar sobre la relació existent entre infermetats i símptomes. Si considerem que f és un conjunt de símptomes i e és una infermetat en concret, des d'un punt de vista mèdic seria molt interessant poder diagnosticar immediatament una infermetat a partir d'uns símptomes, és a dir, modelar directament $p(e \mid f)$. No obstant, és molt complicat construir directament aquest model, ja que hi ha moltes infermetats que poden provocar els mateixos símptomes. En canvi, sí que es coneix amb certesa quins símptomes són provocats per una infermetat en concret, amb la qual cosa és molt més viable construir un model $p(f \mid e)$. D'altra banda, també seria factible modelar $p(e)$, la freqüència d'aparició d'una infermetat en la població. Així doncs, per poder solucionar el problema de modelar directament $p(e \mid f)$, podem raonar sobre la probabilitat de que una infermetat e s'esdevinga ($p(e)$), així com en la probabilitat de que els símptomes f que pateix un pacient siguen provocats per eixa infermetat e en concret ($p(f \mid e)$).

Tanmateix, des d'un punt de vista lingüístic no sembla tant obvi que modelar $p(f \mid e)$ resulte menys complex que $p(e \mid f)$: en realitat, i com veurem més endavant, a partir de la mateixa informació podrem construir indistintament ambdós models. L'única raó per la qual s'aplica la regla de Bayes és que s'inclou una font d'informació addicional i independent, $p(e)$, que ens permetrà millorar la qualitat de les traduccions, i que a més resulta molt més fàcil de construir, ja que aquest model depèn únicament del llenguatge destí i per tant pot ser construït a partir d'un corpus monolingüe (com és obvi, és molt més fàcil obtindre text monolingüe que text bilingüe).

Aportacions dels models al procés de traducció

Anem a explicar de forma intuïtiva allò que aporten cadascun dels "mòduls" que hem obtingut al raonar mitjançant la regla de Bayes: el model de llenguatge $p(e)$, i el model de traducció $p(f \mid e)$ [Kni99b].

D'una banda, suposem que s'assigna un valor alt a $p(f \mid e)$ sols si les paraules de la frase f són en general traduccions de les paraules de la frase e. Baix aquesta assumpció, les paraules de f podrien aparèixer en qualsevol ordre, així que $p(f \mid e)$ no seria un bon model per traduir frases de $\mathcal{F}$ a frases de $\mathcal{E}$. D'altra banda, suposem que assignarem un valor alt a $p(e)$ si i sols si e és una frase gramaticalment correcta, cosa raonable però molt complicada en la pràctica.

Aleshores, si utilitzem ambdues fonts d'informació en el procés de convertir la frase observada f en la seva traducció més probable e, d'acord amb l'Equació (1.6) a cada possible e se li assignarà una puntuació $p(e)\,p(f \mid e)$. El model $p(f \mid e)$ ens garantirà que una hipòtesi e contindrà paraules que en general seran traduccions de les paraules de f. Per exemple, si considerem la frase $f =$ *la bruixa verda*, les frases

$e_1 =$ *the green witch* i $e_2 =$ *witch the green* són dues traduccions a les quals el model que estem considerant atorgarà una bona puntuació (molt possiblement la mateixa), però és obvi que e_2 no és una bona traducció de f. És en aquest aspecte on intervé el factor $p(e)$, ja que aquest assignarà una puntuació més baixa a aquelles frases no gramaticals, i viceversa. Aleshores, en el còmput global de $p(e \mid f)$, e_1 obtindrà una puntuació major que e_2.

Donada aquesta perspectiva, $p(e)$ efectivament es preocupa per l'ordre de les paraules de la frase d'eixida, mentre que $p(f \mid e)$ no, amb la qual cosa sembla més fàcil construir $p(f \mid e)$ que modelar directament $p(e \mid f)$. No obstant, seguir aquesta assumpció presenta certes deficiències. Imaginem que el nostre model de traducció $p(e \mid f)$ ens proporciona la següent seqüència de paraules: *vaig del i vaig blava l'altre meu a anar una amic dia Toni serp casa trobar*. Sembla bastant complex reordenar aquesta frase, fins i tot per a un ésser humà, el qual disposa d'una font de coneixement molt més completa que la d'una màquina. Per tant, arribem a la conclusió de que $p(f \mid e)$ deuria de saber almenys un poc sobre l'ordre de les paraules, i no limitar-se a proporcionar un conjunt de paraules sense més.

De la mateixa forma, resultaria molt útil si el model de llenguatge $p(e)$ aporta informació sobre quines paraules elegir en la traducció final. Per posar un exemple, la preposició en anglès *in* pot traduir-se al català de forma equivalent com *a* o *en*. Imaginem que, com a traducció a la frase *I live in Benicolet*, el model de traducció ens proporciona dues possibles hipòtesis equiprobables: *Visc en Benicolet* i *Visc a Benicolet*. La segon frase es troba en un català més correcte, amb la qual cosa el model de llenguatge li assignarà major probabilitat, i per tant s'erigirà com la traducció més probable.

Fins al moment hem presentat de forma intuïtiva com es modelitza el procés de traducció, en el qual intervenen dos "mòduls" independents: un model de llenguatge $p(e)$ i un model de traducció $p(f \mid e)$. Tot seguit descrivim més a fons aquests dos models.

1.2.3 Models de llenguatge

Com ja hem esmentat adés, el model de llenguatge és "l'especialista" de l'idioma destí, doncs és capaç de donar major importància a aquelles traduccions gramaticalment correctes, en detriment d'altres traduccions més imprecises. Ara bé, com obté aquest model tot el coneixement que té al seu abast? Els cas ideal seria conferir al model informació sobre les propietats gramaticals, sintàctiques i semàntiques del llenguatge, i inclús informació sobre allò que diuen i escriuen les persones. Ara bé, obtindre i representar tota aquesta informació és una tasca tant excessivament complexa que no és assumible. En la pràctica, resulta molt més senzill emmagatzemar frases que es poden escriure o pronunciar en el llenguatge que es modelitza, recopilades a partir d'un corpus monolingüe d'exemples de frases, estimant la distribució de probabilitat de les frases en $\mathcal{E}$ per màxima versemblança. Per exemple, si al recopilar 10000 frases en català detectem que la frase *M'agrada la cassola al forn* apareix 12 vegades, aleshores la probabilitat de que aquesta frase s'esdevinga és $p($*M'agrada la cassola al forn*$) =$

$12/10000 = 0.0012$.

Seguir aquesta aproximació presenta un problema: és impossible capturar totes les frases que es poden generar en un llenguatge (p.e. la frase *El meu besnét menja gossos verds i àguiles blaves* és completament correcta, però poc freqüent). Aleshores, cap la possibilitat que durant el procés de traducció s'assigne una $p(e)$ igual a zero a frases perfectament correctes degut a que mai s'han vist en l'entrenament del model. Per tant, no serviria de res tenir un bon model de traducció $p(f \mid e)$ si el model de llenguatge $p(e)$ assigna probabilitat zero a frases correctes, ja que en el còmput total de $p(e \mid f)$ obtindríem una probabilitat nul·la.

Arribats a aquest punt, ens donem compte que no és necessari emmagatzemar una gran base de dades de frases per jutjar si una frase és gramaticalment correcta o no (pensem en l'exemple anterior). En canvi, si trossegem la frase i els segments resultants són gramaticalment correctes, i després aquests segments els podem combinar de forma raonable, aleshores podem afirmar que la frase és bona. I és precisament aquesta aproximació la que més bones prestacions confereix als models de llenguatges i la més comunament emprada als sistemes estadístics de traducció automàtica [GV03].

Models d'n-grames

D'acord amb l'aproximació que acabem d'exposar, la frase d'eixida e es trosseja en segments o subcadenes. Una subcadena d'n paraules s'anomena n-grama, de forma que, per exemple, si $n = 2$ parlem de bigrames, si $n = 3$ parlem de trigrames, o si $n = 1$ parlem d'unigrames (o simplement paraules). Aleshores, si una frase està formada per una sèrie d'n-grames raonables, és molt probable que siga una frase gramaticalment correcta [Kni99b].

En general, i més formalment, un model d'n-grames es defineix de la següent forma [GV03]:

$$p(e) = \prod_{i=1}^{I} p(e_i \mid e_1, \ldots, e_{i-1}) \tag{1.7}$$

on I és la longitud de la frase e, e_i és la paraula i-èsima de la frase e, i $e_0 = e_{-1} = \ldots = e_{-n+1} = e_{I+1} = \$$, que és un símbol emprat per delimitar les frases.

En realitat, l'Equació (1.7) és el resultat aplicar a $p(e)$ la regla de la cadena a nivell de paraula [Koe10]:

$$p(e) = p(e_1, e_2, ..., e_I) = p(e_1)p(e_2 \mid e_1)p(e_3 \mid e_2, e_1) \ldots p(e_I \mid e_1, ..., e_{I-1}) \tag{1.8}$$

D'aquesta forma, la probabilitat del model $p(e_1, e_2, ..., e_I)$ és el producte de les probabilitats de cada paraula e_i donada la història de la mateixa[12]. Ara bé, és molt costós i

[12] Conjunt de paraules precedents a la paraula considerada.

poc pràctic emmagatzemar la història completa, motiu pel qual la història és limitada a $m = n - 1$ paraules:

$$p(e_i \mid e_1, ..., e_{i-1}) \simeq p(e_i \mid e_{i-m}, ..., e_{i-2}, e_{i-1}) \tag{1.9}$$

Històricament, la majoria dels sistemes de TA han emprat trigrames [NGW95], però en l'actualitat solen emprar models de 5-grames, atès que es treballa en corpus més grans. En el cas dels trigrames, s'empra una història de dues paraules per predir la probabilitat de la tercera. Formalment, i a partir de l'Equació (1.7):

$$p(e) = \prod_{i=1}^{I} p(e_i \mid e_{i-2}, e_{i-1}) \tag{1.10}$$

Per exemple, si considerem la frase *"visc a Guadassèquies ."*, la seva probabilitat en un model de llenguatge de trigrames es calcularia de la següent forma:

$$\begin{aligned} p(\textit{visc a Guadassèquies .}) = {} & p(\textit{visc} \mid \$) \cdot p(\textit{a} \mid \$\ \textit{visc}) \cdot \\ & \cdot p(\textit{Guadassèquies} \mid \textit{visc a}) \cdot \\ & \cdot p(. \mid \textit{a Guadassèquies}) \end{aligned}$$

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 C [GV03]: per estimar $p(e_i \mid e_{i-2}, e_{i-1})$ s'ha d'analitzar la freqüència d'aparició de la paraula e_i just després de la seqüència de paraules $e_{i-2}\ e_{i-1}$, i normalitzar aquest resultat pel nombre total d'aparicions de la història $e_{i-2}\ e_{i-1}$, com veiem a la següent equació: (per simplificar la notació, $e_i = e'$, $e_{i-1} = e''$ i $e_{i-2} = e'''$)

$$p(e' \mid e''', e'') = \frac{N(e''', e'', e')}{N(e''', e'')} \tag{1.11}$$

on $N(e''', e'', e')$ és el nombre de vegades que la seqüència de paraules $e'''\ e''\ e'$ apareix al corpus C. En termes de normalització, cal notar que:

$$N(e''', e'') = \sum_{e' \in \mathcal{E}} N(e''', e'', e') \tag{1.12}$$

de forma que es segueix una distribució de probabilitat, doncs:

$$\sum_{e' \in \mathcal{E}} p(e' \mid e''', e'') = 1 \tag{1.13}$$

A la Taula 1.1 podem observar un exemple d'estimació de probabilitats de trigrames a

Paraula	Ocurrències	Prob.
cross	123	0.547
tape	31	0.138
army	9	0.040
card	7	0.031
,	5	0.022

Taula 1.1: Exemple d'anàlisi freqüencial de trigrames i estimació de probabilitats d'ocurrència d'una paraula donada la història *the red* al corpus *Europarl* (on *the red* s'esdevé 225 vegades).

partir de dades reals [Koe10], en aquest cas el corpus del Parlament Europeu anomenat *Europarl*. Dels 225 trigrames de l'*Europarl* que comencen per *the red*, en 123 ocasions els segueix la paraula *cross*, i per tant la probabilitat d'aquest trigrama $p(\textit{cross} \mid \textit{the red}) = \frac{123}{225} \simeq 0.547$. De forma anàloga, $p(\textit{tape} \mid \textit{the red}) = \frac{31}{225} \simeq 0.138$, a l'igual que per a la resta de paraules que apareixen a la taula.

Suavitzat de models

Doncs bé, com ja s'ha parlat adés, seguir una aproximació basada en n-grames evita el problema s'assignar probabilitat zero a frases no vistes en anterioritat (és a dir, no vistes en el corpus d'entrenament). No obstant, això no és del tot cert, ja que cap la possibilitat d'assignar probabilitat zero a una frase en la qual apareix un n-grama no vist anteriorment. Per tal d'evitar això, s'apliquen tècniques de suavitzat.

Amb un exemple entendrem fàcilment com funciona el suavitzat en models d'n-grames. Desitgem calcular la probabilitat del trigrama *canta bon dia*, aparentment poc probable. Si en el nostre corpus d'entrenament la paraula *dia* mai segueix a la història *canta bon* (trigrama), aleshores podem preguntar-nos si almenys *dia* ha aparegut en alguna ocasió després de *bon* (bigrama), i en tal cas, probablement el trigrama *canta bon dia* no és del tot dolent. És més, fins i tot podriem preguntar-nos si *dia* és una paraula comuna o no (unigrama). I finalment, si cap d'aquestes casos s'esdevé, es pot assignar una xicoteta probabilitat al trigrama per evitar donar-li probabilitat zero i penalitzar tota la frase. Aleshores, per estimar la probabilitat dels trigrames, en lloc d'emprar l'Equació (1.11), és preferible utilitzar aquesta [Kni99b]:

$$p(e_3 \mid e_1, e_2) = \lambda_1 \cdot \frac{N(e_1, e_2, e_3)}{N(e_1, e_2)} + \lambda_2 \cdot \frac{N(e_2, e_3)}{N(e_2)} + \lambda_3 \cdot \frac{N(e_3)}{N} + \lambda_4 \qquad (1.14)$$

on $0 \leq \lambda_i \leq 1$, $\sum_i \lambda_i = 1$, i N és el total de paraules existents al corpus.

Aquesta tècnica de suavitzat s'anomena interpolació lineal, la qual permet definir la distribució de probabilitat d'un model a partir de combinacions lineals de les distribucions d'altres models. D'aquesta forma, la distribució de probabilitat suavitzada

d'un trigrama ve donada per la contribució ponderada d'un model de trigrames, bigrames, i unigrames, a més d'una massa de probabilitat fixa (λ_4). Els paràmetres λ_i s'empren per donar major o menor pes a la contribució de cadascun d'aquests models atenent a les característiques del corpus i/o del llenguatge, amb la qual cosa resulta convenient ajustar aquests valors de la forma més òptima possible, generalment amb un conjunt de validació.

Aleshores, si reprenem l'exemple anterior, la probabilitat suavitzada del trigrama *canta bon dia* s'estimarà de la següent forma:

$$\begin{aligned} p(dia \mid canta\ bon) = \lambda_1 \cdot \frac{N(canta\ bon\ dia)}{N(canta\ bon)} + \\ + \lambda_2 \cdot \frac{N(bon\ dia)}{N(bon)} + \lambda_3 \cdot \frac{N(dia)}{N} + \lambda_4 \end{aligned} \tag{1.15}$$

D'aquesta forma, a l'aplicar tècniques de suavitzat ens assegurem que frases, i per extensió, n-grames no vists en la fase d'entrenament, tindran almenys una xicoteta massa de probabilitat que garantirà la seva consideració com a hipòtesis de traducció en el còmput global de $p(e)\,p(f \mid e)$. Cal dir que existeixen nombroses tècniques de suavitzat de models de llenguatge. A [MS99] podem trobar un estudi detallat sobre tècniques de suavitzat i tècniques d'estimació de models d'n-grames.

1.2.4 Models de traducció

Com s'ha esmentat anteriorment, el model de traducció és l'element encarregat de valorar com de bona és una traducció. Formalment, un model de traducció p d'un vocabulari $\mathcal{E}$ a un vocabulari $\mathcal{F}$ amb una sèrie de paràmetres θ, és una distribució de probabilitat $p(f \mid e)$ que satisfà les següents propietats [GV03]:

$$\begin{aligned} & p(f \mid e) \geq 0,\ p(error \mid e) \geq 0, \\ & p(error \mid e) + \sum_f p(f \mid e) = 1 \end{aligned} \tag{1.16}$$

on $p(f \mid e)$ representa la probabilitat de que un frase e es traduïsca en f, i $p(error \mid e)$ la probabilitat de que no es puga obtindre cap traducció per a la frase e. Un model de traducció és deficient si $\exists e : p(error \mid e) > 0$.

Aleshores, es pretén trobar aquells paràmetres θ que maximitzen el logaritme de la versemblança de les dades o corpus d'entrenament $C = (f_n, e_n) : n = 1, \ldots, N$. Aquesta funció objectiu es defineix d'aquesta forma [GV03]:

$$L(p) = \frac{\sum_{i=1}^{N} \log p(f_i \mid e_i)}{N} = \sum_{f,e} q(f,e) \log p(f \mid e) \tag{1.17}$$

on $q(f,e)$ és la distribució empírica de la mostra, que és igual a $1/N$ vegades el nombre

de vegades (0 ó 1, en general no existeixen repeticions) que el parell (f, e) apareix al corpus C, és a dir: $q(f, e) = \frac{N(f,e)}{N}$.

1.3 Aproximacions a la traducció automàtica estadística

En aquest capítol introduirem algunes de les aproximacions més destacables de la traducció automàtica estadística. Cal notar que les diferencies existents entre aquestes resideixen en el disseny i/o implementació dels respectius models de traducció. D'altra banda, els models de llenguatge poden assumir qualsevol aproximació, encara que en general solen estar basats de models d'n-grames. D'aquesta forma, trobem models de traducció basats en arbres (o jeràrquics), transductors estocàstics, paraules, o seqüències de paraules. Prestarem més atenció al models basats en paraules i en especial als basats en seqüències de paraules, ja que les millores que es plantegen a aquest treball estan pensades per ser aplicades en eixa aproximació.

Cal mencionar que, per donar una visió més intuïtiva de tots els conceptes que tot seguit s'exposen, recorrerem a l'ús d'exemples enfocats a traduir del català ($\mathcal{F}$) a l'anglès ($\mathcal{E}$) per modelar el model de traducció invers $p(f \mid e)$ (traducció d'anglès a català). Cal notar que la construcció del model directe $p(e \mid f)$, es realitzaria de forma anàloga a l'invers.

1.3.1 Models basats en paraules

Els models basats en paraules[13], com el seu propi nom indica, restringeixen el procés de traducció a nivell de paraula, donant lloc al que anomenem traducció lèxica. Cal notar que aquesta aproximació no representa l'estat d'art actual de la TA. És més, ens podem adonar fàcilment que no és molt bona idea afrontar el procés traducció d'aquesta forma. No obstant, aquests models són la base de les tècniques i mètodes emprats en l'actualitat (i en aquest treball), fet pel qual el seu estudi és imprescindible a la vegada que obligat.

Traducció lèxica

Si ens proposarem traduir una frase en anglès (e) al català (f) sense tenir massa coneixements d'anglès, probablement consultaríem un diccionari bilingüe i buscaríem, per a cada paraula en anglès, la seva traducció corresponent en català. Molt prompte ens adonaríem d'un problema: hi ha paraules en anglès que es poden traduir en múltiples d'un altre idioma atenent al seu significat[14]. Per exemple, considerem la paraula en anglès *glass*. Si consultem un diccionari bilingüe anglès - català[15], les possibles traduccions d'aquesta paraula són *vidre, espill, got, vas, tassó,* o *copa*. Òbviament

[13] *Word-Based Models*, en anglès.
[14] És al que anomenem accepcions.
[15] http://www.diccionaris.net

Paraula	Ocurrències	Prob.
vidre	450	0.45
got	300	0.30
espill	150	0.15
tassó	70	0.07
vas	30	0.03

Taula 1.2: Exemple d'estimació de probabilitats de traducció lèxiques per a la paraula *glass* (on $N(\text{glass}) = 1000$).

elegirem aquella accepció que s'ajuste millor al context (si tenim els suficients coneixements per esbrinar-lo, és clar). Però, posem-nos per moments en la "pell" d'una màquina. D'acord amb el model que estem plantejant, un computador no sap res més que la pròpia paraula que li subministrem, en aquest cas *glass*. Al no disposar de cap informació del context, assumirem que la màquina apostarà per la traducció més probable. Per tant, en aquest sentit, un computador deuria conèixer, per a cada paraula de la frase d'eixida e, les seves possibles traduccions en $\mathcal{F}$ i la probabilitat associada a cadascuna d'elles (recordem que estem considerant el model de traducció invers).

L'estimació d'aquesta distribució de probabilitat $p(f \mid e)$ es realitza per màxima versemblança a partir d'un corpus d'entrenament de parells de paraules $\{(f_n, e_n) \in C : n = 1, \ldots N\}$. Per exemple, seguint amb l'exemple anterior, imaginem que la paraula *glass* apareix 1000 vegades al nostre corpus d'entrenament, i de totes aquestes, en 450 ocasions es tradueix en *vidre*, 300 en *got*, 150 en *espill*, 70 en *tassó*, i 30 en *vas*. L'estimació de les probabilitats és trivial. Per al cas concret de *vidre*:

$$t(\textit{vidre} \mid \textit{glass}) = \frac{450}{1000} = 0.45 \tag{1.18}$$

I de forma general:

$$t(f \mid e) = \frac{N(f, e)}{N(e)} \tag{1.19}$$

on $N(f, e)$ és el nombre de vegades que les paraules f i e s'esdevenen conjuntament al corpus d'entrenament, i $N(e)$ el nombre total d'ocurrències de la paraula e. A la Taula 1.2 podem observar les probabilitats estimades per a totes les possibles traduccions de *glass*.

Ampliem ara la nostra perspectiva del procés i considerem la traducció de frases completes paraula per paraula. En aquest sentit, anem a plantejar-nos la traducció de la frase *the glass is full of milk*. Podem aterrir-nos per moments amb la possibilitat que el nostre model ens proporcione la traducció *el vidre està ple de llet*, entre altres coses perquè la traducció més probable de *glass* és *vidre*. Afortunadament, el model

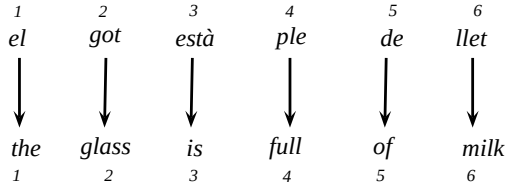


Figura 1.3: Exemple senzill d'alineament de paraules entre dues frases.

de llenguatge ens cobrirà les esquenes, apostant poc (presumiblement) per aquesta hipòtesi, i conferint major probabilitat a frases com *el got està ple de llet*, que seria una traducció correcta.

Alineaments de paraules

Fins al moment hem assumit que, per estimar les probabilitats de traducció a partir d'un corpus d'entrenament, coneixem en quina o quines paraules del vocabulari d'eixida s'ha alineat (o traduït) cada paraula d'entrada. Doncs bé, en realitat no disposem d'aquesta informació, ja que etiquetar tot un corpus d'entrenament, que està format per milers de parells de frases, amb aquesta informació, seria una tasca molt costosa. Per tant, donada una frase d'eixida $e = e_1\, e_2 \ldots e_i \ldots e_I$ i una frase d'entrada $f = f_1\, f_2 \ldots f_j \ldots f_J$, desconeixem quines paraules d'e són traducció de f, i viceversa. Aquestes "associacions" entre les paraules de f i d'e s'anomenen alineaments. Obtindre els alineaments de paraules per a cada parell de frases $(f_n, e_n) \in C$ constitueix un problema complex al qual s'hem d'enfrontar per construir aquests models.

Per exemple, si considerem el parell de frases $e =$ *the glass is full of milk* i $f =$ *el got està ple de llet*, un un possible alineament de paraules entre aquestes dues frases és el que podem observar a la Figura 1.3.

Un alineament es defineix com una funció a que relaciona posicions de paraules de la frase d'entrada (j), amb posicions de paraules de la frase d'eixida (i) [Koe10]:

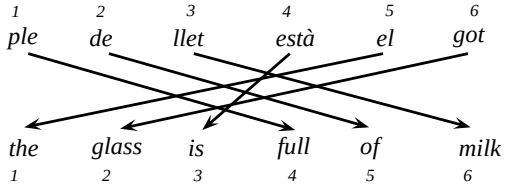
$$a : j \to i \tag{1.20}$$

Aquesta funció està completament definida: per a tota paraula de la frase d'entrada f en la posició j existeix una paraula de la frase origen e a la posició i amb la qual està relacionada:

$$\{\, j \;:\; \exists i \in a_j = i \,\}\; \forall j \tag{1.21}$$

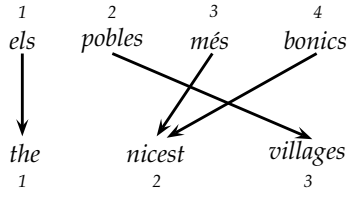
Per a l'exemple anterior, la funció d'alineament proporciona aquestes assignacions:

$$a : \{1 \to 1, 2 \to 2, 3 \to 3, 4 \to 4, 5 \to 5, 6 \to 6\}$$



$$a : \{1 \rightarrow 4, 2 \rightarrow 5, 3 \rightarrow 6, 4 \rightarrow 3, 5 \rightarrow 1, 6 \rightarrow 2\}$$

Figura 1.4: Exemple d'alineament en el que les paraules alineades ocupen posicions diferents en cadascuna de les frases.



$$a : \{1 \rightarrow 1, 2 \rightarrow 3, 3 \rightarrow 2, 4 \rightarrow 2\}$$

Figura 1.5: Exemple d'alineament en el que una paraula d'eixida es troba relacionada amb més d'una paraula d'entrada.

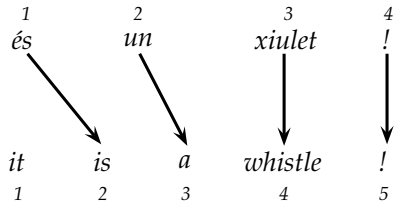
Cal notar que, a pesar d'estar modelitzant la traducció d'anglès a català (model invers), la funció d'alineament assigna posicions de paraules en català a posicions de paraules en anglès.

Com podem observar, l'alineament de la Figura 1.3 és molt senzill, ja que les paraules alineades presenten el mateix ordre a la frase d'entrada i d'eixida. En general, podem trobar-nos en aquests casos:

1. Les paraules alineades apareixen en el mateix ordre a la frase d'entrada i d'eixida (Figura 1.3).

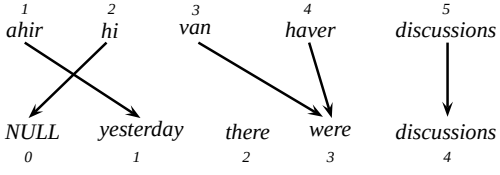
2. Les paraules alineades poden aparèixer en ordres diferents a cadascuna de les frases (Figura 1.4).

3. Una paraula de la frase d'eixida e pot estar alineada amb més d'una paraula de la frase d'entrada f per expressar el mateix concepte (Figura 1.5).



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

Figura 1.6: Exemple d'alineament en el que paraules de la frase d'eixida no han estat alineades amb cap paraula de la frase d'entrada.



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

Figura 1.7: Exemple d'alineament en el que paraules de la frase origen es troben alineades amb la paraula destí especial *NULL*.

4. Poden haver-hi paraules de la frase d'eixida e que no tenen un clar equivalent en la frase d'entrada f, i que per tant no són alineades. (Figura 1.6).

5. De la mateixa forma, poden haver-hi paraules de la frase d'entrada f que no tenen relació amb cap paraula de la frase d'eixida. Aquest cas rep un tracte especial: donat que la funció d'alineament és completa i per tant totes les paraules de la frase d'entrada deuen estar alineades amb exactament una paraula de la frase d'eixida, s'introdueix una paraula especial, *NULL*[16], que s'alinea amb les paraules de la frase d'entrada que es troben en aquest cas (Figura 1.7).

Com hem esmentat abans, els nostres corpus d'entrenament no es troben alineats a nivell de paraula, sinó a nivell de frase, amb la qual cosa resulta necessari calcular aquests alineaments per poder estimar els paràmetres del model de traducció. Aleshores, un model de traducció basat en paraules presenta de forma implícita un model d'alineament que haurem d'estimar, de forma que estarem modelitzant

[16] Paraula nul·la.

$p(f, a \mid e)$ (és a dir, donada la frase e, quina és la probabilitat d'obtenir la frase f amb un alineament a).

Models d'IBM

Amb tot, acabem de proposar un model de traducció $p(f, a \mid e)$ que considera l'alineament existent entre les paraules de les frases d'entrada i d'eixida com una variable oculta (no disposem d'aquests alineaments). Aquest model és conegut com el Model 1 d'IBM, proposat per P.F. Brown i altres investigadors d'IBM a l'any 1990 [B$^+$90, B$^+$93], i que va suposar una gran revolució en el camp de la TA.

El primer model d'IBM defineix la probabilitat de traduir una frase d'eixida $e = e_1 \dots e_I$ de longitud I en una frase destí $f = f_1 \dots f_J$ de longitud J , amb una funció d'alineament a que relaciona cada posició (paraula) de la frase d'entrada f_j amb una posició (paraula) de la frase d'eixida e_i, d'aquesta forma [Koe10]:

$$p(f, a \mid e) = \prod_{j=1}^{J} \frac{1}{I+1}\, t(f_j \mid e_{a(j)}) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} t(f_j \mid e_{a(j)}) \tag{1.22}$$

és a dir, la probabilitat de traduir una frase e en una frase f d'acord amb un alineament a és el producte de les probabilitats del model de traducció lèxica per traduir cadascuna de les J paraules de la frase d'entrada a la posició j en les seves respectives paraules d'eixida a la posició $a(j)$, normalitzat pel nombre total d'alineaments possibles (notar que per la inclusió de la paraula *NULL* hi ha un total de $I + 1$ paraules d'eixida, a combinar amb J paraules d'entrada).

Si a l'exemple de la Figura 1.3 li apliquem la definició d'aquest model, la probabilitat de traduir la frase *the glass is full of milk* en *el got està ple de llet* és:

$$p(f, a \mid e) = \frac{1}{7^6} \cdot t(el \mid the) \cdot t(got \mid glass) \cdot t(està \mid is) \cdot t(ple \mid full) \cdot t(de \mid of) \cdot t(llet \mid milk)$$

La relació existent entre un model de traducció i el model d'alineament subjacent ve donat per la següent expressió [Koe10, GV03]:

$$p(f \mid e) = \sum_{a} p(f, a \mid e) \tag{1.23}$$

és a dir, la probabilitat de que la frase e es traduisca en la frase f considera tot possible alineament entre e i f.

En la pràctica sols es considera l'alineament més probable (l'anomenat alineament de Viterbi) per aproximar el càlcul del sumatori definit sobre a [GV03]:

$$\sum_{a} p(f, a \mid e) \approx \max_{a} p(f, a \mid e) \tag{1.24}$$

Cal notar que l'alineament més probable és aquell que maximitza la probabilitat de que la frase e es traduisca en la frase f:

$$\hat{a} = \underset{a}{\operatorname{argmax}}\, p(f, a \mid e) \tag{1.25}$$

Un model de traducció basat en paraules depèn d'una sèrie de paràmetres θ que són estimats maximitzant la versemblança per a un corpus d'entrenament $\{(f_n, e_n) \in C : n = 1, \ldots, N\}$ [GV03]:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \prod_{n=1}^{N} \sum_{a} p_\theta(f_n, a \mid e_n) \tag{1.26}$$

Ara bé, per a realitzar aquesta estimació de paràmetres tenim un problema. Des d'un primer moment hem assumit que disposàvem dels alineaments paraula per paraula per a cada parell de frases del corpus d'entrenament, amb la qual cosa estimar les probabilitats de traducció lèxica (o siga, estimar els paràmetres del model de traducció lèxica) per màxima versemblança ha estat quasi trivial. Però com sabem, no disposem de cap model d'alineament. No podem estimar el nostre model de traducció a partir d'informació incompleta: és per això que l'alineament entre paraules és considerat una variable oculta al nostre model. D'una banda, si coneixerem els alineaments entre paraules, resultaria fàcil estimar el model de traducció de paraules. Per l'altra banda, si disposarem del model ja estimat, seria possible obtindre els alineaments de paraules més probables per a cada parell de frases. Malauradament, no disposem ni d'una cosa ni de l'altra.

Ara bé, existeix una tècnica amb la qual sí podem fer front a aquest problema. L'algoritme EM[17] [DLR77] permet encarar el problema de la informació incompleta i de l'estimació de paràmetres de models probabilístics amb variables ocultes. Podem trobar més informació al respecte a [Koe10, B$^+$93].

A més del Model 1 d'IBM, existeixen quatre models de complexitat creixent que introdueixen millores respecte als models anteriors. Aquests són:

- **Model 2 d'IBM**: Afegeix al model original un model d'alineament absolut.
- **Model 3 d'IBM**: Afegeix un model de fertilitats[18].
- **Model 4 d'IBM**: Afegeix un model d'alineament relatiu.
- **Model 5 d'IBM**: Esmena la deficiència que presenten conjuntament els models 3 i 4.

Per a més informació sobre els models d'IBM, remetem al lector a [Koe10, GV03, B$^+$90, B$^+$93].

[17] *Expectation-Maximization Algorithm.*

[18] Modela la probabilitat de que una paraula de la frase origen s'alinee amb n paraules de la frase destí.

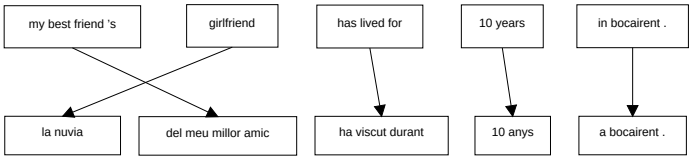


Figura 1.8: Exemple del procés de traducció automàtica en sistemes basats en seqüències de paraules.

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. Ara bé, com ja s'ha esmentat, traduir a nivell de paraula no és molt recomanable, doncs hi ha paraules d'un llenguatge origen que es poden traduir en dues o més paraules del llenguatge destí, i viceversa. Considerar solament paraules aïllades, a més, impedeix capturar el context en que es troben, i en conseqüència, la semàntica o significat de la paraula, informació que resultaria molt valuosa a l'hora d'escollir entre múltiples accepcions. Aquests són els principals motius pels quals resulta més convenient escollir una unitat bàsica de traducció de major grandària que permeta capturar informació contextual. Aquest element bàsic, que pot englobar una o més paraules, s'anomena seqüència de paraules[19]. Les seqüències de paraules s'obtenenen dividint una frase completa en segments de paraules contigues i de longitud variable. La Figura 1.8 il·lustra com funcionen aquests sistemes: la frase d'eixida e és segmentada en seqüències de paraules, de forma que cada seqüència $\overline{e}$ és traduïda al llenguatge origen $\mathcal{F}$, originant seqüències $\overline{f}$ que són reordenades a posteriori (si s'escau). Si ens fixem en l'exemple, la seqüència de paraules en anglès *girlfriend* es tradueix de forma òptima al català com *la núvia*. Per tal de conèixer les traduccions més probables per a cada seqüència de paraules, haurem de disposar d'una distribució de probabilitat que ens proporcione la probabilitat d'obtindre una seqüència de paraules d'entrada $\overline{f}$ com a traducció d'una seqüència $\overline{e}$ d'eixida (atès que considerem el model de traducció invers).

Cal destacar que els sistemes de TA basats en seqüències de paraules no empren cap mètode lingüístic per segmentar una frase. Per exemple, si ens fixem novament en l'exemple de la Figura 1.8, una de les seqüències de paraules obtingudes és *has lived for*, la qual, baix un d'un punt de vista sintàctic, representa un agrupament incorrecte, doncs seria més lògic segmentar de forma separada verb (*has lived*) i complement (*for 10 years*). No obstant, realitzar aquest tipus d'agrupacions (*has lived for*) resulta molt útil, ja que permet capturar fàcilment el significat sempre confús de les preposicions. Vegem un exemple: si tractarem de traduir de forma aïllada la preposició *for* podríem escollir entre *per, per a, a causa de, en, a favor de, durant,* , etc. Ara bé, atenent al context en que apareix, la traducció més apropiada sembla ser *durant*. Aquesta valuosa informació es captura agrupant verb i preposició en la segmentació, i generant

[19] *Phrase*, en anglès.

la corresponent seqüència de paraules traducció (en el nostre cas *has lived for*, en *ha viscut durant*). Per tant, realitzar agrupaments de paraules en lloc de considerar paraules aïllades ens permetrà resoldre moltes de les ambigüitats que se'ns puguen presentar al procés de traducció.

Addicionalment, emprar aquest enfocament basat en seqüències de paraules comporta un benefici extra. El poder obtenir segments de longitud variable a partir de les frases origen del corpus d'entrenament ens permet "memoritzar" les seves respectives traduccions, podent arribar a emmagatzemar fins i tot traduccions de frases completes. Ara bé, 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

Formalment, un model de traducció invers $p(f \mid e)$ basat en seqüències de paraules defineix la probabilitat de traduir una frase d'eixida $e = \overline{e}_1, \overline{e}_2, \ldots, \overline{e}_K$, segmentada en K seqüències de paraules, en una frase d'entrada $f = \overline{f}_1, \overline{f}_2, \ldots, \overline{f}_K$ de la següent forma:

$$p(f \mid e) = \prod_{k=1}^{K} p(\overline{f}_k \mid \overline{e}_k) \tag{1.27}$$

és a dir, $p(f \mid e)$ es defineix com el producte de les probabilitats de traduir cada seqüència de paraules $\overline{e}_k$ en $\overline{f}_k$. Cal notar que ambdues frases e i f presenten el mateix nombre de segmentacions, fet pel qual tota seqüència de paraules d'eixida $\overline{e}_k$ genera una seqüència de paraules d'entrada $\overline{f}_k$ durant el procés de traducció.

Extracció de seqüències de paraules

Com ja sabem, partim d'un corpus de parells de frases $\{(f_n, e_n) \in C : n = 1, \ldots N\}$, a partir del qual s'obtenen els parells de seqüències de paraules $(\overline{f}, \overline{e})$, que són els paràmetres del nostre model. Com s'obtenen aquestes seqüències de paraules? És obvi que un criteri aleatori no té gens de sentit, doncs a banda de que per cada parell de frases es podrien extraure desenes, centenars, o inclús milers de parells de seqüències, cosa que seria impossible de gestionar, extraure per exemple un hipotètic parell de seqüències de paraules (*10 anys*, *my best friend*), és inútil i contraproduent. Un possible mètode seria considerar la segmentació 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* [DLR77]. No obstant, emprar aquesta tècnica d'estimació és una tasca molt costosa, així que optarem per una estratègia alternativa més manejable basada en tècniques heurístiques, a costa de perdre certa correcció i rigor estadístic. Es tracta de partir dels alineaments de paraules per a cada parell de frases (que es poden obtindre mitjançant els models d'IBM, Secció 1.3.1) en ambdues direccions de traducció, combinar-los

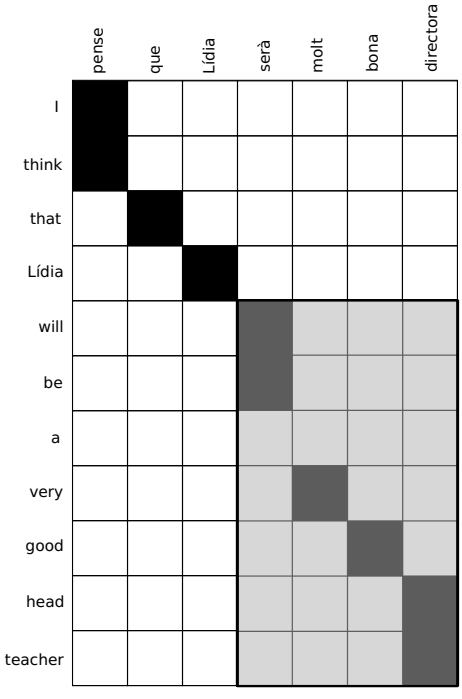


Figura 1.9: Exemple d'extracció de parells de seqüències de paraules a partir de l'alineament entre paraules d'un parell de frases.

aplicant un algorisme heurístic, i extraure aquells parells de seqüències de paraules que siguen consistents amb els alineaments combinats.

Un parell de seqüències de paraules $(\overline{f}, \overline{e})$ és vàlid i consistent amb un alineament A si totes les paraules $f_1, \ldots, f_n \in \overline{f}$ que presenten alineaments en A troben totes les seves paraules alineades en la seqüència $\overline{e}$, i viceversa. Més formalment [Koe10]:

$$\begin{aligned}(\overline{f}, \overline{e}) \text{ és consistent amb } A \Leftrightarrow \quad & \forall f_j \in \overline{f} : (f_j, e_i) \in A \Rightarrow e_i \in \overline{e} \wedge \\ \wedge \quad & \forall e_i \in \overline{e} : (f_j, e_i) \in A \Rightarrow f_j \in \overline{f} \wedge \\ \wedge \quad & \exists f_j \in \overline{f},\, e_i \in \overline{e} : (f_j, e_i) \in A \end{aligned} \tag{1.28}$$

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 (1.28). 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.

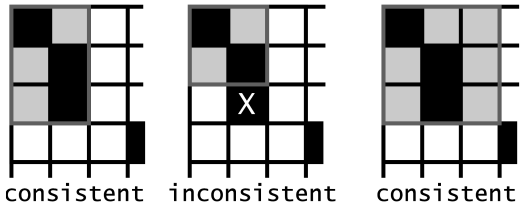


Figura 1.10: Exemples de possibles seqüències de paraules consistents i inconsistents, depenent de l'alineament entre les paraules d'un parell de frases qualsevol.

A la Figura 1.9 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 (1.28)) 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. Podem observar, a continuació, part dels possibles parells de seqüències de paraules que es poden extreure de forma vàlida:

pense – I think
pense que – I think that
que – that
pense que Lídia – I think that Lídia
Lídia – Lídia
que Lídia – that Lídia
pense que Lídia serà – I think that Lídia will be
...
pense que Lídia serà molt bona directora – I think that Lídia will be a very good teacher

D'altra banda, la Figura 1.10, extreta de [Koe10], mostra casos en els que un possible parell de seqüències de paraules és consistent o no donat l'alineament entre paraules d'un parell de frases. Tenint en ment que totes les paraules que formen part del parell de seqüències deuen d'estar alineades les unes amb les altres (si és que presenten algun alineament), podem determinar que el primer exemple (esquerra) és completament vàlid; el segon exemple (centre) viola una de les condicions, ja que

un punt d'alineament marcat amb una creu es troba fora del rang de paraules que abasta la seqüència; i per últim, el tercer exemple (dreta) representa una extracció vàlida, doncs inclou la paraula de la tercera columna que no es troba alineada amb cap paraula.

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. D'altra banda, cal tenir en compte que les seqüències de paraules llargues permeten capturar major informació de context, motiu pel qual aquestes són especialment útils en la traducció d'expressions fetes (seria un sistema de TA capaç de traduir correctament l'expressió feta en català *a burro barra*, que significa *aleatòriament*?). 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 $\{(f_n, e_n) \in C : n = 1, \dots N\}$, de la següent forma [Koe10]:

$$p(\overline{f} \mid \overline{e}) = \frac{N(\overline{f}, \overline{e})}{\sum_{\overline{f}'} N(\overline{f}', \overline{e})} \tag{1.29}$$

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

Reordenament de seqüències de paraules

Com hem vist a l'Equació 1.27, el model de traducció s'ha definit únicament a partir de les probabilitats de traducció de cada seqüència de paraules en que s'ha segmentat la frase origen. Com hem esmentat al principi d'aquesta secció, durant el procés de traducció la frase origen és segmentada en seqüències de paraules que posteriorment són traduïdes i, possiblement, reordenades, ja que l'ordre en que apareixen les paraules pot ser diferent en cadascun dels idiomes considerats. En general, preferim que les seqüències de paraules destí no es reordenen (és a dir, que les seqüències es traduïsquen de forma monòtona, seguint el mateix ordre que les seqüències de paraules de la frase origen), o bé que canvien la seva posició el mínim possible, doncs salts llargs són infreqüents i, per tant, poc probables (tot i que això depèn del parell de llenguatges). Per tant, premiarem la immobilitat de les seqüències de paraules traduïdes, i penalitzarem els salts llargs al reordenar-les.

En aquest sentit, s'introdueix un model de distorsió o reordenament que es defineix com una funció exponencial $d(x) = \alpha^x$ amb un valor apropiat del paràmetre $\alpha \in [0, 1]$, on x és la distància en paraules del salt efectuat per la seqüència de paraules al ser reordenada. La distància del salt x es calcula de la següent forma:

$$x = |\text{inici}_k - \text{fi}_{k-1} - 1| \tag{1.30}$$

on inici_k és la posició, respecte a la frase origen, de la primera paraula de la seqüència de paraules d'entrada que es tradueix a la k-èsima seqüència de paraules d'eixida, i fi_{k-1} és la posició de l'última paraula de la seqüència de paraules d'entrada que es tradueix a la $(k-1)$-èsima seqüència de paraules d'eixida (és a dir, la seqüència d'eixida precedent a la considerada). Aleshores, si combinem aquest model amb el model de traducció de seqüències de paraules, obtenim un nou model proporcional al presentat a l'Equació (1.27):

$$p(e \mid f) \propto \prod_{k=1}^{K} p(\overline{e}_k|\overline{f}_k)\ d(|\text{inici}_k - \text{fi}_{k-1} - 1|) \tag{1.31}$$

Com veiem, 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. Per exemple, poden haver-hi certes seqüències que, al traduir-les a un llenguatge concret, siguen més propenses a ser reordenades que altres. Un cas típic és el que ocorre amb els noms i els adjectius, que al traduir de l'anglès al català canvien molt freqüentment la seva posició. Existeixen altres models de reordenament més complexos que permeten capturar aquesta informació que estudiarem al següent capítol.

Per concloure aquesta secció, remarcar que, en el moment en que s'escriuen aquestes línies, els sistemes de TA basats en seqüències de paraules són els que millors prestacions ofereixen en comparació amb altres aproximacions, i un dels principals objectes de la comunitat investigadora [CBKM+10, Koe10].

1.3.3 Models basats en transductors

Aquesta aproximació tracta de modelar la relació entre el llenguatge origen i el llenguatge destí mitjançant autòmates o transductors d'estats finits estocàstics[20] (TEFS). La característica més interessant d'aquesta aproximació és que els TEFS inclouen de forma implícita un model de llenguatge de l'idioma destí. Conseqüentment, es pot modelar directament la probabilitat conjunta de la frase origen f i destí e. Aleshores, baix aquest principi, la traducció més probable e ve donada per la següent expressió:

$$\hat{e} = \underset{e}{\text{argmax}}\, p(f, e) \tag{1.32}$$

[20] Amb probabilitats associades a transicions i/o estats.

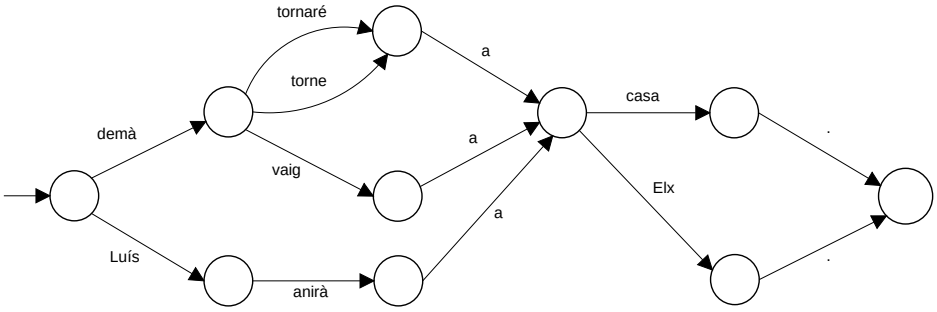


Figura 1.11: Exemple d'un transductor estocàstic d'estats finits (sense probabilitats associades).

Llavors no és necessari aplicar la regla de Bayes i modelar de forma separada un model de traducció i un model de llenguatge, doncs d'acord amb el que hem vist a la Secció 1.2.1, $p(f, e) = p(e)\, p(f|e)$.

Formalment, un transductor d'estats finits estocàstic T es defineix com una tupla [BBC+09]:

$$T = \langle \Sigma, \Delta, Q, q_0, p, f \rangle \tag{1.33}$$

on Σ és un conjunt finit de símbols del llenguatge origen, Δ és un conjunt finit de símbols del llenguatge destí, Q és un conjunt finit d'estats de l'autòmata, q_0 és l'estat inicial, i p i f són dues funcions $p : Q \times \Sigma \times \Delta^* \times Q \Rightarrow [0, 1]$ (probabilitats de les transicions) i $f : Q \Rightarrow [0, 1]$ (probabilitats dels estats finals) respectivament que satisfan, per a tot estat q:

$$\forall q \in Q : f(q) + \sum_{(f, \tilde{e}, q') \in \Sigma \times \Delta^* \times Q} p(q, f, \tilde{e}, q') = 1 \tag{1.34}$$

és a dir, la suma de les probabilitats de totes les possibles transicions d'un estat q (incloent la probabilitat de finalitzar el recorregut en dit estat) és 1.

Considerem ara un camí P, associat a parells de símbols $(f, e) \in \Sigma^* \times \Delta^*$, i format per una seqüència de J transicions $\phi = (q_0, f_1, \tilde{e}_1, q_1)$, $(q_1, f_2, \tilde{e}_2, q_2)$, $\ldots$, $(q_{J-1}, f_J, \tilde{e}_J, q_J)$, en la que $f_1\ f_2\ \ldots\ f_J = f$ i $\tilde{e}_1\ \tilde{e}_2\ \ldots\ \tilde{e}_J = \tilde{e}$. La probabilitat d'un camí és el producte de la probabilitat de cadascuna de les transicions, incloent la probabilitat d'estat final de l'últim estat [BBC+09]:

$$p(\phi) = \prod_{j=1}^{J} p(q_j - 1, f_j, \tilde{e}_j, q_j) f(q_J) \tag{1.35}$$

La probabilitat conjunta del parell (f, e) d'acord amb un camí P és la suma de les probabilitats de tots els camins associats al parell [BBC+09]:

$$p(f, e) = \sum_{\phi} p(\phi) \tag{1.36}$$

A la Figura 1.11 podem observar un transductor estocàstic d'estats finits que modela simultàniament un model de llenguatge en català i un model de traducció per a la frase en anglès *Luís will go to Elx.* (No s'inclouen les probabilitats de les transicions, ni els símbols del llenguatge origen). Per aprofundir més en aquesta aproximació, remetem al lector a [CV04, CVP05, CV07].

Existeixen moltes tècniques per entrenar aquests models. Una d'elles és l'anomenada *Grammatical Inference and Alignments for Transducer Inference* (GIATI), una tècnica híbrida que empra tècniques estadístiques per construir l'estructura dels transductors i estimar les probabilitats de les transicions. Podem trobar més informació al respecte a [CV04, CVP05, BBC+09].

1.3.4 Models jeràrquics o basats en arbres

Aquesta aproximació a la TAE, al contrari que totes les precedents vistes a aquest capítol, té com objectiu introduir informació lingüística en el procés de traducció. Aquests models es basen en la representació del llenguatge a través d'arbres que permeten explotar tota relació sintàctica entre paraules i frases. Per representar aquests arbres s'empren el que s'anomenen *gramàtiques d'estructuració de frases*[21], que són unes gramàtiques incontextuals on:

- **Símbols no terminals (NT)**: Representen etiquetes de sintagmes o classes gramaticals de paraules, com per exemple VP (*verb phrase* o sintagma verbal) o NP (*noun phrase* o sintagma nominal) quant a etiquetes de sintagmes, o bé DET (determinant) o NN (nom) quant a categories gramaticals.
- **Símbols terminals (T)**: Representen les paraules d'un vocabulari $\mathcal{V}$ (paraules del llenguatge).
- **Símbol inicial**: S, que representa a la frase completa (*Sentence*).
- **Regles**: Del tipus $NT \rightarrow [NT, T]^+$. Per exemple, la regla $NP \rightarrow DET\ NN$, denota que, en un determinat llenguatge, un sintagma nominal (NP) és format per un determinant (DET) seguit d'un nom (NN).

A la Figura 1.12 podem observar un arbre generat per una gramàtica d'estructuració de frases, extret de [Koe10]. Com veiem, l'arbre representa el modelat gramatical i sintàctic de la frase en anglès *I shall be passing on to you some comments*,

[21] En anglès, *phrase structure grammars*.

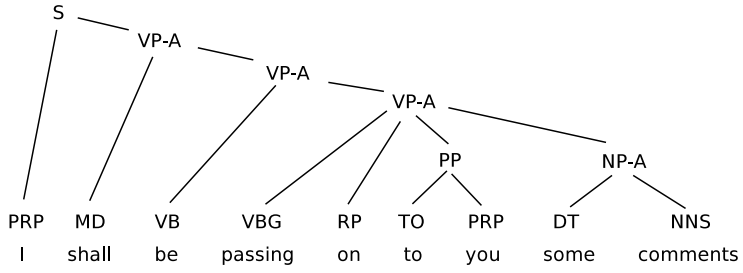


Figura 1.12: Exemple d'arbre que representa la frase en anglès *I shall be passing on to you some comments* modelada amb una gramàtica d'estructuració de frases.

de forma que les paraules són etiquetades per la seva categoria lingüística amb símbols no terminals de la gramàtica, i les regles de derivació s'expliciten amb les arestes que uneixen els diferents símbols no terminals. L'element més elevat de l'arbre és el símbol inicial de la gramàtica S, que representa la frase completa.

Fins al moment hem vist com s'estructura sintàcticament i gramaticalment una frase d'un llenguatge en concret en un arbre modelat per una gramàtica. Però realment, l'interès recau en poder representar l'equivalència sintàctica entre parells d'idiomes. Precisament és aquest el propòsit de les *gramàtiques síncrones d'estructuració de frases*[22], que permeten representar i relacionar parells de frases en parells d'arbres sintàctics.

Amb un petit exemple entendrem fàcilment aquest propòsit. Un possible sintagma nominal en anglès ve donat per la seqüència determinant, adjectiu i nom, com mostra, per exemple, la frase *the blue cat*, sent la seva regla gramatical associada $NP \rightarrow DET\ ADJ\ NN$. L'equivalència d'aquest sintagma al català ve donat per la seqüència determinant, nom i adjectiu, com per exemple *el gat blau*, sent la seva regla gramatical $NP \rightarrow DET\ NN\ ADJ$. Una gramàtica síncrona ens permet capturar la variació sintàctica existent entre els dos llenguatges:

$$NP \rightarrow DET_1\ ADJ_2\ NN_3 \mid DET_1\ NN_3\ ADJ_2$$

Cal notar que cada símbol no terminal es troba indexat per poder identificar les correspondències úniques entre les paraules d'ambdós llenguatges.

En realitat, aquestes gramàtiques són estocàstiques (les regles duen associades probabilitats d'aplicació), sent les probabilitats de les regles estimades a partir d'un corpus d'entrenament. Aleshores, en termes generals, la probabilitat de traduir una frase d'eixida e en una frase d'entrada f a partir d'un arbre t que representa la gramàtica síncrona que relaciona sintàcticament ambdues frases, ve donada per el producte de les probabilitats de les regles r_i emprades per construir l'arbre:

[22] En anglès, *syncronous phrase structure grammars*.

$$p(f, t \mid e) = \prod_i r_i \tag{1.37}$$

En l'actualitat s'ha demostrat que els sistemes de TA basats en arbres sintàctics ofereixen prestacions similars, i en certs casos puntuals, superiors, als sistemes basats en seqüències de paraules. Podem trobar més informació sobre aquesta aproximació a [Koe10].

Capítol 2

Traducció automàtica estadística basada en seqüències de paraules

Al capítol introductori, entre d'altres coses, hem donat una visió general de la disciplina de la traducció automàtica, prenent especial atenció a l'aproximació estadística. De forma similar, a l'explorar els diferents enfocaments estadístics de la TA, hem donat major èmfasi als models basats en seqüències de paraules, ja que en general, el sistemes que segueixen aquesta aproximació són els que millors prestacions ofereixen [CBKM+10]. Llavors, amb la intenció de millorar l'estat d'art actual, per a la realització d'aquest treball s'ha emprat, com a base per a la implementació de possibles millores, un sistema de TAE basat en seqüències de paraules anomenat *Moses* [KHB+07].

En aquest capítol introduirem, en primer lloc, el sistema *Moses*, situant-lo en el marc teòric de la TAE basada en seqüències de paraules, i explicarem les extensions que afegeix al model estàndard presentat al capítol 1.3.2. També detallarem com entrenar el sistema i com optimitzar-lo. Posteriorment, analitzarem les mancances que presenta el model de traducció implementat en aquest sistema i presentarem les possibles millores que es pretenen implantar per millorar les seves prestacions.

2.1 El sistema de TA Moses

Moses és un sistema de TA basat en seqüències de paraules, de codi obert, i desenvolupat per una gran comunitat coordinada per la Universitat d'Edimburg. Donat que aquest sistema implementa un model logarítmic-lineal, en primer lloc introduirem a nivell general aquest model per a posteriorment instanciar-lo al cas particular de *Moses*. Tota la informació que apareix en aquesta secció es pot ampliar consultant [Koe10].

2.1.1 Models logarítmic-lineals

Un model logarítmic-lineal[1] (log-lineal, d'ara endavant) és un model emprat recentment al món del reconeixement de formes que permet construir una distribució de probabilitat integrant diverses distribucions de probabilitat (o models), que en aquest context s'anomenem característiques o *features*, tot mitjançant una combinació lineal dels logaritmes de les probabilitats en forma exponencial. Formalment, es defineixen de la següent manera:

$$p(x) = \exp\left(\sum_{i=1}^{N} \lambda_i h_i(x)\right) \frac{1}{Z} \tag{2.1}$$

on $p(x)$ és la distribució de probabilitat a modelar, N és el nombre de característiques a considerar en el model complet, h_i és la característica i-èsima, els λ_i són els paràmetres del model associats a la característica i-èsima, i Z és el factor de normalització, definit així:

$$Z = \sum_{x'} \exp\left(\sum_{i=1}^{N} \lambda_i h_i(x')\right) \tag{2.2}$$

En el cas concret de la TA basada en seqüències de paraules, d'acord amb l'Equació (1.5) la distribució de probabilitat que volem modelar amb un model log-lineal és $p(e \mid f)$:

$$p(e \mid f) = \exp\left(\sum_{k=1}^{K} \sum_{i=1}^{N} \lambda_i h_i(f, e, \overline{f}_k, \overline{e}_k)\right) \frac{1}{Z(f)} \tag{2.3}$$

on K és el nombre de seqüències de paraules en que es descomposa la frase d'entrada, de forma que cada característica $h_i(f, e, \overline{f}_k, \overline{e}_k)$ depèn del k-èsim segment, mentre que $Z(f)$ es defineix de forma a anàloga a l'Equació (2.2). Si integrem aquest model en el criteri de decisió de l'Equació (1.5), tenim que:

$$\hat{e} = \operatorname*{argmax}_{e} \left[\exp\left(\sum_{k=1}^{K} \sum_{i=1}^{N} \lambda_i h_i(f, e, \overline{f}_k, \overline{e}_k)\right) \frac{1}{Z(f)}\right] \tag{2.4}$$

Donat que busquem aquella frase e que maximitze la probabilitat del model mitjançant la funció argmax, podem simplificar aquesta expressió: en primer lloc podem prescindir de $Z(f)$, ja que és una constant respecte a e, i per tant no afecta al criteri de decisió argmax; i en segon lloc, podem eliminar la funció exp, doncs és una funció monòtona creixent, i per tant tampoc afecta al resultat de la funció argmax, ja que tots els termes sobre els quals es calcularia el màxim serien proporcionals a la funció exp.

[1] *Log-linear model*, en anglès.

$$\hat{e} = \underset{e}{\operatorname{argmax}} \left[\sum_{k=1}^{K} \sum_{i=1}^{N} \lambda_i h_i(f, e, \overline{f}_k, \overline{e}_k) \right] \tag{2.5}$$

Si instanciem un model log-lineal emprant els models vists a la Secció 1.3.2, tenim un total de tres característiques, que són:

- $h_1(f, e, \overline{f}_k, \overline{e}_k)$: Model de traducció de seqüències de paraules $\log p(\overline{f}_k \mid \overline{e}_k)$
- $h_2(f, e, \overline{f}_k, \overline{e}_k)$: Model de distorsió $\log d(|\text{inici}_k - \text{fi}_{k-1} - 1|)$
- $h_3(f, e, \overline{f}_k, \overline{e}_k)$: Model de llenguatge $\log p(\overline{e}_k)$

Integrar tota aquesta informació en un model log-lineal presenta una sèrie d'avantatges: en primer lloc, permet ajustar l'aportació de cadascun dels models, amb els paràmetres λ_i. En segon lloc, se'ns permet afegir de forma natural nous models que poden contribuir a millorar la qualitat de les traduccions. Per últim, se'ns dona la possibilitat d'entrenar cadascun dels models de forma separada, ja que s'assumeix que són independents els uns dels altres.

A la Secció 2.1.2 descriurem les característiques addicionals, no presentades encara, que s'inclouen al model log-lineal de *Moses*, i posteriorment, a la Secció 2.1.3 descriurem el model complet d'aquest sistema, que integra totes les característiques que s'exposen a aquest document.

2.1.2 Extensions del model original

En aquesta secció descriurem les característiques, no presentades encara, que implementa *Moses* al model log-lineal.

Model de traducció de seqüències de paraules directe

Com hem vist al capítol introductori, a l'aplicar la regla de Bayes (veure Equació (1.6)) hem considerat un model de traducció invers $p(f \mid e)$, el qual hem aproximat a la traducció de seqüències de paraules (veure Equació (1.27)). Al passar d'un model purament generatiu a un model log-lineal, s'habilita la possibilitat d'incloure el model de traducció directe $p(\overline{e} \mid \overline{f})$ com una nova característica del model log-lineal, ja que podem trobar casos particulars en els que disposar del raonament directe és molt útil.

En *Moses* s'empren ambdues direccions del model de traducció de seqüències de paraules, $p(\overline{f} \mid \overline{e})$ i $p(\overline{e} \mid \overline{f})$, que amb un ajust adequat dels pesos associats s'ha demostrat que conjuntament milloren de forma significativa les prestacions globals del sistema [Koe10].

Model de suavitzat lèxic

A l'entrenar el model de traducció de seqüències de paraules podem trobar-nos amb casos no desitjats relacionats amb el tractament de seqüències de paraules poc freqüents, com és el cas en que un parell de seqüències $\overline{f}$ i $\overline{e}$ s'han extret una única vegada del corpus d'entrenament i a més de forma concurrent, fet que implica una sobreestimació del model, doncs $p(\overline{f} \mid \overline{e}) = p(\overline{e} \mid \overline{f}) = 1$. En aquest context s'introdueix un mètode de suavitzat com una característica addicional en el model log-lineal, anomenat *model de suavitzat lèxic*[2], el qual modela la probabilitat de traducció, paraula per paraula, de la seqüència de paraules $\overline{f}$ en la seqüència $\overline{e}$ a partir d'un alineament a entre les paraules d'ambdós seqüències. Dit model es troba clarament inspirat en el model 1 d'IBM vist a la Secció 1.3.1. Així, la falta d'informació que es deriva, en certs casos, en una sobreestimació del model de traducció de seqüències de paraules, es compensa incloent un model de traducció lèxic, estadísticament més significatiu i robust.

L'estimació d'aquest model ve donada per la següent equació [Koe10]:

$$lex(\overline{e} \mid \overline{f}, a) = \prod_{i=1}^{I} \frac{1}{|\{j \mid (i,j) \in a\}|} \sum_{\forall (i,j) \in a} t(e_i \mid f_j) \tag{2.6}$$

on $t(e_i \mid f_j)$ és la probabilitat de traduir la paraula f_j en e_i, i a és un alineament entre les paraules de les seqüències $\overline{f}$ i $\overline{e}$. Donat que una paraula e_i pot estar alineada amb més d'una paraula de $\overline{f}$, la probabilitat de traducció lèxica per a cada e_i és normalitzada pel nombre d'alineaments que presenta la paraula e_i.

El sistema *Moses* inclou ambdues direccions del model lèxic: $lex(\overline{e} \mid \overline{f}, a)$ i $lex(\overline{f} \mid \overline{e}, a)$.

Penalització per paraula

Una característica del model log-lineal de *Moses* que en ocasions dona problemes és el model de llenguatge. Com hem vist a la Secció 1.2.3, la majoria dels sistemes TA inclouen un model de llenguatge d'n-grames, com és el cas de *Moses*. El que no hem esmentat a dita secció és que els models de llenguatge basats en n-grames són deficients per naturalesa, doncs aquests modelen tant frases correctes com incorrectes del llenguatge destí, des d'un punt de vista lingüístic. En altres paraules, podem veure el model de llenguatge com un element que modela $\mathcal{E}^*$ (totes les possibles combinacions de paraules del vocabulari destí $\mathcal{E}$), de forma que pot conferir probabilitats altes a frases sintàcticament no vàlides.

A més, els models de llenguatge basats en n-grames són deficients per un altre motiu: en general, assignen major probabilitat a frases curtes que a frases llargues. Això és així perquè, a major longitud de frase (major nombre de paraules), major nombre d'n-grames a considerar, fet que es tradueix en un major nombre de termes de probabilitats multiplicant-se, obtenint un valor de probabilitat cada cop més baix. En

[2] *Lexical weighting*, en anglès.

definitiva, podem considerar que un model de llenguatge basat en n-grames assumeix, per a qualsevol llenguatge, que les frases curtes són molt més probables que les frases llargues.

L'efecte 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 (2.5)).

Per corregir aquest comportament no desitjat del sistema podem considerar dues opcions: bé emprar un model de llenguatge més complex (no necessàriament basat en n-grames) que reduïsca o elimine per complet la deficiència que presenta el model actual, o bé incloure al model log-lineal un factor o característica que compense l'efecte negatiu del model de llenguatge. El sistema *Moses* implementa la segon opció, amb la inclusió al model log-lineal d'un factor ω anomenat *penalització per paraula*[3], 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 hi és beneficiós afavorir construcció d'hipòtesis a partir de seqüències de paraules de major o menor longitud, mitjançant la inclusió d'un factor ρ anomenat *penalització per seqüències de paraules*[4]. Açò ve motivat perquè totes les possibles segmentacions de la frase origen són equiprobables, sent la resta de factors (models de traducció, reordenament, llenguatge) els qui determinen, de forma indirecta, la millor segmentació possible.

Més concretament, la inclusió d'aquest factor ve motivada perquè l'ús de seqüències de paraules molt llargues per generar la frase d'eixida provoca, en molts casos, que les seqüències de paraules emprades per completar la traducció siguen de poca longitud (possiblement paraules aïllades) i per tant de mala qualitat, generant finalment una traducció dolenta, però que, no obstant, pot tenir assignada una probabilitat alta, degut a que les seqüències de paraules molt llargues solen tenir associades probabilitats de traducció molt altes (inclús la màxima probabilitat, 1). Llavors, aquestes traduccions es poden perfilar com serioses candidates a erigir-se com les traduccions més probables, fet que no és desitjable.

En aquest sentit, la finalitat que es persegueix amb la inclusió d'aquest factor és afavorir aquelles segmentacions de la frase d'entrada que originen un major nombre de seqüències de paraules, i en conseqüència, que generen seqüències de paraules de longitud moderada. Així, per cada seqüència de paraules emprada per construir la frase d'eixida, s'afegeix un factor de bonificació ρ a la puntuació o probabilitat

[3] *Word penalty*, en anglés.
[4] *Phrase penalty*, en anglès.

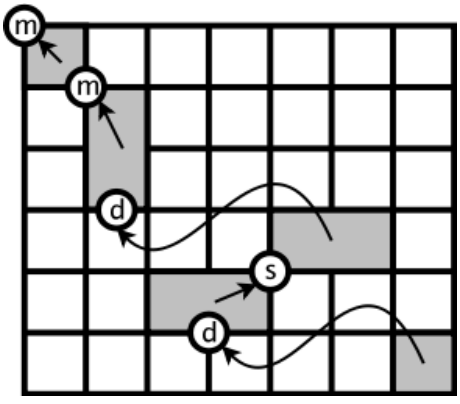


Figura 2.1: Exemples dels tres tipus d'orientacions que poden donar-se lloc a un model de reordenament lexicalitzat.

associada a la traducció. D'aquesta forma, s'atenua l'efecte negatiu del model de traducció, que sobrevalora traduccions possiblement incorrectes generades a partir de l'aplicació de seqüències de paraules molt llargues (opcions de traducció de gran longitud).

Model de reordenament lexicalitzat

A la Secció 1.3.2 hem presentat un model de distorsió $d(|\text{inici}_k - \text{fi}_{k-1} - 1|)$ molt senzill que tant sols es troba condicionat per la distància del moviment realitzat per cada seqüència de paraules de la frase destí al ser reubicat, i posteriorment hem discutit sobre la conveniència de condicionar el model a seqüències de paraules concretes, ja que poden haver-hi certes seqüències més susceptibles a ser reordenades, com havem vist que passa, per exemple, amb els sintagmes nominals formats per nom i adjectiu en llenguatges com el català o l'espanyol, on ambdues categories gramaticals intercanvien les seves posicions al traduir a l'anglès. Aquest és el motiu pel qual *Moses* inclou un model de reordenament addicional que permet aportar aquest tipus d'informació, rep el nom de *model de reordenament lexicalitzat*[5].

En termes generals, aquest nou model contempla tres tipus diferents de reordenament d'una seqüència de paraules respecte a la seqüència anterior: monòton (m), intercanvi (s) i discontinu (d)[6], el significat dels quals es pot entendre de forma intuïtiva a la Figura 2.1.

Més formalment, presentem una distribució de probabilitat que prediu el tipus d'orientació més probable que seguirà un parell de seqüències de paraules $\overline{f}, \overline{e}$:

[5] *Lexicalized reordering model*, en anglès.
[6] En anglès: *monotone* (m), *switch* (s), i *discontinous* (d).

$$\begin{aligned} &orientació \in \{m, s, d\} \\ &p(orientació \mid \overline{f}, \overline{e}) \end{aligned} \tag{2.7}$$

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: en primer lloc s'ha de determinar el tipus d'orientació que segueix cada parell de seqüències de paraules, per a posteriorment obtindre estadístiques que revelen amb quina freqüència un parell de seqüències de paraules concret ha seguit un determinat tipus de reordenament, és a dir:

$$p(orientació \mid \overline{f}, \overline{e}) = \frac{N(orientació, \overline{f}, \overline{e})}{\sum_o N(o, \overline{f}, \overline{e})} \tag{2.8}$$

on $N(o, \overline{f}, \overline{e})$ és el nombre de vegades que s'ha vist al corpus el parell de seqüències de paraules $(\overline{f}, \overline{e})$ seguint una orientació o respecte al parell de seqüències anterior. En la pràctica, per evitar la sobreestimació del model, es suavitza el nombre d'ocurrències dels diferents esdeveniments amb la probabilitat *a priori* de l'orientació, estimada de forma similar a la probabilitat condicionada, comptant esdeveniments i normalitzant:

$$p(orientació) = \frac{\sum_{\overline{f}} \sum_{\overline{e}} N(orientació, \overline{f}, \overline{e})}{\sum_o \sum_{\overline{f}} \sum_{\overline{e}} N(o, \overline{f}, \overline{e})} \tag{2.9}$$

i per tant, l'estimació suavitzada del model de l'Equació, (2.8) amb un valor apropiat del factor σ, es realitza d'aquesta forma:

$$p(orientació \mid \overline{f}, \overline{e}) = \frac{\sigma\, p(orientació) + N(orientació, \overline{f}, \overline{e})}{\sigma + \sum_o N(o, \overline{f}, \overline{e})} \tag{2.10}$$

Existeixen moltes variants d'aquest model de reordenament: per exemple, podem considerar el tipus de reordenament existent no sols respecte a la seqüència de paraules anterior, sinó també respecte a la seqüència de paraules posterior. Per obtindre informació més detallada sobre aquestes variants i ampliar allò exposat a aquesta secció, remetem al lector a [Koe10].

2.1.3 El model logarítmic-lineal de Moses

Com hem vist, a la Secció 2.1.1 s'han introduït els models log-lineals, instanciant-los per al cas de la TA. Posteriorment, a la Secció 2.1.2 s'han exposat les característiques que formen part del model log-lineal de *Moses*, però de forma aïllada. Resta per tant integrar aquestes característiques en el model log-lineal de Moses, que sumen un total de 14, sis de les quals pertanyen al model de reordenament lexicalitzat.

- Associades al model de traducció:

  - Model de traducció de seqüències de paraules directe i invers: $p(\overline{e}_k \mid \overline{f}_k)$ i $p(\overline{f}_k \mid \overline{e}_k)$.
  - Model de suavitzat lèxic directe i invers: $lex(\overline{e}_k \mid \overline{f}_k)$ i $lex(\overline{f}_k \mid \overline{e}_k)$.
  - 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 (configuració *msd-bidirectional-fe*):
  - Models d'orientació (*m*,*s*,*d*) respecte al parell de seqüències de paraules anterior: $p(m, p \mid \overline{f}_k, \overline{e}_k)$, $p(s, p \mid \overline{f}_k, \overline{e}_k)$, i $p(d, p \mid \overline{f}_k, \overline{e}_k)$
  - Models d'orientació (*m*,*s*,*d*) respecte al parell de seqüències de paraules posterior: $p(m, n \mid \overline{f}_k, \overline{e}_k)$, $p(s, n \mid \overline{f}_k, \overline{e}_k)$, i $p(d, n \mid \overline{f}_k, \overline{e}_k)$
- Model de llenguatge $p(e)$.

2.1.4 Entrenament del model

En la present secció descriurem el procés d'entrenament d'un sistema *Moses*, dividit en 8 etapes. En primer lloc cal resoldre la dependència dels models de *Moses* amb l'alineament entre paraules de les frases d'entrenament (passos 1, 2 i 3) mitjançant el programari *GIZA++* que és una implementació lliure dels models d'IBM. Una vegada estimats els alineaments, el sistema pot obtindre els models de traducció lèxica (pas 4), de traducció de seqüències de paraules (passos 5 i 6), i el model de reordenament (pas 7). Per últim, es genera un fitxer que arreplega la configuració del sistema.

1. **Preparar dades per al GIZA++ toolkit**: Es processa el corpus paral·lel d'entrenament per a poder ser utilitzat amb el programari *GIZA++*.
2. **Executar GIZA++**: Obté els alineaments més probables entre les paraules de cada parell de frases del corpus d'entrenament en ambdues direccions de traducció d'acord amb els models d'IBM (veure Secció 1.3.1).
3. **Obtindre alineaments bidireccionals**: S'aplica un mètode heurístic per combinar els alineaments d'ambdós direccions de traducció, obtenint els alineaments finals necessaris per a la extracció de seqüències de paraules a partir del corpus.
4. **Generar taula de traducció lèxica**: S'estimen, a partir dels alineaments de paraules obtinguts als pas 3, els models de traducció lèxica $lex(f \mid e)$ i $lex(e \mid f)$ (veure Secció 2.1.2).
5. **Extraure seqüències de paraules**: S'obtenen tots els parells de seqüències de paraules consistents amb els alineaments de paraules (veure Secció 1.3.2) obtinguts al final del pas 3, els quals es desen a un fitxer comprimit anomenat

extract.gz. Aquest fitxer presenta, línia a línia, tots els parells de seqüències de paraules extrets del corpus d'entrenament. Les línies d'aquest fitxer tenen un aspecte similar a aquest:

```
resumption ||| reanudación ||| 0-0
resumption of the ||| reanudación del ||| 0-0 1-1 2-1
resumption of the session ||| reanudación del período de sesiones ||| 1-1 2-1 3-2 3-4
of the ||| del ||| 0-0 1-0
of the session ||| del período de sesiones ||| 0-0 1-0 2-1 2-3
session ||| período de sesiones ||| 0-0 0-2
```

Les seqüències de paraules apareixen delimitades pel símbol |||, de forma que la primera seqüència de paraules representa a la seqüència en el llenguatge origen $\overline{f}$, mentre que la segona seqüència de paraules representa a la seqüència en el llenguatge destí $\overline{e}$. Per últim, l'últim camp delimitat per ||| mostra els alineaments existents entre les paraules que conformen ambdues seqüències.

6. **Generar taula de seqüències de paraules**: A partir del fitxer *extract.gz* es genera la taula de seqüències de paraules[7], que és la implementació física de les característiques del model log-lineal associades al model de traducció de *Moses* (veure Secció 2.1.3). En aquesta taula apareixen tots els parells de seqüències de paraules observats al fitxer *extract.gz* sense repeticions, acompanyats de les probabilitats associades a les característiques del model de traducció, per al cas particular del parell de seqüències de paraules considerat. Les línies d'aquesta taula presenten un aspecte com aquest:

```
a gradual resumption of ||| una reanudación paulatina de ||| 1 0.0407673 1 0.011889 2.718
a gradual resumption ||| una reanudación paulatina ||| 1 0.123856 1 0.0183073 2.718
a legal presumption of ||| una presunción judicial de ||| 1 0.0173598 1 0.00322516 2.718
a legal presumption ||| una presunción judicial ||| 1 0.0527413 1 0.00496625 2.718
a presumption against ||| una presunción en contra de ||| 1 0.0811048 0.5 0.00329944 2.718
a presumption against ||| una presunción en contra ||| 1 0.0811048 0.5 0.0152981 2.718
```

Podem observar tres camps delimitats pel separador |||. En primer lloc, trobem la seqüència de paraules en el llenguatge origen $\overline{f}$, en segon lloc, la seqüència de paraules en el llenguatge destí $\overline{e}$, i en tercer i darrer lloc, les probabilitats, separades per espais en blanc, de les característiques associades al model de traducció. En l'exemple trobem, d'esquerra a dreta, $p(\overline{f} \mid \overline{e})$, $lex(\overline{f} \mid \overline{e})$, $p(\overline{e} \mid \overline{f})$, $lex(\overline{e} \mid \overline{f})$ i ρ, que en el nostre cas sempre és $\exp(1) = 2.718$.

7. **Construir model de reordenament**: Es genera un fitxer comprimit que conté la informació relacionada amb el model de reordenament estimat d'acord amb la configuració elegida (veure Secció 2.1.2). En el nostre cas emprarem la configuració per defecte (*msd-bidirectional-fe*).

[7] *Phrase Table*, en anglès.

8. **Crear arxiu de configuració**: Per finalitzar el procés d'entrenament del sistema, es crea un fitxer de configuració que arreplega tota la informació necessària per fer funcionar el sistema: pesos associats a les característiques del model log-lineal, rutes on es troben la taula de seqüències de paraules, el model de reordenament, etc.

D'aquest procés cal tenir en compte dos aspectes importants: en primer lloc, els pesos assignats a les característiques del model log-lineal després d'entrenar el model prenen uns valors predefinits (típicament 0.1 ó 0.2 per a la majoria de característiques), els quals s'han d'ajustar per balancejar les aportacions de cadascuna d'aquestes característiques al model global, en termes de millorar les prestacions finals del sistema, tot d'acord amb les característiques del corpus (aquest procediment és explicat breument a la Secció 2.1.7). En segon lloc, cal construir un model de llenguatge amb alguna ferramenta externa, com per exemple la ferramenta SRILM [Sto02], la qual permet construir un model d'n-grames a partir de la part monolingüe del corpus paral·lel emprat per entrenar el sistema, si bé podria emprar-se un corpus monolingüe independent.

2.1.5 Procés de traducció

La tasca de traducció o descodificació és el procés en el que el sistema genera un gran nombre de possibles traduccions de la frase d'entrada, elegint aquella que maximitza la probabilitat o puntuació conferida pel model log-lineal:

$$\hat{e} = \underset{e}{\operatorname{argmax}} \left[\sum_{k=1}^{K} \sum_{i=1}^{N} \lambda_i h_i(f, e, \overline{f}_k, \overline{e}_k) \right] \tag{2.11}$$

instanciat amb les característiques exposades a la Secció 2.1.3. No obstant, aquesta tasca és molt complexa, tant que existeix un nombre exponencial de possibles traduccions respecte a la longitud de la frase d'entrada. De fet s'ha demostrat que el problema de la cerca de la traducció més probable és NP-Complet [Kni99a], amb la qual cosa, l'exploració de tot l'espai de cerca en busca de la traducció més probable és una tasca tant costosa com impossible.

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 1.8). Com veiem, existeix un gran nombre d'alternatives (segmentació, traducció i reordenament) a considerar, o el que és el mateix, un gran espai de cerca, fet que demostra clarament la complexitat de trobar la traducció més probable.

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*. Els algorismes heurístics utilitzen una funció de puntuació $f(n)$, que serveix per valorar el prometedor que és un estat o hipòtesi n, i que es defineix com el cost per arribar des de l'estat inicial fins l'estat n, denotat per $g(n)$, més una estimació heurística del cost restant per trobar la solució (traducció) òptima, denotada per $h(n)$. És a dir, de forma matemàtica: $f(n) = g(n) + h(n)$. Cal notar que, en aquest context, a menor cost, major probabilitat associada a la possible traducció, i viceversa.

L'algorisme heurístic emprat al sistema *Moses* presenta les següents característiques:

- **No Complet**: L'algorisme no pot garantir trobar la traducció d'una frase d'entrada encara que el model siga capaç de proveir-la, ja que durant el procés de cerca s'apliquen tècniques de poda (explicades a continuació) que poden eliminar els estats que condueixen a alguna solució factible (traducció completa).

- **Admissible**: La funció d'estimació del cost restant per arribar a la traducció òptima no sobreestima el cost restant real per arribar a la mateixa, de forma que es compleix que:

$$\begin{aligned} F(n) &\geq g(n) + h(n) \quad \forall n \\ F(n) &= g(n) + h(n) \quad n \in S \end{aligned}$$

  on $F(n)$ representa el cost real d'arribar a l'estat o hipòtesi n, i S representa al conjunt traduccions completes (estats solució) de la frase d'entrada.

- **Estratègia de cerca per amplària**: S'expandeixen totes les hipòtesis de nivell i abans d'expandir hipòtesis de nivell $i+1$. El nivell i-èsim engloba totes aquelles hipòtesis generades després d'aplicar i opcions de traducció.

Per tal d'acotar l'espai de cerca i, en conseqüència, reduir la complexitat computacional del 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 menor funció de cost $g(n)$.

- **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. Així doncs, aquesta limitació es tradueix en una acotació de l'espai de cerca, degut a que en el procés d'expansió d'hipòtesis es descarta l'aplicació d'opcions de traducció que excedirien el límit de distorsió.

- **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 amb menor valor per a la funció de puntuació $f(n)$ són descartades. Cal notar que, a pesar d'emprar una funció d'informació heurística $h(n)$ admissible, 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, o pitjor encara, de podar tots els estats que poden conduir a un estat solució. No obstant, la probabilitat de que això succeïsca és molt baixa, i en aquests casos l'algorisme de cerca proporciona una traducció molt propera a l'òptima (en termes de cost). Cal tenir en compte que, a menor capacitat d'aquestes piles, major acotament de l'espai de cerca 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 [Koe10].

2.1.6 Avaluació de la qualitat de la traducció

Amb el sistema entrenat i preparat per descodificar qualsevol frase d'un vocabulari origen $\mathcal{F}$, ens trobem amb la necessitat de mesurar les prestacions del mateix, en termes d'avaluar la qualitat de les traduccions que ofereix. En aquest sentit, la forma ideal d'avaluar un sistema de TA seria l'avaluació humana: l'eixida proporcionada pel sistema seria analitzada per lingüistes experts, els quals determinarien el nivell de correcció de la frase (en aspectes sintàctics, gramaticals, semàntics, etc.). Això seria el que anomenem avaluació subjectiva. Ara bé, aquest tipus d'avaluació presenta una sèrie d'inconvenients:

- **Procés exhaustiu**: Per determinar si un sistema presenta bones prestacions o no, no és suficient amb analitzar la traducció d'unes poques frases d'entrada. Estem parlant de considerar grans quantitats de frases, de l'ordre de milers, doncs sols d'aquesta forma podríem estimar de forma de robusta les prestacions reals del sistema. Avaluar de forma manual la traducció de tal quantitat de frases pot ocupar períodes de temps exageradament grans, i possiblement no seria viable.

- **Pèrdua de la noció d'automatització**: Es perd l'automatització completa del procés de construcció d'un sistema de TA, ja que la fase d'avaluació es realitza de forma manual, retardant la posada en marxa del sistema.

- **Dificultat de comparació amb altres sistemes**: Degut a la component purament subjectiva introduïda a l'hora d'avaluar els sistemes, dos sistemes diferents sols es podrien comparar si han estat analitzats pel mateix grup d'experts, en el mateix moment i baix les mateixes circumstàncies, cosa que és quasi impossible de garantir.

Per tots aquests motius, l'avaluació dels sistemes de TA es realitza també de forma automàtica. L'avaluació automàtica es realitza comparant l'eixida del sistema (hipòtesi) amb la referència associada a la frase d'entrada, valorant la qualitat de la traducció mitjançant alguna mètrica existent. El problema que trobem en l'avaluació automatitzada és que, en general, sols es proporciona una única referència per realitzar l'anàlisi comparatiu, quan en realitat una frase pot tenir més d'una traducció completament vàlida. No obstant això, la rapidesa en que es realitza aquest anàlisi així com la facilitat en que es pot comparar un sistema amb altres ha propiciat que la comunitat investigadora es decante per l'avaluació automàtica envers l'avaluació subjectiva.

Algunes de les mètriques emprades per avaluar de forma automàtica les prestacions d'aquests sistemes són les següents:

- **WER** (*Word Error Rate*) [ON03]: Aquesta fou la primera mètrica d'avaluació automàtica, adoptada directament dels mètodes d'avaluació dels sistemes de reconeixement de la parla. Aquesta mètrica 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*) [SDS+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*) [PRWZ01]: La taxa BLEU, una de les més populars 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.

Existeixen altres mètriques com PER [TVN+97], o NIST [Dod02]. Per a l'avaluació dels nostres sistemes emprarem la mètrica BLEU, donada la gran popularitat que té en aquesta disciplina.

2.1.7 Ajustament de paràmetres

Com hem comentat al final de la Secció 2.1.4, els pesos inicials associats a les característiques del model log-lineal de *Moses* són de molt mala qualitat i requereixen ser ajustats. L'aprenentatge d'aquests pesos es realitza mitjançant l'*entrenament per mínima taxa d'error*, de l'anglès *Minimum Error Rate Training* [Och03], 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'entrenament (ni del sistema *Moses* ni del model 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 no es modifiquen en absolut), o bé s'arriba a un nombre màxim d'iteracions. La forma de realitzar l'exploració en l'espai de possibles valors per als paràmetres és una tasca molt complexa. Existeixen algorismes com el *Simplex Algorithm* o *Powell Search* que resolen aquest problema [Koe10].

2.2 Mancances del model de seqüències de paraules

A la Secció 2.1.3 hem exposat el model complet de *Moses*, basat en un model log-lineal que integra una sèrie de característiques exposades a la Secció 2.1.2. Una cosa que es troba a faltar a aquest model és la inclusió de més informació sobre la longitud de les frases, i en especial, de les seqüències de paraules. El model log-lineal de *Moses* tant sols incorpora una característica relacionada amb la longitud, que és el *factor de penalització per seqüències de paraules*, però la seva aportació és insuficient baix aquest propòsit, doncs no pren en compte exactament la longitud de les seqüències de paraules, sinó més bé el nombre de seqüències de paraules en que es segmenta la frase d'entrada. La idea és modelar les distribucions de probabilitat que segueixen les longituds de les seqüències de paraules del corpus d'entrenament, ja que aquesta informació podria ser de gran utilitat en el procés de traducció.

De fet, en [AFJ09, AF10] es proposa un model de traducció monòton de seqüències de paraules purament estadístic basat en semi-models ocults de Markov, en el que l'autor anomena *Phrase-Based Semi-Hidden Markov Models* (PBSHMM), on es modela de forma explícita la forma en que es segmenten (en seqüències de paraules) les frases d'entrada i d'eixida mitjançant dos vectors l i m respectivament. Aquests vectors són dues variables ocultes al model (donat que el corpus no està etiquetat amb les segmentacions més idònies, com passa de forma anàloga amb els alineaments de paraules), motiu pel qual l'estimació del model es realitza amb l'algorisme *Expectation-Maximization* [DLR77].

El model de traducció invers proposat en [AFJ09, AF10] es defineix com una exploració de totes les possibles segmentacions, definides per l i m:

$$p(f \mid e, J) = \sum_m \sum_l p(f, l, m \mid e, J) \tag{2.12}$$

on f és la frase d'entrada, e és la frase d'eixida, i J és la longitud de la frase d'entrada. El model incomplet queda a expenses, doncs, del model complet $p(f, l, m \mid e, J)$ amb l i m com variables ocultes, el qual es descompon, per la regla de la cadena i d'esquerra a dreta:

$$p(f, l, m \mid e, J) = p(m \mid e, J)\, p(l \mid m, e, J)\, p(f \mid l, m, e, J) \tag{2.13}$$

Donada la dificultat d'estimar cadascun dels termes en que es descompon el model complet, l'autor realitza una sèrie d'assumpcions per simplificar-lo:

$$p(m \mid e, J) := \prod_t p(m_t) \tag{2.14}$$

$$p(l \mid m, e, J) := \prod_t p(l_t \mid m_t) \tag{2.15}$$

$$p(f \mid l, m, e, J) := \prod_t p(f(t) \mid e(t)) \tag{2.16}$$

on t pren com a valors les posicions de m i l on es denota l'inici d'una seqüència de paraules, de forma que que $f(t)$ i $e(t)$ representen la t-èsima seqüència de paraules d'f i de e respectivament, i l_t i m_t les seves respectives longituds. Així, cada submodel ve donat per una exploració de tota seqüencia de paraules t explicada per les variables de segmentació l i m.

Baix aquestes assumpcions, el model de traducció complet simplificat es defineix com segueix:

$$p(f, l, m \mid e, J) := \prod_t p(m_t)\, p(l_t \mid m_t)\, p(f(t) \mid e(t)) \tag{2.17}$$

Per la seva banda, el model incomplet quedaria definit d'aquesta forma:

$$p(f \mid e, J) := \sum_m \sum_l p(f, l, m \mid e, J) = \sum_m \sum_l \prod_t p(m_t)\, p(l_t \mid m_t)\, p(f(t) \mid e(t)) \tag{2.18}$$

Per mesurar les prestacions d'aquest model, l'autor realitza un estudi experimental en el que es comparen, per a un mateix corpus d'entrenament, validació i test, tres sistemes diferents: un primer lloc, un sistema base *Moses* com el presentat a la Secció 2.1.3, a excepció que no inclou les característiques del model de reordenament lexicalitzat; en segon lloc, el sistema anterior però reemplaçant ambdós models de traducció

de seqüències de paraules pels models proposats a l'Equació (2.18); i en darrer lloc, un sistema com el primer però afegint les versions directa i inversa del model proposat com dues característiques addicionals al model log-lineal. Els resultats mostren com els dos sistemes que incorporen aquest nou model milloren el sistema base, encara que no de forma significativa en termes estadístics [AF10].

Així doncs, vist que aquest model pot donar peu a una millora de les prestacions, tractarem d'adaptar-lo i integrar-lo al nostre sistema. Cal tenir en compte que el model proposat per l'autor està ideat en un context de traducció monòtona, i no pas en el context de *Moses*, en el que s'efectuen de freqüentment operacions de reordenament dels segments de la frase origen al ser traduïts. Aquesta limitació es pot evitar si efectuem un reordenament inicial de la frase origen, per a posteriorment realitzar el procés de traducció de forma monòtona amb un sistema PBSHMM. Aleshores, sorgeix la necessitat d'introduir al model original una component de distorsió o reordenament inicial de la frase origen (que degut a l'aplicació de la regla de Bayes és la frase e del llenguatge d'eixida), que permeta condicionar el model PBSHMM a la frase d'entrada ja reordenada. Per tant, si manipulem l'Equació (2.12) amb aquest propòsit, tenim que:

$$p(f \mid e, J) = \sum_{\tilde{e}} \sum_{m} \sum_{l} p(f, \tilde{e}, l, m \mid e, J) \tag{2.19}$$

on havem inclòs la variable $\tilde{e}$ que representa la frase e reordenada d'alguna forma d'entre totes les permutacions possibles. El nou model, al que considerem una extensió del model PBHSMM proposat originalment, el definim com segueix:

$$p(f, \tilde{e}, l, m \mid e, J) := p(\tilde{e} \mid e, J)\; p(f, l, m \mid \tilde{e}, J) \tag{2.20}$$

on $p(\tilde{e} \mid e, J)$ ve donat per un model de distorsió o reordenament, el qual ens indica la probabilitat de permutar e donant lloc a $\tilde{e}$, mentre que $p(f, l, m \mid \tilde{e}, J)$ és el model PBSHMM (veure Equació (2.17)), però instanciat per a la frase permutada $\tilde{e}$. Anem a assumir, en primer lloc, que el model de distorsió és uniforme i proporcional al model de reordenament bàsic $d(|\text{inici}_k - \text{fi}_{k-1} - 1|)$ que incorpora el sistema *Moses* (presentat a la Secció 1.3.2), amb la qual cosa podrem prescindir d'ell; i en segon lloc, que $\overline{e}_t$ és la t-èsima seqüència de paraules en que és segmentada la frase reordenada $\tilde{e}$. Llavors, si desenvolupem el model PBSHMM definit a l'Equació (2.17) i l'integrem amb la definició del model incomplet, tenim que:

$$p(f \mid e, J) := \sum_{\tilde{e}} \sum_{m} \sum_{l} \prod_{t} p(m_t)\; p(l_t \mid m_t)\; p(\overline{f}_t \mid \overline{e}_t) \tag{2.21}$$

Donat que utilitzem la notació simplificada dels vectors de segmentació m i l, on m_t i l_t representen les longituds de les t-èsimes seqüències de paraules origen i destí, podem emprar en el seu lloc la longitud de les seqüències de paraules, ja que representen el mateix concepte, i per tant, podem prescindir dels vectors m i l al nostre model. A

més, com t explora les K segmentacions de la frase d'entrada, podem fer un canvi d'índex per homogeneïtzar la notació amb el nostre discurs. Aleshores, reescriurem l'Equació (2.21) de la següent manera:

$$p(f \mid e, J) := \sum_{\tilde{e}} \prod_{k} p(|\overline{e}_k|)\; p(|\overline{f}_k| \;/\; |\overline{e}_k|)\; p(\overline{f}_k \mid \overline{e}_k) \tag{2.22}$$

Per últim, assumirem que el nostre model no depèn de cap reordenament inicial de la frase origen, ja que en *Moses* cada segment $\overline{e}_k$, proporcionat per un algorisme de cerca heurístic, és reordenat durant el procés de traducció. Per tant:

$$p(f \mid e, J) := \prod_{k} p(|\overline{e}_k|)\; p(|\overline{f}_k| \;/\; |\overline{e}_k|)\; p(\overline{f}_k \mid \overline{e}_k) \tag{2.23}$$

Aquesta equació defineix el que al Capítol 3 presentarem com el model de longitud estàndard.

D'aquest model que acabem de presentar podem considerar una variant en la que assumim que la probabilitat de segmentació de la frase d'eixida m_t donada la frase d'eixida e i la longitud de la frase d'entrada J és uniforme respecte al total de segmentacions M que es poden realitzar a la frase d'eixida e. Llavors, l'assumpció explicitada en l'Equació (2.14) es convertiria en aquesta:

$$p(m \mid e, J) := \frac{1}{M} \tag{2.24}$$

i per tant, la variant del model de longitud estàndard que únicament considera el model de longitud condicional quedaria expressat de la següent forma:

$$p(f \mid e, J) := \prod_{k} p(|\overline{f}_k| \;/\; |\overline{e}_k|)\; p(\overline{f}_k \mid \overline{e}_k) \tag{2.25}$$

D'altra banda, l'autor del citat treball proposa una possible variant d'aquest model en la que es pot assumir que la longitud de la t-èsima seqüència de paraules origen no sols depèn de la longitud de la t-èsima seqüència de paraules destí (veure Equació (2.15)), sinó també de la t-èsima seqüència de paraules destí:

$$p(l \mid m, e, J) := \prod_{t} p(l_t \mid m_t, e(t)) \tag{2.26}$$

Seguint aquesta nova assumpció i aplicant de nou les mateixes transformacions, el model $p(f \mid e, J)$ quedaria reescrit de la següent forma:

$$p(f \mid e, J) := \prod_{k} p(|\overline{e}_k|)\; p(|\overline{f}_k| \;/\; \overline{e}_k)\; p(\overline{f}_k \mid \overline{e}_k) \tag{2.27}$$

Cal notar que s'ha prescindit de la longitud de la t-èsima seqüència de paraules destí

m_t (o $|\overline{e}_k|$), ja que és redundant si es coneix $y(t)$ (o $\overline{e}_k$). Aquesta variant del model és la que al Capítol 3 presentarem com el model de longitud especialitzat.

De la mateixa forma que al model de longitud estàndard, podem obtindre una variant del model especialitzat considerant que la probabilitat de la segmentació de la frase origen $p(f \mid e, J)$ és uniforme, de forma que el model quedaria reescrit de la següent forma:

$$p(f \mid e, J) := \prod_k p(|\overline{f}_k| \;/\; \overline{e}_k)\; p(\overline{f}_k \mid \overline{e}_k) \tag{2.28}$$

En definitiva, tant l'Equació (2.23) com l'Equació (2.27) ens mostren com modelar la longitud dels segments juntament amb un model de traducció invers de seqüències de paraules. Al Capítol 3 veurem com implementar i integrar aquests models al sistema *Moses*.

Capítol 3

Models de Longitud

En el capítol anterior hem presentat els sistemes log-lineals, els quals ens brinden la possibilitat d'afegir fonts d'informació addicionals al procés de traducció. En aquest context, a la Secció 2.2 hem discutit la possibilitat d'incloure informació sobre la longitud de les seqüències de paraules al model, donades les bones perspectives que es desprenen del treball mencionat. Aquesta possibilitat es converteix en realitat en aquest Capítol, on proposarem la inclusió del modelat de les longituds de les seqüències de paraules al model de traducció de *Moses*.

3.1 Model de longitud estàndard

El primer dels dos models de longitud que hem considerat a aquest treball és aquell que modela les relacions existents entre les longituds de les seqüències de paraules d'entrada i d'eixida a partir d'un corpus d'entrenament. D'acord amb l'Equació (2.23), el model de traducció invers $p(f \mid e)$ ve donat per tres termes (que en realitat podríem considerar-ne únicament dos, donat que $p(|\overline{e}|)\ p(|\overline{f}| \ / \ |\overline{e}|) = p(|\overline{f}|, |\overline{e}|)$):

- Un **model de traducció de seqüències de paraules** $p(\overline{f} \mid \overline{e})$ (veure Secció 1.3.2).

- Un **model de longitud incondicional** $p(|\overline{e}|)$, que representa la probabilitat de trobar una seqüència de paraules del llenguatge destí de longitud $|\overline{e}|$. Aquest model és capaç de respondre a preguntes com la següent: *quina és la probabilitat d'observar una seqüència de 5 paraules de longitud en anglès?*

- Un **model de longitud condicional** $p(|\overline{f}| \ / \ |\overline{e}|)$, que representa la probabilitat de traduir una seqüència de paraules del llenguatge destí de longitud $|\overline{e}|$ en una seqüència de paraules del llenguatge origen de longitud $|\overline{f}|$. Aquest model podria respondre, per exemple, la següent pregunta: *quina és la probabilitat de traduir una seqüència (qualsevol) de 6 paraules de longitud en anglès, en una seqüència (qualsevol) de 4 paraules de longitud en català?*

Cal notar que, si invertim la direcció de la traducció, obtindrem models de longitud anàlegs: $p(|\overline{f}|)$ i $p(|\overline{e}| \mid |\overline{f}|)$, en addició al model de traducció de seqüències de paraules invers $p(\overline{e} \mid \overline{f})$.

3.1.1 Estimació del model i implementació

El model de traducció de seqüències de paraules s'estima conforme a l'esmentat a la Secció 1.3.2, mentre que que els models relacionats amb la longitud s'estimen per màxima versemblança respecte a un conjunt de seqüències de paraules $(\overline{f}, \overline{e})$ extretes d'un corpus d'entrenament de parells de frases $\{(f_n, e_n) \in C : n = 1, \ldots N\}$.

D'una banda, el model de longitud incondicional s'estima de la següent forma:

$$p(|\overline{e}|) = \frac{N(|\overline{e}|)}{N} \tag{3.1}$$

on $N(|\overline{e}|)$ és el nombre de vegades que s'han observat seqüències de paraules de longitud $|\overline{e}|$ en el llenguatge d'eixida, i N és el nombre total de seqüències de paraules observades. No obstant, per evitar una sobreestimació del model es realitza un suavitzat mitjançant la interpolació lineal entre el model de longitud i la distribució uniforme de la longitud, amb un valor adequat d'ϵ:

$$p_\epsilon(|\overline{e}|) = (1-\epsilon)\frac{N(|\overline{e}|)}{N} + \epsilon\,\frac{1}{L} \tag{3.2}$$

on L és la longitud màxima que poden assolir les seqüències de paraules (recordar que la longitud està acotada per termes d'eficiència, veure Secció 1.3.2). L'estimació de $p(|\overline{f}|)$ es realitza de forma similar.

D'altra banda, l'estimació del model de longitud condicional s'efectua com segueix:

$$p(|\overline{f}| \mid |\overline{e}|) = \frac{N(|\overline{f}|, |\overline{e}|)}{N(|\overline{e}|)} \tag{3.3}$$

on $N(|\overline{f}|, |\overline{e}|)$ és el nombre de vegades que s'han observat de forma conjunta seqüències de paraules de longitud $|\overline{e}|$ en el llenguatge d'eixida i seqüències de paraules de longitud $|\overline{f}|$ en el llenguatge d'entrada. L'estimació de $p(|\overline{e}| \mid |\overline{f}|)$ també es realitza de forma similar. De nou, aquesta distribució de probabilitat és suavitzada per evitar la seva sobreestimació:

$$p_\epsilon(|\overline{f}| \mid |\overline{e}|) = (1-\epsilon)\frac{N(|\overline{f}|, |\overline{e}|)}{N(|\overline{e}|)} + \epsilon\,\frac{1}{L} \tag{3.4}$$

Fins al moment hem vist com estimar aquests models, però no d'on extraure la informació necessària per estimar-los. El procés descrit no és més que una simplificació del

procés d'estimació dels models de longitud, doncs resulta complex estimar-lo directament per màxima versemblança a partir del corpus d'entrenament (s'hauria d'aplicar un entrenament basat en l'algorisme *Expectation-Maximization* [DLR77], ja que no disposem de les segmentacions de les frases). En el context de *Moses*, trobem dues maneres d'obtindre seqüències de paraules a partir del corpus d'entrenament: d'una banda, a partir de les seqüències de paraules extretes de forma consistent respecte a un heurístic d'alineaments de paraules (veure Secció 1.3.2), i d'altra banda, a partir de les segmentacions de Viterbi proveïdes pel sistema en un procés de traducció restringit (dirigit) a obtindre la traducció correcta. Ambdues alternatives es basen en l'ús de tècniques heurístiques, tal i com hem vist a les Seccions 1.3.2 i 2.1.5 respectivament, amb la qual cosa l'estimació dels models no serà estrictament correcta. Anem a detallar cadascuna de les vies proposades.

Estimació a partir de seqüències de paraules extretes en la fase d'entrenament

Com ja sabem, el sistema *Moses*, en lloc d'obtindre les seqüències de paraules a partir de les segmentacions més probables estimades per màxima versemblança a partir del corpus d'entrenament, aplica un algorisme que extrau, per a cada parell de frases del corpus, totes les seqüències de paraules consistents amb un heurístic dels alineaments entre paraules estimats prèviament mitjançant una implementació dels models d'IBM. Això es tradueix en que, per cada parell de frases del corpus, existeixen moltes segmentacions possibles que donen lloc a múltiples parells de seqüències de paraules, la majoria de les quals no tenen per què explicar la forma més idònia de traduir la frase origen en la frase destí, però tot i això, formen part dels paràmetres del model de traducció de seqüències de paraules, i per extensió, multipliquen el nombre d'esdeveniments a observar en la estimació del model de longitud. A pesar de la incorrecció que presenta aquesta tècnica des d'un punt de vista teòric, la primera de les dues fonts d'informació que considerarem per estimar els nostres models és aquest conjunt de seqüències de paraules extretes de forma heurística.

Com hem vist a la Secció 2.1.4, els parells de seqüències de paraules extretes amb aquest heurístic es desen, línia per línia, al fitxer *extract.gz*. Cal doncs processar cada línia d'aquest fitxer, calcular la longitud de les seqüències de paraules i incrementar els comptes dels esdeveniments observats. Una vegada s'ha processat tot el fitxer, es normalitzen els comptes i es suavitzen les distribucions de probabilitat resultants conforme a les Equacions (3.2) i (3.4).

Estimació a partir de les segmentacions de Viterbi

Com ja s'ha esmentat adés, la forma més natural i estadísticament correcta d'estimar els models que ací proposem és l'entrenament per màxima versemblança a partir del corpus d'entrenament de parells de frases. Mitjançant aquest mètode considerem que el nostre model és un model paramètric, els paràmetres del qual són denotats per θ i estimats a partir de les dades del corpus aplicant una funció de versemblança L, que en el nostre cas es defineix així:

$$L(\theta) = \prod_{n=1}^{N} p_\theta(f_n \mid e_n, J_n) \tag{3.5}$$

on p_θ és el model de traducció paramètric i N és el nombre de mostres del corpus. Ja que el nostre criteri és maximitzar la versemblança respecte el corpus d'entrenament, necessitem trobar els paràmetres òptims $\hat{\theta}$ que maximitzen aquest criteri:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta) \tag{3.6}$$

En aquesta literatura és habitual emprar la funció log-versemblança, atès que la funció log és monòtona creixent i no afecta al còmput de la funció argmax. Aleshores, tenim que:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, (\log L(\theta)) = \underset{\theta}{\operatorname{argmax}} \left(\sum_{n=1}^{N} \log p_\theta(f_n \mid e_n, J_n) \right) \tag{3.7}$$

Ara bé, en realitat el nostre model de traducció paramètric $p_\theta(f \mid e, J)$ és incomplet: no explica la forma de segmentar les frases d'eixida i d'entrada, com tampoc la forma de reordenar els segments de la frase d'eixida per generar, de forma monòtona, la frase d'entrada (recordem que estem modelant el model de traducció invers). Per tant, el model requereix tres variables addicionals: l i m, que denoten respectivament la segmentació de les frases f i e en seqüències de paraules, i r, que representa un reordenament de les seqüències de paraules de la frase d'eixida e explicades per m. Aquestes variables són ocultes al nostre model, doncs el corpus d'entrenament no es troba etiquetat amb tal informació, i per tant necessitem estimar aquesta informació per poder posteriorment estimar el model. Llavors, si destapem les variables ocultes del model de traducció, tenim que:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left(\sum_{n=1}^{N} \log \left(\sum_{l} \sum_{m} \sum_{r} p_\theta(f_n, l, m, r \mid e_n, J_n) \right) \right) \tag{3.8}$$

Així, no és possible estimar el model aplicant una maximització del criteri de la log-versemblança, doncs no tenim informació sobre les variables ocultes l, m i r. Hauríem de recórrer a l'aplicació d'un entrenament *Expectation-Maximization* [DLR77], ja que sols aquesta tècnica ens permet estimar el nostre model amb variables ocultes per màxima versemblança. No obstant, aquest tipus d'estimació és molt costosa, així que optarem per realitzar una aproximació heurística d'aquest procés. La idea és que la traducció més probable d'una frase, d'acord amb el model log-lineal de *Moses*, és aquella que explica la forma més òptima de segmentar i reordenar la frase d'eixida per generar de forma monòtona les seqüències de paraules que conformen la frase d'entrada. En altres paraules, la traducció més probable f d'una frase d'eixida e proporciona una aproximació heurística dels valors òptims de les variables ocultes del nostre model:

$$\hat{l}, \hat{m}, \hat{r} = \underset{l,m,r}{\operatorname{argmax}}\, p(f, l, m, r \mid e, J) \tag{3.9}$$

Les segmentacions definides per $\hat{l}$ i $\hat{m}$ són aquelles que expliquen la generació de la traducció més probable segons el model log-lineal de *Moses*, i reben el nom de segmentacions de Viterbi. Per tant, podem aconseguir extraure parells de seqüències de paraules del corpus d'entrenament d'una forma més precisa emprant el descodificador de *Moses*: cal proporcionar-li tant el conjunt de frases d'entrenament per a ser traduïdes com les seves respectives referències, orientant el procés de descodificació a la producció de les referències proporcionades. La generació per part del descodificador de la traducció de referència (la millor traducció possible) permet obtindre les segmentacions de les frases d'eixida i d'entrada i les relacions o associacions entre les seqüències de paraules resultants, que és precisament una aproximació a la informació que proporcionen les variables ocultes del model l, m i r, respectivament. Cal recordar que el procés de traducció es basa en una tècnica de cerca heurística que tracta de trobar més probable segons el model log-lineal (veure Secció 2.1.5).

Aleshores, per poder realitzar aquest procés es requereix un sistema base *Moses* entrenat i optimitzat. En primer lloc, cal entrenar el sistema amb un conjunt d'entrenament (veure Secció 2.1.4). Posteriorment, és aconsellable ajustar els pesos de les característiques del model log-lineal amb un conjunt de validació (veure Secció 2.1.7), deixant així el sistema enllestit per abordar el següent pas, que és traduir el conjunt d'entrenament, obligant al descodificador a obtindre les traduccions de referència, de forma que el sistema proporcionarà les segmentacions de les frases d'entrada i d'eixida que millor expliquen el procés de traducció d'acord amb els paràmetres del model. Llavors, a partir d'aquesta informació es poden estimar els models de longitud conforme a les Equacions (3.2) i (3.4).

Un exemple d'eixida proveïda pel descodificador, al traduir de l'anglès a l'espanyol el corpus *Europarl-v3* (veure Secció 4.1), és la següent:

aprobación |0-0| del |1-2| acta |3-3|

Aquesta eixida indica, per exemple, que la seqüència de paraules *del* es troba alineada amb la seqüència de paraules determinada pel rang de posicions de paraules en la frase origen 1-2 (les paraules en les posicions 1 i 2), que consultant la referència és *of the*. En realitat no ens preocupa saber quina és la seqüència de paraules relacionada amb la seqüència origen, sinó la seva longitud, que en el cas de l'exemple es pot calcular de forma trivial: $(2 - 1) + 1 = 2$. Hem implementat un programa que permet processar aquesta eixida de forma adient per obtindre la informació necessària per estimar els models d'una forma similar a la realitzada a partir del fitxer *extract.gz*.

Cal destacar un aspecte important relacionat amb el procés de traducció guiat per les referències: el sistema no sempre serà capaç de generar com a traducció més probable la referència proporcionada, i en aquests casos, el descodificador no mostrarà

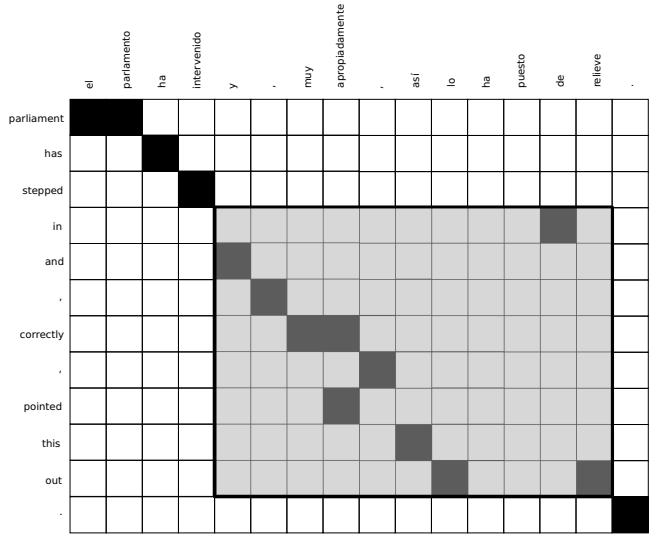


Figura 3.1: Exemple d'alineament entre paraules d'un parell de frases, extret del corpus Europarl-v3 per al parell de llenguatges anglès - espanyol (veure Secció 4.1). Aquest alineament requereix, en el procés de traducció, l'extracció d'una seqüència d'11 paraules de longitud com a mínim (denotada per la franja grisa), segons l'algorisme d'extracció de seqüències de paraules de *Moses*.

cap eixida: simplement ignorarà la frase processada. Això pot ocórrer principalment per dos motius: bé pel problema derivat de la incompletitud de l'algorisme heurístic de cerca de la traducció més probable (generalment motivat pel límit de distorsió en la cerca, veure Secció 2.1.5), o bé degut a que poden haver-hi seqüències de paraules que no es poden extraure de forma consistent de cap de les maneres respecte a l'alineament entre paraules proporcionat, generalment a causa de que la longitud màxima a la que estan limitades les seqüències de paraules impedeix abastar una seqüència de paraules de suficient longitud que respecte la propietat de consistència amb els alineaments. És per això que aquest efecte negatiu és inversament proporcional a la longitud màxima de les seqüències de paraules: a menor longitud màxima, major és el nombre de frases que no es poden traduir forçadament en les seves referències, i viceversa.

A la figura 3.1 podem observar un exemple d'alineament[1] entre paraules d'un parell de frases extret del corpus Europarl-v3 per al parell de llenguatges anglès - espanyol (veure Secció 4.1). Si ens fixem, aquest alineament no permetria obtindre una segmentació vàlida d'ambdós frases si considerarem seqüències de paraules amb una longitud màxima limitada, per exemple, a 7 paraules, doncs es requereix l'extracció d'un parell de seqüències de paraules (denotat per la franja grisa) format per una seqüència en espanyol d'almenys 11 paraules de longitud i una seqüència en anglès d'almenys 8 paraules de longitud. Llavors, en aquest cas el descodificador de *Moses*

[1] Cal recordar que aquests alineaments són estadístics, és dir, no s'estableixen a partir de coneixements lingüístics.

no seria capaç d'obtindre la traducció de referència si el model s'entrena limitant la longitud màxima de les seqüències de paraules a menys d'11 paraules.

En definitiva, aquesta forma d'obtindre la informació necessària per estimar els models presenta un gran avantatge i un gran inconvenient. D'una banda, considerar sols les seqüències de paraules que millor expliquen les traduccions correctes (referències), o en altres paraules, les segmentacions més probables, implica augmentar la precisió del model de longitud. Però d'altra banda, el fet que el descodificador no puga, en certes ocasions, obtindre com a traducció més probable la referència proporcionada, a part que és un mostra de la deficiència del model de traducció de *Moses* (veure Secció 1.2.4), això provoca que el nombre d'events disponibles per estimar els models disminuïsca críticament, amb la qual cosa el sistema serà menys confiable i robust. No obstant, en el cas del model de longitud estàndard, al considerar únicament esdeveniments de tipus longitud de seqüències de paraules, l'impacte negatiu no és tant significatiu com en el cas del model de longitud especialitzat, com comprovarem més endavant.

3.1.2 Integració en Moses

La inclusió d'aquests models al sistema es realitza modificant el model log-lineal de *Moses* (veure Secció 2.1.3). Es consideren tres variants del sistema original:

- **Moses-LConjunta**: Substitueix les característiques de traducció de seqüències de paraules directa i inversa pels models de traducció directe i invers proposats a l'Equació (2.23).

- **Moses-LCondicional**: Aquest sistema és similar a l'anterior, però implementa la variant del model de longitud estàndard presentada a l'Equació (2.25) (no es modela la probabilitat de longitud incondicional). Aquest nou sistema permetrà conferir major expressivitat al model de longitud condicional, de forma que, si la informació que aporta és realment útil, aleshores major efecte benigne tindrà sobre el procés de traducció, doncs considerar el model conjunt implica que els models de traducció de seqüències de paraules directe i invers aporten la mateixa funció de probabilitat de longitud: $p(|\overline{e}|)\, p(|\overline{f}| \;/\; |\overline{e}|) = p(|\overline{f}|)\, p(|\overline{e}| \;/\; |\overline{f}|) = p(|\overline{f}|, |\overline{e}|)$.

- **Moses + LCondicional**: Aquest sistema inclou les característiques del sistema base *Moses* sense modificar, més dues característiques addicionals: les probabilitats condicionades de les longituds en ambdues direccions de la traducció, és a dir, els models $p(|\overline{f}| \;/\; |\overline{e}|)$ i $p(|\overline{e}| \;/\; |\overline{f}|)$. Aquest sistema ens permetrà avaluar d'una forma més transparent l'aportació real del model de longitud al procés de traducció, donat que es tracta d'una característica (en realitat dues) addicional i independent. A més, cal destacar que amb aquest sistema ens evitem pertorbar el model de traducció de seqüències de paraules de *Moses*.

Com ja havem vist a la Secció 2.1.4, les característiques que formen part del model de traducció de seqüències de paraules s'integren a la taula de seqüències de paraules,

en l'apartat de puntuacions. Hem desenvolupat una ferramenta que processa íntegrament aquesta taula, de forma que, per a cada parell de seqüències de paraules $\overline{f}$ i $\overline{e}$ es calculen les seves respectives longituds $|\overline{f}|$ i $|\overline{e}|$ i es modifiquen les probabilitats de les característiques en funció del parell de seqüències de paraules considerat i del sistema que s'està construint (*Moses-LConjunta, Moses-LCondicional, Moses + LCondicional*). Per exemple, al construir el sistema *Moses-LCondicional*, per a tot parell de seqüències de paraules es modificaran les puntuacions $p(\overline{f} \mid \overline{e})$ i $p(\overline{e} \mid \overline{f})$ per $p(|\overline{f}| \ / \ |\overline{e}|)\, p(\overline{f} \mid \overline{e})$ i $p(|\overline{e}| \ / \ |\overline{f}|)\, p(\overline{e} \mid \overline{f}))$ respectivament.

Al Capítol 4 s'analitzaran les prestacions d'aquests tres sistemes en comparació amb el sistema *Moses* convencional.

3.2 Model de longitud especialitzat

L'altre model de longitud que proposem a aquest capítol, originat com una variant del primer, és el que anomenem model de longitud especialitzat. Respecte al model de longitud estàndard, i d'acord amb la definició formal del model especialitzat (veure Equació (2.27)), l'únic terme que canvia és el model de longitud condicional, que en aquest cas és condicionat a una seqüència de paraules: aquest model l'anomenarem model de longitud condicionat a seqüències de paraules $p(|\overline{f}| \ / \ \overline{e})$, que representa la probabilitat de traduir una seqüència de paraules $\overline{e}$ concreta del llenguatge destí en una seqüència de paraules del llenguatge origen de longitud $|\overline{f}|$. Aquest model permet respondre a preguntes com aquesta: *quina és la probabilitat de traduir la seqüència de paraules en anglès "La meva mare és universitària" en una seqüència (qualsevol) de 2 paraules de longitud en català?*

Es pot comprovar com la naturalesa d'aquest model és la que motiva la nomenclatura del mateix (model de longitud especialitzat), doncs aporta informació sobre la longitud dels segments en que es pot traduir cada seqüència de paraules concreta.

Si invertim la direcció de la traducció, obtindrem models de longitud anàlegs: $p(|\overline{f}|)$ i $p(|\overline{e}| \ / \ \overline{f})$, en addició al model de traducció de seqüències de paraules directe $p(\overline{e} \mid \overline{f})$.

3.2.1 Estimació del model i implementació

L'estimació d'aquests models es realitza, a l'igual que els models de longitud estàndard, per màxima versemblança respecte a un conjunt de seqüències de paraules $(\overline{f}, \overline{e})$ extretes d'un corpus d'entrenament de parells de frases $\{(f_n, e_n) \in C : n = 1, \ldots N\}$ i consistents amb uns alineaments entre paraules per a cada parell de frases.

D'una banda, el model de longitud incondicional s'estima tal i com s'indica a l'Equació (3.2). D'altra banda, l'estimació del model de longitud condicionat a seqüències de paraules es realitza de forma similar al model de longitud condicionat del model estàndard:

$$p(|\overline{f}| \ / \ \overline{e}) = \frac{N(|\overline{f}|, \overline{e})}{N(\overline{e})} \tag{3.10}$$

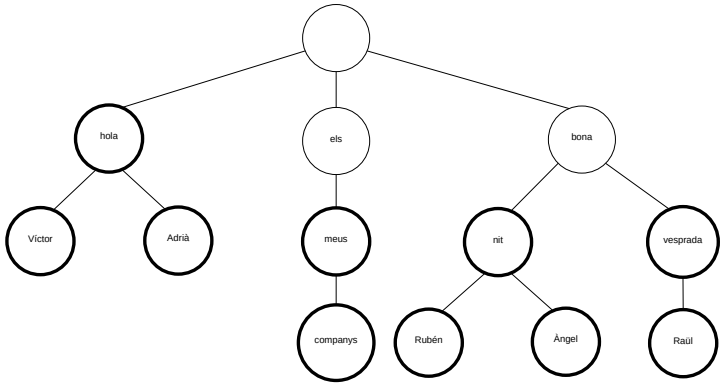


Figura 3.2: Exemple d'una estructura de dades trie.

on $N(|\overline{f}|, \overline{e})$ és el nombre de vegades que s'han observat conjuntament seqüències de paraules de longitud $|\overline{f}|$ en el llenguatge d'entrada i la seqüència de paraules $\overline{e}$ en el llenguatge d'eixida. Novament suavitzarem aquest model per evitar problemes de sobreestimació, mitjançant una interpolació lineal amb el model de longitud condicional estàndard amb un valor d'ϵ adequat:

$$p_\epsilon(|\overline{f}| \ / \ \overline{e}) = (1-\epsilon)\frac{N(|\overline{f}|, \overline{e})}{N(\overline{e})} + \epsilon \, \frac{N(|\overline{f}|, |\overline{e}|)}{N(|\overline{e}|)} \tag{3.11}$$

Cal notar que l'estimació de $p(|\overline{e}| \ / \ \overline{f})$ es realitza de forma similar.

Ara bé, estimar el model condicional presenta una complexitat afegida, doncs per a la part dreta del model (la condició) no considerem longituds de segments (valors enters), sinó seqüències de paraules (cadenes de text de longitud variable). L'especialització del model de longitud implica una major complexitat de l'algorisme d'estimació, però sobretot, de l'estructura de dades encarregada d'emmagatzemar, per a totes les seqüències de paraules $\overline{e}$, els comptadors $N(|\overline{f}|, \overline{e})$, necessaris per poder estimar dit model (veure Equació (3.11)). Per poder satisfer aquest requeriment de la forma més eficient possible, s'ha implementat una estructura de dades anomenada *trie*, una estructura en forma d'arbre molt apropiada per emmagatzemar grans quantitats de dades que presenten la propietat de prefix entre elles, com és el cas particular de les seqüències de paraules (recordar la forma en que s'extrauen, veure Secció 1.3.2). Emprar una estructura jeràrquica d'aquest tipus en lloc de cap altra estructura de dades permet reduir la complexitat espacial del model, ja no s'emmagatzemen prefixos repetits, i per extensió, també redueix la complexitat temporal de les operacions elementals a realitzar sobre l'arbre, ja que existiran menys elements a processar.

Podem observar un exemple d'un *trie* a la Figura 3.2. En una estructura de dades *trie* adaptada i implementada conforme a les nostres necessitats, cada node representa una única cadena de text (una paraula, signe de puntuació, etc.). L'arbre de prefixos

comença amb un node principal, que no representa cap cadena (en realitat representa el prefix buit), a partir del qual pengen tots els nodes. Cada camí que parteix del node principal i arriba fins un node qualsevol representa el prefix d'una seqüència de paraules observada (seguint l'exemple, la seqüència *els meus* és un prefix d'alguna seqüència de paraules observada). Les seqüències de paraules observades es poden identificar a l'arbre com tot camí que parteix del node principal que arriba a un node no nul (representat a l'exemple pels nodes de contorn més gros). La diferència principal entre un node nul i un node no nul és que aquest darrer representa una seqüència de paraules observada, determinada pel camí que parteix del node principal i arriba al node considerat, de forma que aquests nodes alberguen els comptadors per a les seqüències de paraules que representen. Dit d'altra forma, un node no nul que representa la seqüència de paraules $\overline{e}$ manté els comptadors associats a eixa seqüència $N(|\overline{f}|, \overline{e})$ i $N(\overline{e})$.

Així doncs, a l'exemple de la Figura 3.2 observem que els nodes ressaltats representen un node no nul (*hola, bon*, etc.), mentre que la resta representen nodes nuls (*els*, node inicial). Aquest *trie* denota que en l'entrenament del model s'han observat les seqüències de paraules *hola, hola Víctor, hola Adrià, els meus, els meus companys, bona nit, bona nit Rubén, bona nit Àngel, bona vesprada, bona vesprada Raül*. Cal notar que en cap cas s'ha observat la seqüència *els* o la seqüència *bona*, donat que els nodes que les representen són nuls.

La reducció de la complexitat espacial en comparació amb altres estructures de dades es posa de manifest en l'exemple de la Figura 3.2, doncs, per exemple, a l'haver observat les seqüències de paraules *bona nit*, *bona nit Àngel*, i *bona nit Rubén*, no ha estat necessari emmagatzemar tres seqüències de paraules de forma independent, o el que és el mateix, 8 paraules diferents $(2 + 3 + 3)$: amb emmagatzemar 4 paraules (en forma de nodes) i establir les correspondències adients al *trie* n'és suficient. En aquest cas, explotant la propietat de prefix de les seqüències de paraules hem aconseguit reduït la informació a emmagatzemar a la meitat.

Cal notar a més que, en termes d'eficiència, en lloc de representar els nodes com una cadena de text, és preferible etiquetar-los amb un identificador numèric per reduir a cost constant la tasca de comparació d'etiquetes de nodes. Això s'aconsegueix emprant un diccionari o taula *hash* que assigna a cada paraula diferent un identificador numèric únic.

Donat el context en que s'empra aquesta estructura de dades, tant sols és necessari implementar les operacions de cerca i d'inserció a l'arbre. Per cada seqüència de paraules observada a la font d'informació, en primer lloc s'ha de consultar el contingut de l'arbre per comprovar si aquesta s'havia observat en anterioritat o no. Si ja s'havia observat, és a dir, existeix a l'arbre el node no nul que representa la seqüència recentment observada, aleshores s'incrementa el comptador d'observacions adequat. Si, per contra, la seqüència de paraules no s'havia observat en anterioritat, aleshores s'insereixen els nodes necessaris per construir el camí que condueix al node (no nul) que representa la frase observada (que òbviament també s'ha d'inserir), i per últim s'incrementa el comptador d'esdeveniments adient. Cal tenir en compte, respecte a

l'operació d'inserció, que és possible que el camí que representa a una seqüència de paraules no observada en anterioritat ja existisca: en eixe cas la seqüència recentment observada era un prefix d'una seqüència de paraules de major longitud observada anteriorment. En tal cas, l'única operació necessària a realitzar és convertir el node nul que representa la seqüència observada en un node no nul, i posteriorment actualitzar els comptadors. Per entendre-ho millor, si reprenem l'exemple de la Figura 3.2, i imaginem que en primer lloc s'observa la seqüència de paraules *els meus companys*, aleshores a l'arbre es crearien dos nodes nuls, *els* i *meus*, i un node no nul que representa la seqüència observada, el node *companys*. Quan, posteriorment, s'observe la seqüència de paraules *els meus*, aleshores només caldrà "transformar" el node *meus* en un node no nul, i incrementar els comptadors pertinents.

Per últim, estudiarem la complexitat temporal de les operacions de cerca i d'inserció en l'arbre, proporcionant els costs associats al pitjor cas (O) i millor cas (Ω) que es contemplen en aquestes operacions.

D'una banda, l'operació de cerca requereix explorar tants nivells de l'arbre com paraules conté la seqüència de paraules, explorant-se, en el pitjor dels casos, L nivells, sent L la longitud màxima que poden assolir les seqüències de paraules. En cada nivell s'exploren els nodes descendents del node predecessor fins trobar el node buscat: en el pitjor dels casos poden arribar a explorar-se tants nodes com paraules té el vocabulari de l'idioma considerat (grandària del vocabulari, $|\mathcal{V}|$). Per tant, la complexitat temporal de l'operació de cerca en un *trie* en el pitjor dels casos és $O(L\,|\mathcal{V}|)$. El millor dels casos, per la seva part, es donarà quan cerquem una seqüència de paraules de longitud 1, i el node a buscar es troba en la primera (o primeres) posicions del contenidor dels nodes del primer nivell (que en la pràctica és un vector), amb un cost $\Omega(1)$.

D'altra banda, l'operació d'inserció requereix, en el pitjor dels casos, explorar L nivells del *trie* i inserir un node al nivell L (és a dir, afegir una seqüència de paraules de longitud L, no observada anteriorment, en un *trie* en el que ja existeix una seqüència de paraules formada per les $L-1$ primeres paraules de la seqüència observada), amb un cost de $O(L\,|\mathcal{V}|)$, ja que la inserció d'un node aïllat té cost constant. Per la seva part, el millor dels casos ve representat per la inserció al *trie* d'una seqüència de paraules de longitud 1 en un *trie* buit, sense cap node existent al primer nivell, amb un cost $\Omega(1)$. Ara bé, donat que aquesta és una instància que sols es pot esdevindre en una ocasió, resulta més realista acotar inferiorment el cost de la inserció amb $\Omega(|\mathcal{V}|)$, ja abans d'inserir un nou node al primer nivell cal assegurar-se que aquest es troba al mateix nivell, fet que requereix l'exploració de tots els nodes.

Determinar la complexitat temporal dels millors i pitjors casos de les operacions elementals a realitzar amb la nostra estructura de dades ens dona una idea del comportament d'aquesta estructura de dades en casos extrems. Ara bé, és preferible realitzar un anàlisi més precís que ens permeta calcular el cost mitjà d'aquestes operacions, per poder així introduir la notació Θ. Per fer-ho, hauríem d'explorar tota possible seqüència de paraules $\overline{x}$, obtindre la seva probabilitat d'ocurrència $p(\overline{x})$, i calcular el seu cost $c(\overline{x})$ associat a la operació a realitzar:

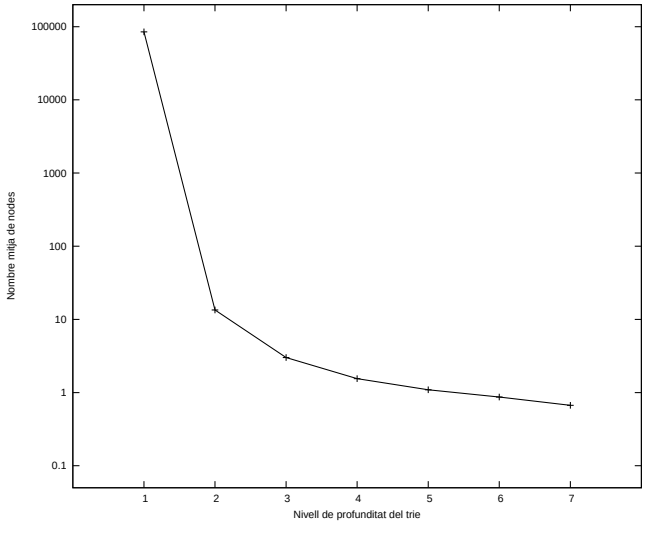


Figura 3.3: Distribució del nombre de nodes emmagatzemats a cada nivell de *trie*. S'aprecia, en escala logarítmica, el nombre mitjà de nodes residents a cada nivell d'un *trie* construït a partir de les seqüències de paraules en espanyol extretes de forma heurística a partir del conjunt d'entrenament del corpus Europarl-v3 (veure Secció 4.1), considerant seqüències de paraules limitades a 7 paraules de longitud. S'observa una clara tendència exponencial inversa del nombre mitjà de nodes conforme s'aprofundeix en l'arbre.

$$\text{cost\_mitjà} = \sum_{\overline{x}} p(\overline{x})\, c(\overline{x}) \tag{3.12}$$

Com es pot imaginar, estimar el cost mitjà de les operacions considerant tota possible seqüència de paraules seria una tasca molt costosa, així que per simplicitat aproximarem el cost mitjà de les operacions elementals considerant les longituds de les seqüències de paraules (és a dir, càlcul de probabilitats i costs associats a cerca / inserció de seqüències de paraules de longitud determinada). Aquesta aproximació al cost mitjà es defineix així:

$$\text{cost\_mitjà} \approx \sum_{l=1}^{L} p(l)\, c(l) \tag{3.13}$$

on l representa longituds de seqüències de paraules. Per estimar el cost associat a cadascuna de les operacions per a tota possible longitud de les seqüències de paraules, hem estudiat com es distribueixen les dades (nodes) en aquesta estructura, construint un *trie* a partir de les seqüències de paraules en espanyol extretes de forma heurística a partir del conjunt d'entrenament del corpus Europarl-v3, per al parell de llenguatges anglès - espanyol (veure Secció 4.1), considerant seqüències de paraules limitades a 7

paraules de longitud. El propòsit d'aquest estudi és mesurar el nombre mitjà de nodes que s'haurien de processar a cada nivell de l'arbre per realitzar una operació de cerca o d'inserció. Per exemple, reprenent el *trie* mostrat a la Figura 3.2, la cerca de la seqüència de paraules *bona nit Àngel* requeriria explorar 3 nodes al primer nivell (els nodes descendents del nivell zero, on figura el node principal), 2 nodes al segon nivell (descendents del node *bona*), i altres 2 nodes al tercer nivell (descendents del node *nit*); però en canvi, la cerca de la seqüència *els meus companys* requereix explorar 3, 1 i 1 nodes a cadascun dels tres nivells, respectivament. Es tracta doncs de determinar quin és el nombre mitjà de nodes a explorar a cada nivell, per fer-nos una idea de quin pot ser el cost mitjà de les operacions de cerca i inserció. Els resultats d'aquest estudi es mostren a la gràfica de la Figura 3.3, on es pot apreciar, en escala logarítmica, el nombre mitjà de nodes a explorar per a cada nivell del *trie* (cal notar que al *trie* hi hauran tants nivells com la màxima longitud de les seqüències de paraules, en aquest cas, 7). S'observa una clara evolució exponencial inversa del nombre mitjà de nodes conforme s'aprofundeix en l'arbre: per exemple, si al primer nivell es podrien arribar a processar vora 90000 nodes, al segon nivell se'n processarien, en mitjana, uns 13 nodes. A la vista d'aquests resultats, podem interpretar el nombre mitjà de nodes presents al primer nivell com una aproximació a la grandària del vocabulari $\mathcal{V}$ del llenguatge sobre el que es construeix el *trie*, és a dir, $|\mathcal{V}|$ (i de fet es tracta d'un valor molt pròxim a la grandària del vocabulari espanyol del conjunt d'entrenament del corpus, consultar Taula 4.1); mentre que el nombre mitjà de nodes a explorar a partir del segon nivell és una aproximació al logaritme de la grandària del vocabulari, és a dir, $\log|\mathcal{V}|$.

Així, ja que partirem d'una observació empírica, podem expressar els costs associats a les operacions elementals $c(l)$ com el cost mitjà real $\Theta(l)$. Considerant que l'operació d'inserció requereix processar en mitjana els mateixos nodes que l'operació de cerca, i que la inserció efectiva d'un node duu associat un cost constant, podem realitzar un raonament comú per esbrinar el cost mitjà d'ambdós operacions. D'acord amb la distribució mitjana de nodes a cada nivell del *trie* (veure Figura 3.3), el nombre de nodes a processar al primer nivell és molt superior al nombre de nodes a processar en la resta de nivells. Per tant, podem diferenciar dos casos: d'una banda, el cas en que la longitud de les seqüències de paraules és igual a 1, amb un cost mitjà $\Theta(|\mathcal{V}|)$; i d'altra banda, el cas en que la longitud és major que 1, amb un cost mitjà $\Theta(l\log|\mathcal{V}|)$. Si instanciem l'Equació (3.13) diferenciant aquests dos casos, tindrem que:

$$\text{cost\_mitjà} \approx\ p(1)\ \Theta(|\mathcal{V}|) + \sum_{l=2}^{L} p(l)\ \Theta(l\log|\mathcal{V}|) \tag{3.14}$$

Sols ens falta conèixer la distribució de la longitud de les seqüències de paraules, que podem estimar-la tal i com hem exposat a la Secció 3.1.1, Equació (3.1). A la Figura 3.4 es mostra una gràfica que representa aquesta distribució, estimada a partir de la longitud de les seqüències de paraules en espanyol extretes de forma heurística a partir del conjunt d'entrenament del corpus Europarl-v3 (veure Secció 4.1), considerant seqüències de paraules limitades a 7 paraules de longitud. D'aquesta gràfica s'infereix

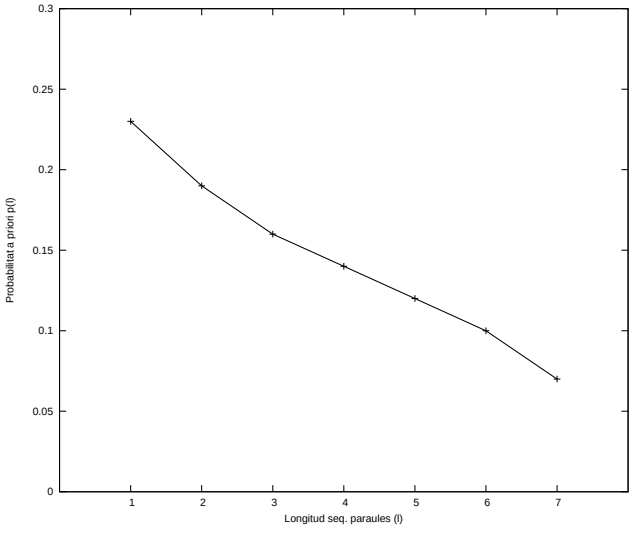


Figura 3.4: Distribució del la longitud de les seqüències de paraules, estimada a partir de les seqüències de paraules en espanyol extretes de forma heurística a partir del conjunt d'entrenament del corpus Europarl-v3 (veure Secció 4.1), considerant seqüències de paraules limitades a 7 paraules de longitud.

que el terme dominant és el cas $l = 1$, així que podem prescindir de l'altre terme (cas $l > 1$) i assumir que el cost mitjà de les operacions de cerca i inserció és $\Theta(|\mathcal{V}|)$.

Com havem vist, el punt crític d'aquesta estructura de dades és el primer nivell, ja que alberga una quantitat de nodes de l'ordre de $|\mathcal{V}|$. D'acord amb aquesta realitat, podem optimitzar aquesta estructura de dades per tal d'accelerar les operacions de cerca i inserció en aquest nivell. La idea és implementar el primer nivell del *trie* (format pels descendents del node principal) com un vector en el que els nodes es troben indexats pel seu identificador numèric, de forma que trobar la primera paraula d'una seqüència de paraules és converteix en una operació de cost constant. Aquesta optimització permet reduir de forma significativa el cost temporal de l'operació de cerca des d'un punt de vista pràctic, però no asimptòtic. El cost associat al pitjor cas de les dues operacions elementals passaria a ser $O((L-1)\,|\mathcal{V}|) = O(L\,|\mathcal{V}| - |\mathcal{V}|) = O(L\,|\mathcal{V}|)$. Per la seva part, el cost associat el millor cas de l'operació de cerca continuaria sent $\Omega(1)$, però amb un petit (i important) matís: aquest cost és aplicable per a tota instància de longitud 1, mentre que al *trie* sense optimitzar, com havem vist, sols unes poques instàncies concretes de longitud 1 donen lloc a un cost constant. Respecte a l'operació d'inserció sí que s'observa una millora del cost asimptòtic inferior, que passa a ser un cost constant, aplicable també a totes les instàncies de longitud 1: $\Omega(1)$.

Si discutim sobre el cost mitjà d'aquesta versió optimitzada del *trie* original, seguint un raonament similar al que hem realitzat anteriorment, determinaríem que:

	cerca			inserció		
trie original	$\Omega(1)$	$O(L\|\mathcal{V}\|)$	$\Theta(\|\mathcal{V}\|)$	$\Omega(\|\mathcal{V}\|)$	$O(L\|\mathcal{V}\|)$	$\Theta(\|\mathcal{V}\|)$
trie optimitzat	$\Omega(1)$	$O(L\|\mathcal{V}\|)$	$\Theta(\log\|\mathcal{V}\|)$	$\Omega(1)$	$O(L\|\mathcal{V}\|)$	$\Theta(\log\|\mathcal{V}\|)$

Taula 3.1: Anàlisi de la complexitat temporal de les operacions de cerca i inserció al *trie* original i al *trie* optimitzat.

$$\text{cost\_mitjà} \approx p(1)\ \Theta(1) + \sum_{l=2}^{L} p(l)\ \Theta(l \log|\mathcal{V}|) \tag{3.15}$$

Novament ens trobaríem amb que el terme $p(1)$ és el dominant, fet que conferiria un cost mitjà constant a les operacions de cerca i inserció, però això no és gens realista (es pot comprovar empíricament). En realitat, com el primer terme (cas $l = 1$) es redueix a una constant, podem despreciar-lo i assumir que el cost mitjà d'aquestes operacions ve donat pel segon terme (cas $l > 1$), i llavors: $\Theta(l \log|\mathcal{V}|) = \Theta(\log|\mathcal{V}|)$ (prescindim d'l ja que és una constant). Aleshores, dissenyar aquesta versió optimitzada del *trie* ens ha permés reduir de forma significativa el cost mitjà de les operacions de cerca i inserció, passant d'un cost lineal a un cost logarítmic. A la Taula 3.1 s'arrepleguen les cotes superior, inferior, i cost mitjà de les operacions de cerca i inserció, tant per al *trie* original com per a l'optimitzat.

Una vegada presentada l'estructura de dades que ens permetrà implementar els comptadors necessaris per estimar els models, detallarem la forma d'obtindre la informació necessària per realitzar aquesta estimació. De nou, ens trobem amb dues alternatives: d'una banda, a partir de les seqüències de paraules extretes de forma consistent respecte als alineaments de paraules donats (veure Secció 1.3.2), i d'altra banda, a partir de les segmentacions de Viterbi proveïdes pel sistema en un procés de traducció restringit (dirigit) a obtindre la traducció correcta. Tot i que la forma de processar la informació en ambdós casos és molt similar a la detallada en el model de longitud estàndard (veure Secció 3.1.1), hi ha petites diferències que exposem tot seguit per a cada cas.

Estimació a partir de seqüències de paraules extretes en la fase d'entrenament

Com ja sabem, una possible font d'informació per estimar els models és el conjunt de seqüències de paraules extretes de forma heurística a partir del corpus d'entrenament, les quals són emmagatzemades al fitxer *extract.gz*. Cal processar el fitxer línia per línia, buscant / afegint al *trie* la seqüència de paraules observada i incrementant comptadors d'events. Una vegada s'ha processat tot el fitxer, es normalitzen els comptes emmagatzemats al *trie* (requereix un processat de tot l'arbre) i es suavitza la massa de probabilitat calculada conforme a les Equacions (3.2) i (3.11).

Estimació a partir de les segmentacions de Viterbi

Una altra font d'informació més precisa però menys robusta són les segmentacions de Viterbi proveïdes pel sistema al traduir el conjunt d'entrenament de forma guiada per les referències de traducció. L'estimació es realitza de forma molt similar a la vista en el cas del model de longitud estàndard, però cal tenir en compte una sèrie de consideracions particulars de l'estructura de dades *trie*.

Com ja sabem, el principal inconvenient que presenta aquesta font d'informació és que el nombre d'esdeveniments a observar es redueix de forma considerable respecte a l'altra font d'informació considerada (seqüències de paraules extretes durant l'entrenament del sistema), fet que pot provocar no sols problemes de sobreestimació dels models (atenuats gràcies a l'aplicació de tècniques de suavitzat), sinó també problemes relacionats amb la no observació de seqüències de paraules que sí s'han observat en l'entrenament del sistema per mitjà de l'heurístic d'extracció de seqüències de paraules, i que per tant formen part dels paràmetres del model de traducció de seqüències de paraules de *Moses* (de la mateixa forma que apareixen al fitxer *extract.gz* o a la taula de seqüències de paraules). Aleshores el problema real el trobarem al tractar d'integrar aquests models al sistema *Moses*, doncs això requereix processar la taula de seqüències de paraules (fitxer *phrase-table.gz*), en la que poden aparèixer seqüències de paraules que no s'han observat en la construcció del model de longitud especialitzat. Teoria en mà, en aquests casos la probabilitat associada al model de longitud especialitzat deuria ser nul·la, però com que això no és cert (realment és degut a una estimació poc robusta del model), i donat que la distribució de probabilitat condicionada és suavitzada amb la probabilitat del model de longitud estàndard condicional (veure Equació (3.4)), en la pràctica s'assigna íntegrament la informació aportada pel model de longitud estàndard condicional, molt millor estimat.

3.2.2 Integració

A l'igual que en el model de longitud estàndard, integrarem el model de longitud especialitzat al model log-lineal de *Moses* de tres maneres diferents:

- **Moses-SConjunta**: Substitueix els models de traducció de seqüències de paraules directe i invers pels models de traducció directe i invers proposats a l'Equació (2.27).

- **Moses-SCondicional**: A l'igual que en el cas del model de longitud estàndard, aquest sistema implementa la variant del model de longitud especialitzat presentada a l'Equació (2.28).

- **Moses + SCondicional**: Aquest sistema inclou les característiques del sistema base *Moses* sense modificar, més dues característiques addicionals: les probabilitats condicionades de les longituds en ambdós direccions de la traducció, és a dir, els models $p(|\overline{f}| \ / \ \overline{e})$ i $p(|\overline{e}| \ / \ \overline{f})$. De nou, aquest sistema ens permetrà avaluar d'una forma més transparent l'aportació real del model de longitud al procés de traducció.

De forma similar al cas en el que integràvem el model de longitud estàndard a *Moses* (veure Secció 3.1.2), hem desenvolupat un programa que processa íntegrament la taula de segments, línia per línia, de forma que, per a cada parell de seqüències de paraules $(\overline{f},\overline{e})$, es busca al *trie* la seqüència de paraules que condiciona el model condicional, i llavors es modifiquen els valors de les probabilitats de les característiques d'acord amb les seqüències de paraules identificades i el sistema que s'està construint (*Moses-SConjunta, Moses-SCondicional, Moses + SCondicional*).

Al Capítol 4 també s'analitzaran les prestacions d'aquests tres sistemes en comparació amb el sistema base de *Moses*.

Capítol 4

Corpora i Experimentació

Al present capítol avaluarem les prestacions dels sistemes proposats al capítol anterior. En primer lloc, presentarem el corpus que s'ha emprat per dur a terme aquestes proves, i en segon lloc, detallarem els experiments realitzats amb els sistemes proposats i els seus respectius resultats.

4.1 Corpora

Com ja sabem, un sistema de TAE basat en seqüències de paraules com *Moses* requereix, en primer lloc, un corpus d'entrenament de parells de frases per poder entrenar els models que s'integren al model log-lineal; en segon lloc, un conjunt de validació per ajustar de forma adequada els paràmetres del model log-lineal; i en tercer i darrer lloc, un conjunt de test per avaluar les prestacions del sistema. Per tal de cobrir aquestes necessitats i poder dur a terme les experimentacions que proposem a la Secció 4.2, s'ha emprat el corpus Europarl-v3 [Koe05].

Aquest corpus arreplega les transcripcions de les sessions plenàries de l'Europarlament i les seves corresponents traduccions a un total d'11 llenguatges: Anglès, Alemany, Francès, Italià, Holandès, Portuguès, Danès, Suec, Finès, Grec i Espanyol. Per afinitat lingüística s'ha escollit el corpus associat al parell de llenguatges Anglès - Espanyol, en la seva versió 3, que arreplega totes les transcripcions i traduccions entre aquests dos llenguatges generades en les sessions esdevingudes entre abril de 1996 i octubre de 2006. A la taula 4.1 podem observar les estadístiques d'aquest corpus per als tres conjunts proveïts (entrenament, validació i test), abans de ser preprocessats[1]: el nombre total de parells de frases, i per a cadascun dels idiomes, la longitud mitjana en paraules de les frases, la grandària del vocabulari (nombre de paraules úniques), el nombre total de paraules, i la perplexitat del conjunt calculada per a un model de llenguatge de 5-grames, entrenat amb el conjunt d'entrenament i suavitzat amb el mètode de Kneser-Ney modificat.

[1] Cal notar que: M = Mega = 1.000.000 i K = Kilo = 1.000

	C. Entrenament		C. Validació		C. Test	
Llenguatge	An	Es	An	Es	An	Es
Nombre Frases	730740		2000		2000	
Longitud Mitjana (paraules)	20.8	21.5	29.3	30.3	30.0	30.2
Tamany Vocabulari	72.7K	113.9K	6.5K	8.2K	6.5K	8.3K
Total paraules	15.2M	15.7M	58.7K	60.6K	58.0K	60.3K
Perplexitat (5-grames)	-	-	79.6	78.8	78.3	79.8

Taula 4.1: Estadístiques del corpus Europarl-v3.

Aquest corpus requereix un preprocés (veure Secció 1.1.2) per ser utilitzat amb el sistema *Moses*, que consisteix en tres passos:

1. **"Tokenitzar" les frases** (*Tokenize*): Converteix la frase en una cadena de text equivalent en la que les unitats bàsiques (paraules, signes de puntuació, etc.) apareixen separades mitjançant espais en blanc. Cada unitat bàsica s'anomena *token*.

2. **Filtrar frases llargues** (*Filter*): S'eliminen del corpus les frases amb un nombre de tokens major o igual a 40.

3. **Transformar caràcters a minúscules** (*Lowercase*): Els caràcters en majúscula es transformen en minúscula.

Després de preprocessar el corpus, ens trobem en disposició d'entrenar els nostres sistemes i provar les seves prestacions.

4.2 Experimentació

En la present secció detallarem cadascun dels experiments realitzats, destinats a avaluar si les millores proposades del sistema base de *Moses* es tradueixen efectivament en una millora de les prestacions. En primera instància experimentarem amb el sistema base, per dues raons: en primer lloc, perquè les seves prestacions seran la referència per avaluar si els nous sistemes proposats representen una millora significativa de l'estat d'art de la disciplina; i en segon lloc, perquè a partir del sistema entrenat ens és molt fàcil obtindre la informació necessària per estimar els models de longitud proposats (veure Seccions 3.1.1 i 3.2.1). Això té un altre avantatge, i és que podem estalviar-nos l'entrenament dels nous sistemes proposats, ja que en tots els casos partirem del sistema base entrenat, al qual se li modificarà la taula de seqüències de paraules i el fitxer de configuració *moses.ini* (veure Secció 2.1.4).

Tots els sistemes s'han avaluat per a les dues direccions de traducció (an-es i es-en), limitant la longitud màxima de les seqüències de paraules amb valors que oscil·len entre 2 i 7. L'avaluació de resultats es realitza mitjançant la mètrica BLEU (presentada a la Secció 2.1.6) aplicant una tècnica anomenada bootstrapping [Koe04,

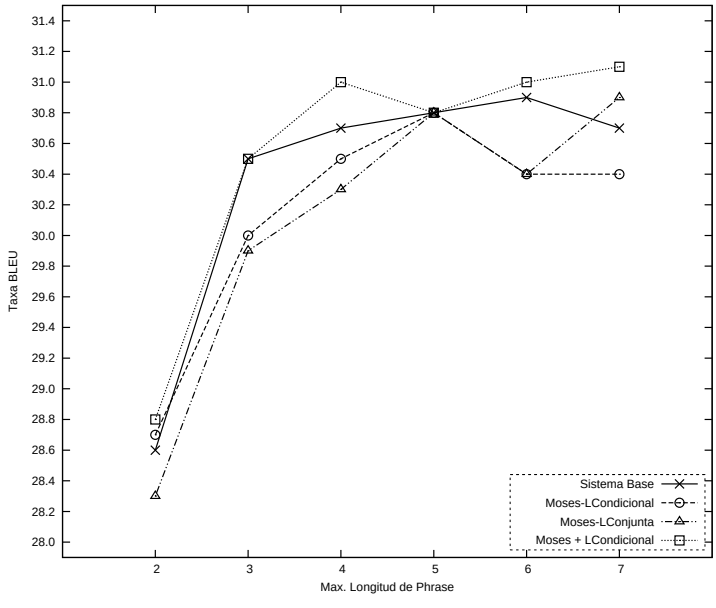


Figura 4.1: Taxa BLEU, en funció de la màxima longitud que poden assolir les seqüències de paraules, del sistema base i dels tres sistemes proposats que implementen el model de longitud estàndard, estimat a partir del conjunt de seqüències de paraules extretes del conjunt d'entrenament del corpus Europarl-v3, per a la direcció de traducció Anglès - Espanyol.

BN04], que ens permet obtindre la taxa BLEU del conjunt d'hipòtesis amb un interval del 95% de confiança. Com a resultats dels experiments proporcionarem la mitjana dels extrems de l'interval, arrodonida a dècimes.

4.2.1 Sistema base

Els experiments amb el sistema base s'han realitzat de la següent forma: per a cada direcció de traducció (an-es i es-an), s'ha entrenat el sistema base limitant la longitud màxima de les seqüències de paraules entre 2 i 7, donant lloc a $2 \times 6 = 12$ sistemes diferents. El model log-lineal del sistema base, com hem descrit a la Secció 2.1.3, presenta les següents característiques:

- Models de traducció de seqüències de paraules directe i invers.
- Models de suavitzat lèxic directe i invers.
- Penalització de seqüències de paraules (amb $\rho = 2.718$).
- Penalització per paraula.
- Model de distorsió uniforme.

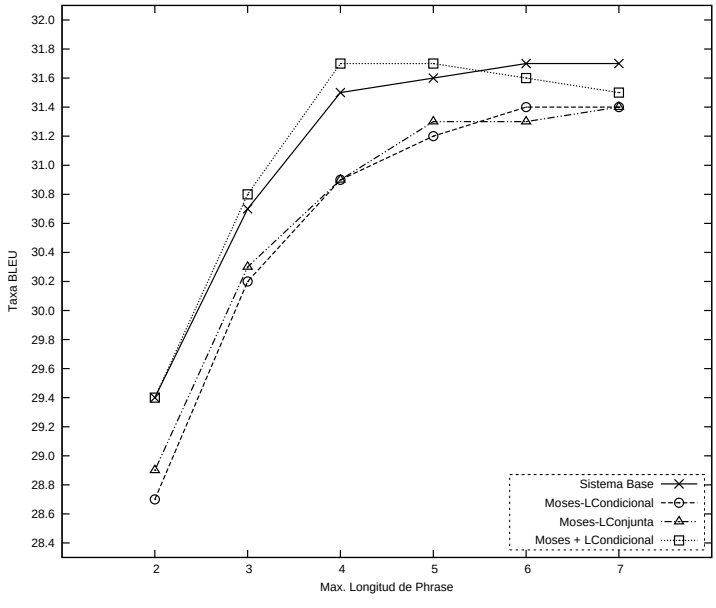


Figura 4.2: Taxa BLEU, en funció de la màxima longitud que poden assolir les seqüències de paraules, del sistema base i dels tres sistemes proposats que implementen el model de longitud estàndard, estimat a partir del conjunt de seqüències de paraules extretes del conjunt d'entrenament del corpus Europarl-v3, per a la direcció de traducció Espanyol - Anglès.

- Model de reordenament (configuració *msd-bidirectional-fe*).

- Model de llenguatge de 5-grames, suavitzat per interpolació lineal amb el descompte modificat de Kneser-Ney.

Per claritat i per evitar redundàncies, els resultats dels experiments amb el sistema base per ambdues direccions de traducció (an-es i es-en) es mostren conjuntament amb els resultats dels experiments realitzats amb els sistemes que implementen ambdós aproximacions del model de longitud. A la Secció 4.2.2 s'exposen els resultats del model de longitud estàndard, mentre que a la Secció 4.2.3 es mostren els resultats del model de longitud especialitzat.

4.2.2 Model de longitud estàndard

En aquesta secció es presenten els resultats obtinguts a l'experimentar amb els sistemes que afegeixen informació del models de longitud estàndard, diferenciant els resultats per a cada via d'estimació dels models i per cada direcció de traducció.

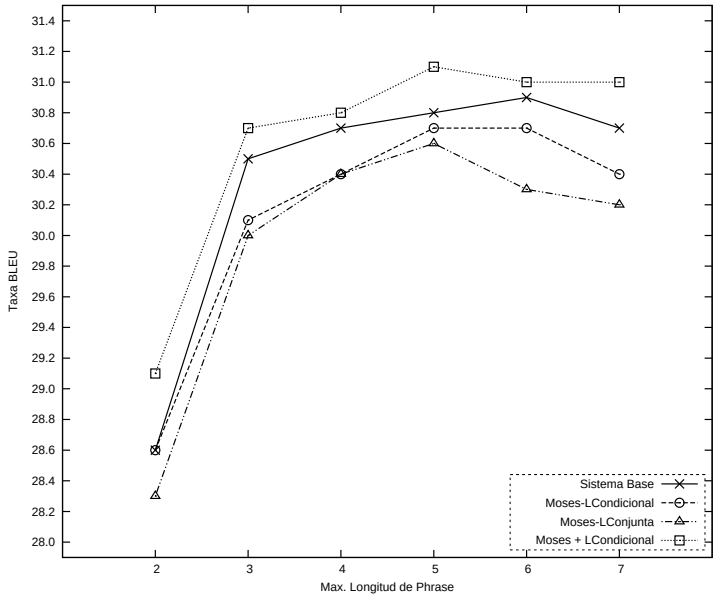


Figura 4.3: Taxa BLEU, en funció de la màxima longitud que poden assolir les seqüències de paraules, del sistema base i dels tres sistemes proposats que implementen el model de longitud estàndard, estimat a partir de les segmentacions de Viterbi extretes del conjunt d'entrenament del corpus Europarl-v3, per a la direcció de traducció Anglès - Espanyol.

Estimació a partir de seqüències de paraules extretes en la fase d'entrenament

En primer lloc considerarem els resultats dels experiments realitzats amb els tres sistemes proposats, estimant els models de longitud a partir del conjunt de seqüències de paraules extretes de corpus.

D'una banda, a la Figura 4.1 podem observar els resultats per a la direcció de traducció Anglès - Espanyol. Com es pot apreciar, els sistemes *Moses-LCondicional* i *Moses-LConjunta* presenten en general prestacions inferiors al sistema base, mentre que el sistema *Moses + LCondicional* presenta un comportament similar al del sistema base, però aportant certes millores puntuals, com s'observa als resultats d'aquest sistema entrenat amb una longitud de seqüències de paraules limitada a 4 i 7, on es millora el sistema base en 3 i 4 dècimes de BLEU, respectivament. Crida especialment l'atenció el punt coincident per a tots els sistemes entrenats limitant la longitud de les seqüències de paraules a un màxim de 5, ja que és molt improbable que aquesta circumstància s'esdevinga. No obstant, en realitat els resultats no són idèntics, doncs es diferencien en centèsimes de BLEU, però l'arrodoniment dels resultats a dècimes ha propiciat aquesta curiosa coincidència de resultats. També s'observa

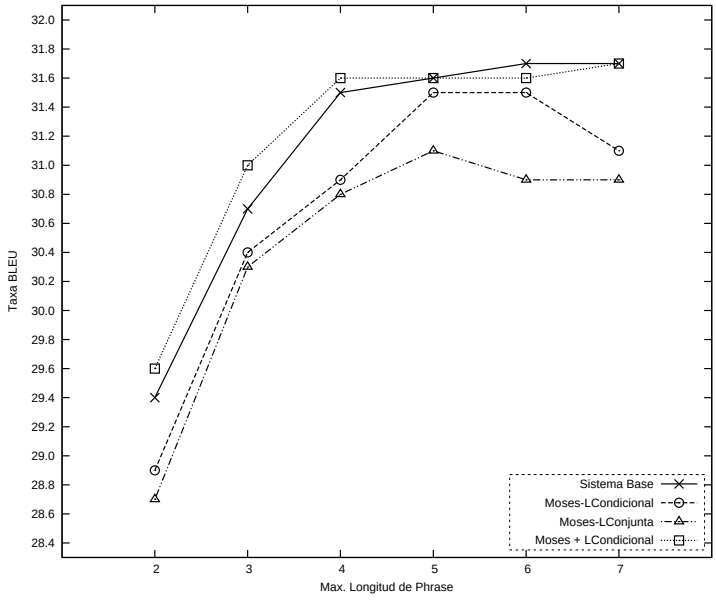


Figura 4.4: Taxa BLEU, en funció de la màxima longitud que poden assolir les seqüències de paraules, del sistema base i dels tres sistemes proposats que implementen el model de longitud estàndard, estimat a partir de les segmentacions de Viterbi extretes del conjunt d'entrenament del corpus Europarl-v3, per a la direcció de traducció Espanyol - Anglès.

una certa inestabilitat en el comportament dels sistemes proposats: per exemple, el sistema *Moses-LConjunta* entrenat amb una màxima longitud de 7 avantatja en 2 dècimes al sistema base i en 5 al sistema *Moses-LCondicional*, quan en la resta de condicions mostra resultats inferiors al sistema base i molt similars al sistema *Moses-LCondicional*. Trobem que això és degut a un procés d'ajustament de paràmetres poc estable i amb moltes fluctuacions que defineix uns pesos poc adequats per a les característiques del model log-lineal.

D'altra banda, a la Figura 4.2 podem observar els resultats per a la direcció de traducció Espanyol - Anglès. Els sistemes *Moses-LCondicional* i *Moses-LConjunta* mostren unes prestacions clarament inferiors al sistema base, mentre que el sistema *Moses + LCondicional* mostra un comportament similar al sistema base, o inclús podríem dir que lleugerament millor, per a longituds màximes més curtes (de 2 a 5 paraules). A partir d'una longitud màxima de 6 les prestacions cauen per baix del sistema base.

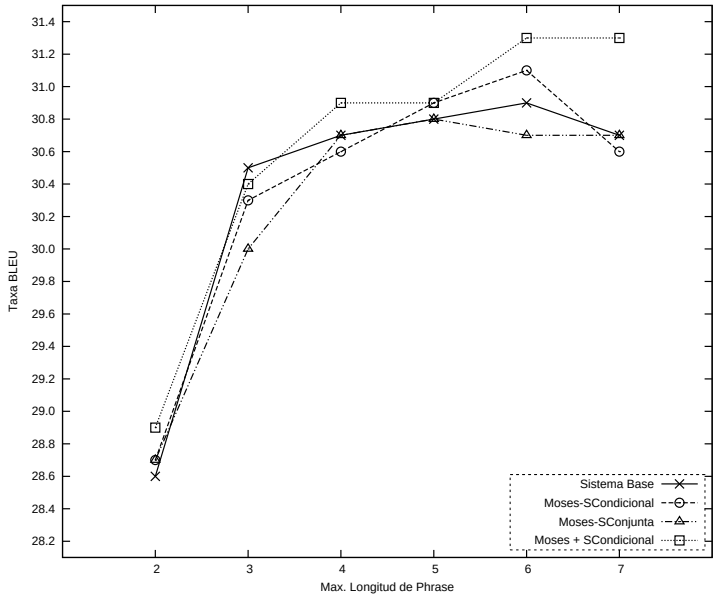


Figura 4.5: Taxa BLEU, en funció de la màxima longitud que poden assolir les seqüències de paraules, del sistema base i dels tres sistemes proposats que implementen el model de longitud especialitzat, estimat a partir del conjunt de seqüències de paraules extretes del conjunt d'entrenament del corpus Europarl-v3, per a la direcció de traducció Anglès - Espanyol.

Estimació a partir de les segmentacions de Viterbi

En segon lloc, es mostren els resultats de la experimentació amb els sistemes proposats, amb els models de longitud construïts a partir de les segmentacions de Viterbi.

D'una banda, a la Figura 4.3 podem observar els resultats per a la direcció de traducció Anglès - Espanyol. De nou ens trobem en que els sistemes *Moses-LCondicional* i *Moses-LConjunta* són comparativament inferiors al sistema base. Ara bé, en aquest cas el sistema *Moses + LCondicional* sí que presenta un comportament clarament millor al sistema base, destacant l'increment de 5 dècimes de BLEU en el cas de la longitud de seqüències limitada a 2 paraules, o l'increment de 3 dècimes per a longituds màximes de 5 i 7 paraules.

D'altra banda, a la Figura 4.4 podem observar els resultats per a la direcció de traducció Espanyol - Anglès. Els resultats són molt similars als observats a la Figura 4.2: els sistemes *Moses-LCondicional* i *Moses-LConjunta* mostren unes prestacions clarament inferiors al sistema base, mentre que el sistema *Moses + LCondicional* presenta una qualitat de resultats molt similar al sistema base, si bé en aquest cas el grau de similitud és major, i mostrant novament que les prestacions d'aquest sistema

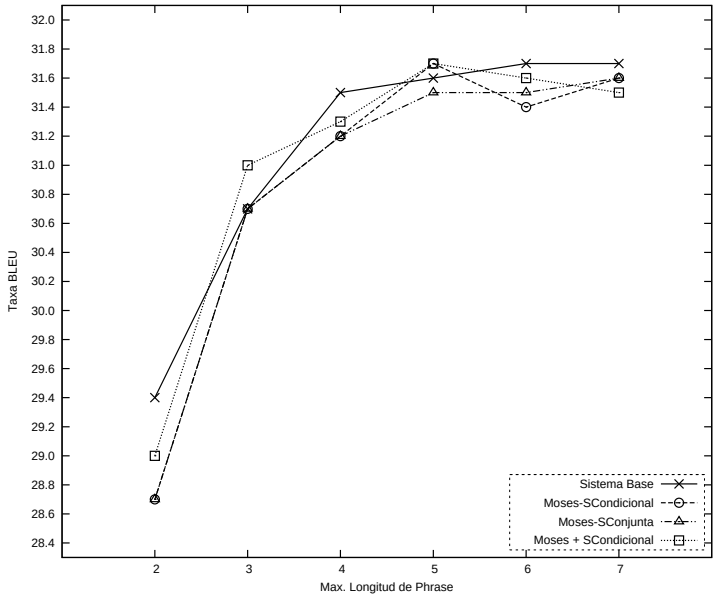


Figura 4.6: Taxa BLEU, en funció de la màxima longitud que poden assolir les seqüències de paraules, del sistema base i dels tres sistemes proposats que implementen el model de longitud especialitzat, estimat a partir del conjunt de seqüències de paraules extretes del conjunt d'entrenament del corpus Europarl-v3, per a la direcció de traducció Espanyol - Anglès.

empitjoren a partir d'una longitud màxima de seqüències de 6 paraules. Destaca sobretot l'empitjorament de les prestacions del sistema *Moses-LConjunta*, mostrant un decrement entre 4 i 8 dècimes de BLEU respecte al sistema base en totes les condicions analitzades.

4.2.3 Model de longitud especialitzat

A continuació s'exposen els resultats dels experiments realitzats amb els sistemes que implementen el model de longitud especialitzat, diferenciant els resultats per a les dues formes diferents d'estimar els models i per cada direcció de traducció.

Estimació a partir de seqüències de paraules extretes en la fase d'entrenament

En primer lloc considerarem els resultats dels experiments realitzats amb els tres sistemes proposats, estimant els models de longitud a partir del conjunt de seqüències de paraules extretes de corpus.

D'una banda, a la Figura 4.5 podem observar els resultats per a la direcció de

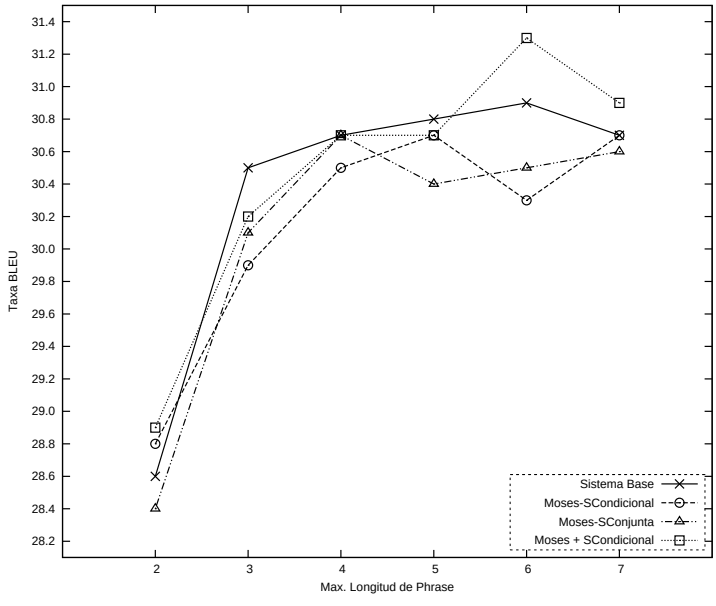


Figura 4.7: Taxa BLEU, en funció de la màxima longitud que poden assolir les seqüències de paraules, del sistema base i dels tres sistemes proposats que implementen el model de longitud especialitzat, estimat a partir de les segmentacions de Viterbi extretes del conjunt d'entrenament del corpus Europarl-v3, per a la direcció de traducció Anglès - Espanyol.

traducció Anglès - Espanyol. A primera vista, observem un comportament força inestable dels sistemes proposats. Cal analitzar, un per un, el comportament de tots tres sistemes: en primer lloc, el sistema *Moses-SConjunta* mostra un funcionament similar al sistema base, excepte en casos puntuals en els que mostra un descens de prestacions, exactament per a una longitud de seqüències limitada a 6 paraules, i sobretot a 3, on s'aprecia un decrement de 5 dècimes de BLEU. En segon lloc, el sistema *Moses-SCondicional* presenta un comportament similar al sistema base, però a la vegada inestable: en certes condicions millora les prestacions, en altres les empitjora. Per últim, el sistema *Moses + SCondicional* mostra en general una millora de les prestacions, especialment quan aquest és entrenat limitant la longitud de les seqüències de paraules a 7, amb un increment de 6 dècimes de BLEU respecte al sistema base.

D'altra banda, a la Figura 4.6 podem observar els resultats per a la direcció de traducció Espanyol - Anglès. Podem observar, a nivell general, que es resultats dels experiments realitzats amb els quatre sistemes es troben en intervals de BLEU d'entre 2 i 3 dècimes, circumstància que encara no s'havia donat. Aquest fet denota que tots tres sistemes presenten unes prestacions similars al sistema base, tot i que dels resultats obtesos es desprén un lleu descens de les prestacions.

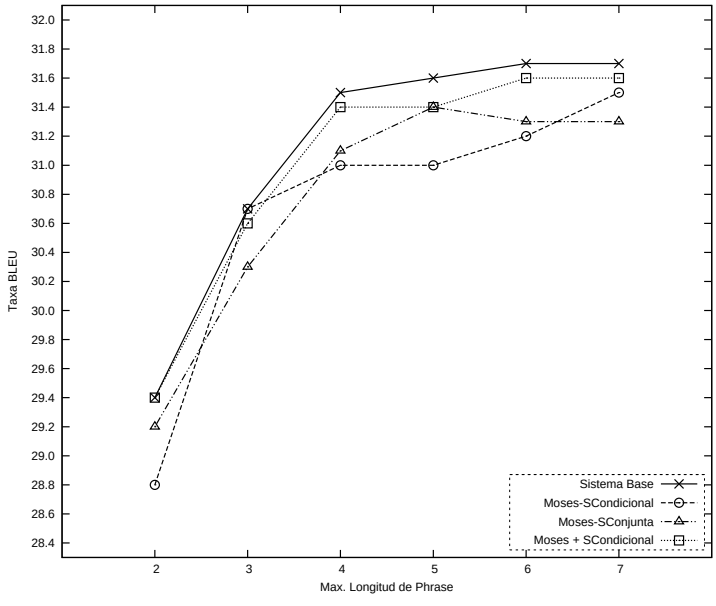


Figura 4.8: Taxa BLEU, en funció de la màxima longitud que poden assolir les seqüències de paraules, del sistema base i dels tres sistemes proposats que implementen el model de longitud especialitzat, estimat a partir de les segmentacions de Viterbi extretes del conjunt d'entrenament del corpus Europarl-v3, per a la direcció de traducció Espanyol - Anglès.

Estimació a partir de les segmentacions de Viterbi

En segon lloc, es mostren els resultats de la experimentació amb els sistemes proposats, amb els models de longitud construïts a partir de les segmentacions de Viterbi.

D'una banda, a la Figura 4.7 podem observar els resultats per a la direcció de traducció Anglès - Espanyol. El que s'observa a primera vista és una gran inestabilitat en els comportaments dels tres sistemes. Si ens fixem en els resultats, s'aprecia que les prestacions de tots tres sistemes són inferiors al sistema base, exceptuant el sistema *Moses + SCondicional*, que mostra, a l'igual que el sistema homònim estimat a partir de les seqüències de paraules extretes del corpus d'entrenament (Figura 4.5), una millora de les prestacions per a longituds màximes de seqüències de 6 i 7 paraules.

D'altra banda, a la Figura 4.8 podem observar els resultats per a la direcció de traducció Espanyol - Anglès. En aquest cas ens trobem amb que tots tres sistemes presentacions clarament inferiors al sistema base. També s'aprecia certa inestabilitat en els resultats dels sistemes *Moses-SCondicional* i *Moses-SConjunta*, amb tendències molt diferenciades conforme augmenta la longitud màxima de les seqüències de paraules.

Capítol 5

Conclusions i treball futur

5.1 Resum

Arribats a aquest punt és moment de recapitular. Al Capítol 1 hem introduït al lector al context de la Traducció Automàtica, mostrant les diferents formes d'enfocar aquest problema, encara per resoldre. La falta de precisió d'aquests sistemes ens ha servit per justificar l'objectiu que es persegueix a aquest treball, que és introduir i avaluar possibles millores en el procés de traducció. Posteriorment ens hem ocupat més a fons de l'aproximació estadística a la Traducció Automàtica, donant una ullada als diferents enfocaments estadístics que ens donen una idea de quant activa és aquesta disciplina en el camp de la investigació, i hem posat especial èmfasi amb els sistemes basats en seqüències de paraules, doncs aquest treball s'ha desenvolupat en aquest context, i no és per casualitat: són els sistemes que millors prestacions ofereixen en l'actualitat. Al Capítol 2 hem centrat la nostra atenció als sistemes de TA basats en seqüències de paraules, i més concretament en el sistema *Moses*, un programari de codi obert basat en aquesta aproximació que ens ha facilitat la implementació i avaluació de la benignitat de les millores proposades, evitant així implementar un sistema complet (amb el conseqüent estalvi de temps). A més, en aquest capítol hem introduït les millores que es proposen a aquest projecte, motivades per unes mancances detectades al sistema *Moses*, i inspirades en un treball anterior que aboca previsions optimistes sobre les mateixes. La idea principal és afegir un modelat de la longitud de les seqüències de paraules al sistema base. Partint d'aquesta idea s'ha presentat formalment el model de longitud, amb les seves dues variants: un model que té en compte únicament les longituds de les seqüències de paraules (model de longitud estàndard), i un altre més restringit que té en compte la longitud de la seqüència de paraules d'eixida del model donada una seqüència de paraules concreta (model de longitud especialitzat). Posteriorment, al Capítol 3 hem estudiat la forma d'estimar aquests models, així com la seva implementació i integració en *Moses*, proposant una sèrie de sistemes que inclouen aquests models per a ser avaluats. Per fi, al Capítol 4 hem avaluat les prestacions dels sistemes proposats al capítol anterior, en comparació al sistema base, no sense abans presentar el corpus d'exemples de traducció amb el

que hem treballat per poder dur a terme aquestes experimentacions. Per últim, en el present capítol exposarem les conclusions a les que havem arribat després de realitzar aquest treball.

5.2 Conclusions

A la vista dels resultats observats al capítol anterior, no podem extreure cap conclusió determinant, doncs no queda clar que el modelat de la longitud constituïsca una millora de les prestacions. Depenent del context, podem trobar-nos amb que els models aporten informació valuosa millorant les prestacions, o bé amb que provoquen un empitjorament de la qualitat de la traducció. No obstant això, sí podem extraure conclusions més concretes, que tot seguit enumerem:

- Les eventuals millores de prestacions observades no són estadísticament significatives. Aquestes millores respecte al sistema base es quantifiquen en la majoria de casos entre 1 i 3 dècimes de BLEU, excepte casos aïllats en els que hem arribat a obtindre una millora de 4, 5 i 7 dècimes. Si a més tenim en compte que la majoria d'experiments han observat un descens de les prestacions, la validesa d'aquestes millores és més que qüestionable.

- Existeix un comportament un tant erràtic i inestable dels sistemes proposats conforme es varia la longitud màxima de les seqüències de paraules, produint-se variacions positives i negatives de la taxa BLEU molt acusades respecte al sistema base i la resta de sistemes. Aquest comportament anormal s'aprecia a les gràfiques en els constants encreuaments entre les línies, quan el perfil esperat de les línies que representen les prestacions dels sistemes proposats deuria de ser aproximadament paral·lel al sistema base, però per baix o per dalt (observar Figura 4.3). Aquesta inestabilitat sembla motivada per la pròpia inestabilitat inherent al procés d'ajust de pesos de les característiques del model log-lineal, el MERT (veure Secció 2.1.7), provocant una assignació de valors no adequats. Aquest inconvenient, aliè al que concerneix a aquest treball, impedeix avaluar d'una forma més precisa i transparent les prestacions reals d'aquests sistemes.

- En general, els models de longitud aporten informació benigna traduint de l'anglès a l'espanyol, però no a l'inrevés, on generalment les prestacions no milloren. Això pot ser degut a característiques intrínseques d'aquest procés, en oposició al procediment invers. Caldria realitzar un estudi exhaustiu per trobar una explicació raonable a aquest fenomen.

- Els sistemes que implementen el model de longitud estàndard ofereixen, en general, millors resultats i més estabilitat que els sistemes que implementen el model de longitud especialitzat. Donat que l'únic que diferencia ambdós aproximacions és el model condicional, aquestes circumstàncies s'expliquen perquè el model de longitud estàndard, al ser més general que el model especialitzat (molt més restringit), es troba millor estimat, ja que, independentment de la font

d'informació, es poden observar molts més esdeveniments per estimar els seus paràmetres. Recordem, a més, que el model de longitud estàndard requereix estimar molts menys paràmetres que el model especialitzat.

- Dels tres sistemes proposats per a cada tipus d'experiment, l'únic que millora eventualment les prestacions del sistema base és el *Moses + /S/L/Condicional*. Els nostres presagis sobre el millor funcionament d'aquests sistemes s'han confirmat: és molt més beneficiós aportar la informació del modelat de la longitud com una característica i font d'informació independent, que integrar-la en el model de traducció de seqüències de paraules de *Moses*. Si bé aquesta segona opció és, baix un punt de vista teòric, més correcta (veure Equació (2.23)), els resultats experimentals mostren un empitjorament de les prestacions; mentre que conferir tota l'expressivitat del model condicional de longitud al model log-lineal presenta empíricament millors resultats que el sistema base. L'empitjorament de les prestacions respecte al sistema base en els models complets es pot interpretar com que el model de traducció de seqüències de paraules, que aporta informació molt concisa i determinant a l'hora d'avaluar una possible traducció, és interferit o pertorbat pel model de longitud, de forma que els dos models en conjunt no són capaços d'aportar informació més precisa, sinó més be al contrari.

5.3 Contribucions científiques

L'aportació d'aquest treball al món de la traducció automàtica és la integració, com una font d'informació addicional i independent en el model log-lineal d'un sistema heurístic de traducció automàtica basat en seqüències de paraules, d'un modelat de la longitud de les seqüències de paraules que formen part dels paràmetres del model de traducció.

5.4 Treball futur

Com hem pogut comprovar, no hem estat capaços de formular una conclusió general respecte a les millores proposades. És per això que el treball realitzat no acaba en aquestes línies, i per tant proposem una sèrie de treballs futurs encaminats a ampliar i millorar el treball ací realitzat. Aquestes són les nostres propostes:

- Emprar tècniques més robustes per a l'ajustament dels pesos de les característiques del model log-lineal que garantisquen una major estabilitat de resultats, a fi de poder avaluar de forma més precisa les prestacions dels sistemes proposats.

- Explorar tècniques d'estimació de major correcció dels models proposats, com per exemple l'aprenentatge per màxima versemblança de les segmentacions de les frases d'entrada i d'eixida directament del corpus d'entrenament, mètode que requereix la realització d'un entrenament *Expectation-Maximization*.

- Emprar altres tècniques alternatives a la interpolació lineal per suavitzar el model de longitud especialitzat (per exemple el descompte de Kneser-Ney emprat als models de llenguatges d'n-grames), donat que presenta majors indicis de sobreestimació.

- Incloure informació de la longitud de les seqüències de paraules al model de reordenament lexicalitzat, de forma anàloga a la inclusió d'aquesta informació al model de traducció.

La realització d'aquestes ampliacions podrien traduir-se en un increment significatiu de les prestacions del sistema base, que és un dels objectius que s'ha perseguit a aquest treball, i que s'ha aconseguit de forma parcial.

Bibliografia

[AdT82] D. Arnold and L. des Tombe. Basic theory and methodology in eurotra. In S. Niremburg, editor, *Machine Translation: Theoretical and Methodological Issues*, pages 114–135, 1982.

[AF10] Jesús Andrés-Ferrer. *Statistical approaches for natural language modelling and monotone statistical machine translation.* PhD thesis, Universidad Politécnica de Valencia, Valencia (Spain), Feb 2010. Advisors: A. Juan and F. Casacuberta.

[AFJ09] Jesús Andrés-Ferrer and 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.

[B+90] P. F. Brown et al. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85, 1990.

[B+93] P. F. Brown et al. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[BBC+09] Sergio Barrachina, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, and Juan M. Vilar. Statistical approaches to computer-assisted translation. *Comput. Linguist.*, 35(1):3–28, 2009.

[BF81] A. Ban and A. Feigenbann. *The Handbook of Artificial Intelligence.* Pitman, 1981.

[BH60] Y. Bar-Hillel. The present status of automatic translation of languages. *Advances in Computers*, 1:91–163, 1960.

[Bil82] R. Billmeyer. Zu den linguistischen Grundlagen von SYSTRAN. *Multilingua*, 2(1):83–96, 1982.

[BN04] M. Bisani and 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.

[CBKM+10] Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and 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.

[Civ08] J. Civera. *Novel statistical approaches to text classification, machine translation and computer-assisted translation.* PhD thesis, Universidad Politécnica de Valencia, Valencia (Spain), Juny 2008. Advisors: A. Juan and F. Casacuberta.

[CV04] Francisco Casacuberta and Enrique Vidal. Machine translation with inferred stochastic finite-state transducers. *Comput. Linguist.*, 30(2):205–225, 2004.

[CV07] F. Casacuberta and E. Vidal. Learning finite-state models for machine translation. *Machine Learning*, 66(1):69–91, 2007.

[CVP05] F. Casacuberta, E. Vidal, and D. Picó. Inference of finite-state transducers from regular languages. *Pattern Recognition*, 38:1431–1443, 2005.

[Dej] Dejavú. http://www.atril.com.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[Dod02] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *HLT '02: Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[FLL02] G. Foster, P. Langlais, and G. Lapalme. User-friendly text prediction for translators. In *Proc. of EMNLP'02*, pages 148–155, Morristown, NJ, USA, July 2002. Association for Computational Linguistics.

[Fos02] G. Foster. *Text Prediction for Translators.* PhD thesis, Université de Montréal, May 2002.

[Goo] Google translate toolkit. http://translate.google.com/support/toolkit/.

[GV03] I. García-Varea. *Traducción automática estadística: modelos de traducción basados en máxima entropía y algoritmos de búsqueda.* PhD thesis, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, December 2003.

[HK] Jochen Hummel and Iko Knyphausen. Trados. http://www.trados.com.

[IC97] P. Isabelle and K. Church. Special issue on new tools for human translators. *Machine Translation*, 12(1–2), 1997.

[Kay97] M. Kay. The proper place of men and machines in language translation. *Machine Translation*, 12:3–23, 1997.

[KHB+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, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL*. The Association for Computer Linguistics, 2007.

[Kni99a] K. Knight. Decoding complexity in word-replacement translation models. *Computional Linguistics*, 25(4):607–615, 1999.

[Kni99b] Kevin Knight. A stadistical machine translation tutorial workbook. http://www.isi.edu/natural-language/mt/wkbk.pdf, August 1999.

[Koe04] Philipp Koehn. Statistical significance tests for machine translation evaluation, 2004.

[Koe05] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proc. of the MT Summit X*, pages 79–86, September 2005.

[Koe10] Philipp Koehn. *Stadistical Machine Translation*. Cambridge University Press, Edinburgh (United Kingdom), 2010.

[LFL00] P. Langlais, G. Foster, and G. Lapalme. Unit completion for a computer-aided translation typing system. *Machine Translation*, 15(4):267–294, 2000.

[LLL02] P. Langlais, G. Lapalme, and M. Loranger. Transtype: Development-evaluation cycles to boost translator's productivity. *Machine Translation*, 15(4):77–98, 2002.

[MS99] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[Mur66] Hubert Murray. *Methods for Satisfying the Needs of the Scientist and the Engineer for Scientific and Technical Communication*. In a Press Release, Washington D.C., 1966.

[NGW95] Hermann Ney, M. Generet, and 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.

[Och03] 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.

[ON03] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29, 2003.

[PC66] John R. Pierce and John B. Carroll. Languages and machines — computers in translation and linguistics. Technical report, Automatic Language Processing Advisory Committe (ALPAC), National Academy of Sciences, 1966.

[PRWZ01] K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176, Thomas J. Watson Research Center, 2001.

[SDS+06] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and 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.

[Slo85] J. Slocum. A survey of machine translation: its history, current status and future prospects. *Computational Linguistics*, 11(1):1–17, 1985.

[Sto02] A. Stolcke. Srilm – an extensible language modeling toolkit. http://www.speech.sri.com, 2002.

[Tih82] B. Tihouin. The Meteo System. In Veronica Lawson, editor, *Proc. of Practical Experience of Machine Translation*, pages 39–44, 1982.

[TVN+97] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. Accelerated dp based search for statistical translation. In *In European Conf. on Speech Communication and Technology*, pages 2667–2670, 1997.

[UoG95] Multilingual Information Processing Department University of Geneva. Evaluation of natural language processing systems. http://www.issco.unige.ch/en/research/projects/ewg95/, 1995.

[VB85] Bernard Vauquois and Christian Boitet. Automated translation at grenoble university. *Comput. Linguist.*, 11(1):28–36, 1985.

[Wea55] W. Weaver. Translation. In W. N. Locke and A. D. Booth, editors, *Machine Translation of Languages: fourteen essays*, pages 15–23. MIT Press, Cambridge, MA., 1955.

[WWC+86] P. J. Whitelock, M. McGee Wood, B. J. Chandler, N. Holden, and H. J. Horsfall. Strategies for interactive machine translation: the experience and implications of the UMIST Japanese project. In *Proc. of COLING'86*, pages 329–334, Bonn, Germany, August 1986.

Índex de figures

Índex de taules