Disseny d'un etiquetari morfosintàctic per al traductor automàtic interNOSTRUM

Raül Canals i Mikel L. Forcada
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E−03071 Alacant

1 Introducció

1.1 Etiquetaris

Un *etiquetari morfosintàctic* és el conjunt d'etiquetes (categories) que es poden assignar als mots d'un idioma, de manera que tots els mots que reben una determinada etiqueta tenen un comportament sintàctic similar. Un mot pot rebre més d'una etiqueta: el mot *ahorro* pot rebre les etiquetes "substantiu masculí singular" i "verb lèxic en present d'indicatiu".
Els etiquetaris poden variar en el seu grau de refinament: poden ser gruixuts ("verb") o fins ("verb lèxic en tercera persona del plural del present d'indicatiu"). L'elecció depén de l'aplicació que es vulga donar a l'etiquetari; una n'és la traducció automàtica. Aquesta comunicació descriu el conjunt d'etiquetes del castellà usat en el traductor automàtic castellà−català *interNOSTRUM* i els criteris usats per a dissenyar−lo.

1.2 El sistema de traducció automàtica *interNOSTRUM*

interNOSTRUM (disponible en línia: `www.internostrum.com`) és un sistema de traducció automàtica indirecta *per transferència*, del castellà al català; fa la traducció en tres fases: l'*anàlisi*, que converteix el text de partida en una representació abstracta que destaca els detalls rellevants per a la traducció; la *transferència*, que converteix la representació abstracta del text de partida en una representació abstracta similar del text d'arribada, i la *generació* que a partir de la representació abstracta del text d'arribada genera el text traduït.

La fase d'anàlisi d'*interNOSTRUM* es compon de tres subfases:

- l'*anàlisi morfològica*, que per a cada forma superficial (forma del mot en el text) entrega una o més formes lèxiques (FL) compostes de *lema* o forma canònica, categoria lèxica, i informació de flexió. Per a la f. superficial *paseábamos* el lema és *pasear*, la categoria *verb lèxic*, i la informació de flexió *pretèrit imperfet, indicatiu, 1a. persona, pl.*

- si un mot té més d'una forma lèxica, el *desambiguador* analitza la categoria gramatical i la informació de flexió de cada FL i li assigna una etiqueta (no té en compte tota la informació present en la forma lèxica); després, aplica criteris estadístics i tria una de les etiquetes per al mot homògraf en qüestió.

- finalment, un programa detecta les construccions de dos o més FL consecutives (sintagmes rudimentaris) que requeriran un tractament especial (concordança de gènere, de nombre, canvis de preposició, reordenament, etc.).

2 Factors que influeixen en el disseny de l'etiquetari

Motivació lingüística: l'etiquetari ha d'agrupar categories de mots amb comportaments sintàctics homogenis i destriar les que tenen comportaments desiguals: l'etiqueta *verb* és massa gruisuda; si hi distingim gerundis, infinitius, participis i verbs en forma personal, podem dir que en castellà un article o una preposició no pot anar seguit d'un verb conjugat, però si d'un infinitiu. Si un mot rep dues etiquetes quasi impossibles de destriar usant informació del context sintàctic immediat, serà molt difícil desambiguar–lo. Així, el verb *cantamos* pot estar en present o en pretèrit indefinit d'indicatiu; la traducció al català és diferent en cada cas (*cantem*, *vam cantar*) però és difícil que un programa puga decidir en quin cas es troba.

Universalitat *versus* adeqüació: Un etiquetari es pot dissenyar per una aplicació concreta o de forma genèrica. Nosaltres hem intentat distingir les categories que, en els mots homògrafs, donen lloc a traduccions diferents però, hem renunciat a diferenciar les que no plantegen problemes de traducció. A més, hem optat per etiquetes que es poden inferir fàcilment a partir dels indicadors categorials i morfològics que entrega l'analitzador morfològic.

Grandària, és a dir, finesa de l'etiquetari: Un etiquetari gran ens permet fer distincions més fines que un de menut. Però augmentar el nombre d'etiquetes multiplica la complexitat del desambiguador. Duplicar l'etiquetari, quadriplica el nombre de paràmetres numèrics per a un model de bigrames (seqüències de dues etiquetes) i, l'octuplica per a un model de trigrames. S'hi ha d'establir un compromís, actualment en *interNOSTRUM* es situa en 60 etiquetes.

Semblança amb altres etiquetaris: Hi ha diversos etiquetaris per al castellà (CRATER/XEROX, IULA). L'etiquetari d'*interNOSTRUM* els ha tingut en compte per poder aprofitar l'experiència continguda en el seu disseny, i per poder usar etiquetadors existents o textos etiquetats (manualment o automàtica) per avaluar el nostre etiquetador.

2.1 Un cas especial: les unitats multimot

Les *unitats multimot* són seqüències de mots tractades com un únic mot: *a través de* com una preposició, *direcció general* com un substantiu, *Sant Vicent del Raspeig* com a nom propi o *tener que* com un verb. Aquestes unitats multimot es corresponen amb parts de l'oració clarament identificables que reben les etiquetes corresponents.

3 Disseny de l'etiquetari guiat pels problemes de traducció

Quins problemes de traducció es poden resoldre amb un etiquetari? Abans de res, cal distingir quins problemes no es podran resoldre:

- La polisèmia, el fenomen pel qual formes superficials que tenen una única forma lèxica són ambigües perquè el lema pot tenir més d'una interpretació (p.ej. *estación*, *destino*).

- Homògrafs on només canvia el lema: *creo* (lemes *crear* i *creer*), *vendo* (lemes *vendar* i *vender*). En alguns casos tractem aquest problema distingint amb categories diferents els verbs modals o auxiliars de la resta dels verbs (lèxics), com ara *fui* (lemes *ir* i *ser*).

Podem intentar resoldre la resta dels homògrafs que afecten la qualitat de la traducció. El cas més obvi són els homògrafs categorials. En altres casos l'homografia no afecta la categoria, però sí la traducció (taula 1). És difícil que un etiquetador basat en regles resolga algunes d'aquestes distincions, ja que pràcticament qualsevol context és compatible amb qualsevol desambiguació. Els mètodes estadístics són els únics que podrien recollir informació per a fer la tria, i per això en *interNOSTRUM* s'ha optat pels etiquetadors estadístics.

TIPUS	HOMÒGRAF	TRADUCCIÓ
CATEGORIALS	*ahorro* (subst./verb) *vino* (subst./verb) *para* (prep./verb.) *como* (conj./verb.)	*estalvi*, *estalvio* *vi*, *va venir* *per (a)*, *para/atura* *com (a)*, *menjo*
NO CATEGORIALS	*està* (imperat./indic.) *cantamos* (pres./perf.) *vivimos* (pres./perf.) *salen* (indic./subj.) *venga* (indic./subj.)	*estigues*, *està* *cantem*, *vam cantar* *vivim*, *vam viure* *surten*, *salin* *venja*, *vingui*

Taula 1: Homògrafs castellans i traduccions al català

En castellà, un de cada tres o quatre mots és homògraf. Si el desambiguador comet errors en mots molt poc freqüents, no afecta massa la qualitat global de la traducció; el problema és quan comet errors en els mots homògrafs més comuns (taula 2).

Els mots *para*, *una* i *como* són especialment problemàtics perquè les traduccions són divergents. Per poc que el desambiguador estadístic s'equivoque en aquests mots es degrada notablement la qualitat de les traduccions. En *interNOSTRUM* aquesta és una causa freqüent d'errors de traducció.

Mot	categories	*f*
la	pronom, article	0,051
que	conj., relatiu	0,027
los	pronom, article	0,027
las	pronom, article	0,018
para	verb, preposició	0,0091
una	verb, article	0,0088
como	verb, conjunció	0,0050

Taula 2: Mots ambigus més freqüents en castellà

Per a dissenyar l'etiquetari, es va construir un catàleg de situacions on l'homografia afectava la traducció; estudiant la casuística existent, i llegint les traduccions produïdes per detectar problemes que es pogueren resoldre fent les distincions oportunes. Es van generar tots els homògrafs que acceptava l'analitzador morfològic que eren susceptibles de solució mitjançant un refinament de l'etiquetari i se'n van destriar aquells que donaven lloc a traduccions divergents.

4 Descripció de l'etiquetari d'interNOSTRUM

La taula 3 descriu l'etiquetari que usa *interNOSTRUM*. En general, la divisió en etiquetes ha estat inspirada per les indicacions de la secció anterior, però algunes en mereixen explicació a banda.

- S'han distingit els antropònims perquè moltes vegades apareixen en grups de més d'un (*Juan López Sanchis*) o darrere d'un substantiu comú (*El profesor López*).

- S'han distingit els topònims perquè requereixen la preposició *a* en les construccions locatives en català (*en Valencia* → *a València*).

- S'han distingit els predeterminants d'altres determinants per evitar seqüències de més d'un determinant. Així, podem tractar bé seqüències com *una los dos cabos y...* on el mot *una* només pot ser verb si es prohibeixen les seqüències de dos determinants.

- S'han distribuït en diverses etiquetes les formes conjugades dels verbs per distingir homògrafs castellans que havien de tenir traduccions divergents al català.

- S'han distingit els verbs lèxics d'altres verbs que actuen com a auxiliars (*ser*, *haber*) o com a modals *poder*, *deber*, per raó de les diferències entre els seus contextos sintàctics habituals i per a poder desambiguar, almenys parcialment, mots com *podemos* (*podar* lèxic però *poder* modal) o *fue* (*ser* auxiliar però *ir* lèxic).

ETIQUETA	DESCRIPCIÓ	EXEMPLES
`ACRONIM`	Acrònim	*CAM*, *ONU*, *sida*
`NOM`	Substantiu comú	*casa*, *dígitos*
`ANTROPONIM`	Nom propi de persona	*Javier*, *López*
`TOPONIM`	Nom propi de lloc	*Valencia*, *Italia*
`NPALTRES`	Altres noms propis	*Moulinex*, *Pentium*
`ADJ`	Adjectius	*rojos*, *mío*
`ADV`	Adverbi	*más*, *ahora*
`PREADV`	Adv. preadjectival o preadverbial	*muy*, *tan*
`CNJCOORD`	Conjunció coordinant	*i*, *e*, *pero*
`CNJSUB`	Conjunció subordinant	*que*
`CNJADV`	Conjunció adverbial	*aunque*, *si*
`DET`	Determinants	*aquel*, *el*, *su*
`DETNEU`	Determinant neutre	*lo*
`INTNOM`	Pronom interrogatiu	*cuál*, *quién*
`INTADJ`	Adjectiu interrogatiu	*cuánto*, *qué*
`INTADV`	Adverbi interrogatiu	*cómo*, *dónde*
`INTERJ`	Interjecció	*adiós*, *ojalá*
`NUM`	Numeral cardinal	*sesenta y cinco*

ETIQUETA	DESCRIPCIÓ	EXEMPLES
PREDET	Predeterminant	*toda*
PREDETNEU	Predeterminant neutre	*todo*
PREP	Preposició	*a*, *de*
PRNTON	Pronom tònic	*ambos*, *nadie*
PRNTONNEU	Pronom tònic neutre	*algo*, *eso*
PRNATNAC	Pronom àton acusatiu	*la*, *me*
PRNATNDAT	Pronom àton datiu	*le*
PRNATNREF	Pronom àton reflexiu	*se*
REL	Relatiu (aa, an, nn)	*cuyo*, *que*, *quien*
RELADV	Relatiu adverbial	*adonde*, *cuando*
Verbs: diferenciem verbs lèxics (*escuchar*), modals (*deber*, *poder*), el verb *ser* i el verb *haber*. En cadascun d'ells diferenciem infinitiu, gerundi i participi i agrupem els temps conjugats en les etiquetes PFCI (present, futur i condicional d'indicatiu), IPI (pretèrit imperfet i indefinit d'indicatiu), SUBJ (subjuntiu), IMP (imperatiu).		
Altres: l'etiqueta LQUEST per al símbol d'apertura d'interrogació (*¿*), RPAR per al parèntesi i claudàtor dret (*)]*), LPAR (*([*), CM per a la coma (,) i SENT per a la fi de frase (*.:;?!*).		

Taula 5: L'etiquetari actual d'interNOSTRUM

5 Conclusió

Hem presentat el conjunt d'etiquetes usades pel desambiguador del traductor automàtic castellà–català interNOSTRUM. La tria d'etiquetes ha estat motivada per criteris sintàctics, agrupant categories gramaticals que apareixen en els mateixos contextos sintàctics en una única etiqueta. També hem considerat la distinció de les categories que, en els mots homògrafs, donen lloc a traduccions diferents. Hem considerat la freqüència d'aparició dels homògrafs per tal de prioritzar la resolució d'aquells que més degraden la qualitat de les traduccions. Quant al nombre d'etiquetes desambiguadores, hem buscat un compromís entre el nombre d'etiquetes i la complexitat del programa desambiguador de manera que tenim 60 etiquetes, el model estadístic té 3.600 bigrames o 216.000 trigrames, segons el cas.

Agraïments i reconeixements:

Per raons de l'organització del *XVI Encuentro de la Asociación de Jóvenes Lingüistas*, l'article no podia anar signat per Amaia Iturraspe Bellver i Anna Esteve Guillén, que són autores de ple dret. El treball ha estat finançat per la Caja de Ahorros del Mediteráneo y pel Vicerectorat de Noves Tecnologies de la Universitat d'Alacant.