

CREACIÓ DE RECURSOS LINGÜÍSTICS PER A LA TRADUCCIÓ AUTOMÀTICA

Victoria Arranz, Núria Castell i Jesús Giménez
Centre de Recerca TALP
Universitat Politècnica de Catalunya
Campus Nord
08034 Barcelona
{varranz,castell,jgimenez}@talp.upc.es

1. INTRODUCCIÓ

En els darrers anys, la construcció de recursos lingüístics s'ha convertit en una tasca cada cop més important ateses les múltiples àrees de l'Enginyeria Lingüística on són necessaris. Congressos com LREC (International Conference on Language Resources and Evaluation)¹, aquest any en la seva quarta edició, posen en evidència l'interès pels recursos lingüístics, la seva evolució i la necessitat d'establir criteris d'avaluació de la qualitat. També aspectes com la creació de recursos per a llengües minoritàries són de continua actualitat en les edicions d'aquest congrés. En aquest article ens centrem en la creació de recursos orientats a la traducció automàtica de la parla basada en tècniques estadístiques.

L'enfocament estadístic està mostrant bons resultats, especialment en dominis restringits. Darrerament s'està incorporant cert coneixement lingüístic a aquests sistemes per tal de millorar els resultats de traducció (cf. [Ueffing i Ney, 2003] i [Toutanova et al., 2002]). S'estima que a mesura que es vagi utilitzant, l'informació lingüística contribuirà més a la millora dels resultats. Per exemple, [de Gispert i Mariño, 2003] mostren una millora inicial en la traducció estadística si integrem informació com la categorització del temps i les expressions de dates, informació que considerem i incloem en la DTD que es presenta en aquest treball.

La traducció automàtica estadística requereix corpus multilingües i paral·lels per a ser usats com entrenament dels sistemes automàtics de traducció. Addicionalment poden estar alineats (a nivell de frase, sintagma o paraula), possiblement etiquetats amb informació lingüística, i ser generats a partir de transcripcions de gravacions (quan es tracta de corpus orals). Aquests recursos són costosos de crear i, especialment, quan es vol treballar amb parla espontània. Si bé existeixen alguns recursos disponibles a través d'organitzacions com ELRA² i LDC³, només algunes llengües tenen presència important en aquests repositoris.

El centre TALP (Tecnologies i Aplicacions del Llenguatge i la Parla)⁴ està treballant en l'àmbit de la traducció automàtica des de diferents vessants. Ens centrarem aquí en el desenvolupament de recursos en el marc del projecte europeu LC-STAR⁵, on treballem tant el castellà com el català, i del projecte espanyol ALIADO⁶, on treballem el castellà.

¹ <http://www.lrec-conf.org>

² <http://www.elra.info>

³ <http://www ldc.upenn.edu>

⁴ <http://www.talp.upc.es>

⁵ <http://www.lc-star.com>

⁶ <http://gps-tsc.upc.es/veu/aliado/>

2. CORPUS DESENVOLUPATS

En el marc d'aquests projectes s'han creat dos tipus de recursos útils per a la traducció automàtica estadística, centrats en el domini turístic. D'una banda, s'ha creat un corpus trilingüe (anglès americà, castellà, català) de diàlegs i vuit corpus bilingües més petits tots ells tenint l'anglès americà d'una banda i una de les altres llengües a l'altra banda (en el nostre cas, doncs, anglès/català i anglès/castellà). S'ha fet també una tasca d'especificació de com han de ser aquests tipus de recursos i properament es faran públics els informes del projecte europeu on es descriuen les especificacions.

2.1 CORPUS BILINGÜES

Els corpus bilingües desenvolupats consisteixen en un recull de frases curtes o trossets de frases en anglès americà i traduïts, en el nostre cas, al català i al castellà. Aquests corpus tenen 10.500 entrades, són emmagatzemats en un format basat en XML i incorporen informació lingüística a nivell de paraula (etiqueta morfosintàctica i lema).

Les frases han estat extretes en part del corpus de diàlegs que descriurem més endavant, en part d'un corpus de textos d'informació turística obtinguts en pàgines web i en part d'exemples creats manualment a partir de llibres de frases típiques (*phrasal books*) que usen els turistes per a viatjar a països estrangers. El criteri per extreure frases dels diàlegs i dels textos ha estat la cobertura de les paraules més freqüents d'ambdós corpus.

2.2 CORPUS TRILINGÜE

El corpus trilingüe s'ha creat a partir de dues fonts [Arranz *et al.*, 2003]. Una primera part del corpus s'ha creat a partir del corpus Verbmobil⁷ que va ser desenvolupat a Alemanya en el marc d'un projecte nacional de traducció automàtica. La segona part del corpus (TALP-tourism) s'ha creat a partir de zero, començant per la gravació de diàlegs en català i castellà.

El corpus Verbmobil és un conjunt de diàlegs relacionats amb la concertació de reunions, trobades, etc. (*appointments*). Hem seleccionat el material disponible en anglès americà i l'hem traduït al català i castellà. El corpus té marques de fenòmens de parla espontània com ara sorolls, repeticions, truncaments, etc. S'han elaborat uns criteris de traducció per a tenir en compte tots aquests fenòmens i que l'objectiu és disposar d'un recurs per a sistemes de traducció estadística. El corpus seleccionat i traduït té més de 200.000 paraules, amb un vocabulari d'unes 3.300 paraules.

El corpus TALP-tourism ha estat creat a partir de zero, prenent model del corpus Verbmobil i centrant-nos en el domini turístic. S'han creat diferents escenaris (situacions típiques) i s'han gravat més de 350 diàlegs (unes 600.000 paraules) utilitzant locutors voluntaris. Aquesta base de dades oral creada conté un vocabulari d'unes 10.000 paraules en català i unes 11.000 en castellà. Els escenaris bàsics eren: hotel, agència de viatges, oficina d'informació turística i companyia aèria o de trens. Per cada escenari, uns quants subescenaris més concrets permetien ubicar-se als locutors, com són, per exemple, fer una reserva d'hotel o un bitllet d'avió. A més a més, es disposava d'unes plantilles que proporcionaven als locutors descripcions exemple sobre les que elaborar les seves converses. Aquestes gravacions s'han fet tant en català com en castellà. Un cop transcrits els diàlegs, el text català ha estat traduït a l'anglès americà i al castellà, mentre el text castellà ha estat traduït a l'anglès americà i al català. Els criteris de traducció han estat els mateixos que per al corpus Verbmobil.

⁷ <http://verbmobil.dfki.de/verbmobil>

3. ENRIQUIMENT DELS RECURSOS

El resultat de la tasca descrita abans és l'existència d'un corpus trilingüe que pot ser usat com a tal o es pot explotar com diferents corpus bilingües. En ambdós casos, es pot incloure informació lingüística utilitzant els analitzadors, etiquetadors i lematitzadors que tenim disponibles per les tres llengües. En el cas dels corpus bilingües, es pot incorporar informació d'alineament entre ambdues llengües. L'alineament a nivell de torns de diàleg és simple i no presenta problemes, però l'alineament a nivell de sintagmes, locucions o paraules és una tasca complexa. Com més informació sobre alineament entres dues llengües tinguem, millor funcionaran els sistemes estadístics de traducció automàtica. En el marc de LC-STAR s'ha fet també una tasca de disseny de formats [Arranz *et al.*, 2004] prou flexible que permeti representar tant els petits corpus bilingües com el corpus trilingüe o els corpus bilingües més grans. El format es basa en l'*Extensible Markup Language* (XML) [Bray *et al.*, 2000] i està pensat per a representar corpus de molts tipus diferents: mono/bi/multi-llengües, textuals o orals, de domini restringit o generals, etc.

3.1 DISSENY DE LA DTD: FORMAT I CONTINGUT

Previ al disseny de la nostra DTD, vàrem fer un estudi dels diferents enfocaments que s'han seguit per a l'emmagatzematge i intercanvi de recursos lingüístics d'aquest tipus. Un d'aquests és el descrit a [Ide *et al.*, 2000], que consisteix en el *Corpus Encoding Standard* (CES), però per a XML: el XCES.

Prenent aquest enfocaments com a referència i amb l'objectiu d'especificar un format adequat per als algorismes de la Traducció Automàtica estadística, hem construït una definició de tipus de document (DTD) basada en XML. Les implicacions d'utilitzar un format adaptat són les següents: a) les etiquetes són menys redundants; b) els elements tenen exactament el significat que necessitem; c) els usuaris del corpus marcat poden llegir-lo fàcilment.

Encara que la utilització de XML representa un consum d'espai de disc i memòria considerable, a més d'un esforç extra per portar les dades al format de XML, també ens aporta uns beneficis importants com ara disposar de documents accessibles per humans i màquines, poder accedir a les dades d'una manera transparenta, i gaudir d'usabilitat i portabilitat, de flexibilitat i escalabilitat, i d'efectivitat i robustesa.

Amb la nostra DTD, el corpus és un repositori de documents, és a dir, una col·lecció de *documents*. Un *document* és un concepte general que pot representar qualsevol tipus de component d'un corpus, com ara un ítem de notícies (en un corpus de notícies) o un diàleg (en un corpus de diàlegs). Cada *document* es pot dividir en *seccions*, com ara un paràgraf (en un corpus de notícies) o una intervenció (en un corpus de diàlegs). Cada *secció* consisteix en un o més *segments* (que poden ser o un passatge o una oració), i cada *segment* es pot dividir en *lsegments* (un segment per llengua). D'aquesta manera, es garanteix l'alineament a nivell de segment. Aquest és el requeriment mínim per fer d'aquestes dades multilingües un corpus valuós. La figura 1 ens dona un exemple de l'estructura més senzilla per a segments alineats, on els segments estan representats per cadenes de text.

La figura 2 mostra un exemple d'una estructura més complexa, amb segments que es descomponen en paraules que porten informació morfològica. Aquesta estructura és susceptible d'incorporar alineaments de nivell més baix, ja sigui a nivell de paraula o de grups sintagmàtics. La **L** es refereix al lema i la **P** a la categoria morfològica (*Part-of-Speech* – *POS*). Per l'etiquetatge dels corpus hem utilitzat dues eines, l'etiquetador d'en Eric Brill per a l'anglès [Brill, 1993] i l'analitzador morfològic MACO+ (*Morphological Analyzer Corpus*

Oriented) [Carmona *et al.*, 1998] i el desambiguador RELAX (*Relaxation Labelling Based Tagger*) [Padró, 1997] del TALP per al català i el castellà.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE lcstar SYSTEM "lcstar.dtd" []>
<DOC_REPOSITORY DATE="10/2/2004">
  <DOC NDOC="000">
    ...
    <SEC N="000" S="1">
      <SGM N="999">
        <LSGM LAN="CA">el dinou , d' aquest mes ?</LSGM>
        <LSGM LAN="EN">nineteenth , of this month ?</LSGM>
        <LSGM LAN="ES">? el diecinueve , de este mes ?</LSGM>
      </SGM>
    </SEC>
    ...
  </DOC>
</DOC_REPOSITORY>
```

Figura 1: Secció trilingüe alineada a nivell de segment

```
<SGM N="999">
<LSGM LAN="CA" S="1" W="7">
  <W L="el" P="DA0MS0">el</W>
  <W L="dinou" P="NCMS000">dinou</W>
  <W L="," P="Fc">,</W>
  <W L="de" P="SPS00">d'</W>
  <W L="aquest" P="DD0MS0">aquest</W>
  <W L="mes" P="NCMS000">mes</W>
  <W L="?" P="Fit">?</W></LSGM>
<LSGM LAN="EN" S="1" W="6">
  <W L="nineteen" P="JJ">nineteenth</W>
  <W L="," P=",">,</W>
  <W L="of" P="IN">of</W>
  <W L="this" P="DT">this</W>
  <W L="month" P="NN">month</W>
  <W L="?" P=".">?</W></LSGM>
<LSGM LAN="ES" S="1" W="8">
  <W L="?" P="Fia">?</W>
  <W L="el" P="DA0MS0">el</W>
  <W L="diecinueve" P="DN0CP0">diecinueve</W>
  <W L="," P="Fc">,</W>
  <W L="de" P="SPS00">de</W>
  <W L="este" P="DD0MS0">este</W>
  <W L="mes" P="NCMS000">mes</W>
  <W L="?" P="Fit">?</W></LSGM>
</SGM>
```

Figura 2: Segment trilingüe

4. CONCLUSIONS

Pensem que la creació de recursos lingüístics és una tasca molt important en l'àmbit de l'enginyeria lingüística que cal portar a terme malgrat el seu alt cost de desenvolupament. Especialment difícil és disposar d'aquest tipus de recursos per a llengües minoritàries, però sense ells no es pot avançar. Els beneficis són clars per aplicacions com la traducció automàtica. En el centre TALP s'han desenvolupat, i es desenvolupen, recursos textuais i orals tant per al català com per al castellà. Per a l'elaboració dels recursos per a la traducció, aquí descrits, s'ha fet una tasca d'especificació en quant a continguts i format amb l'objectiu de que puguin servir tant de referència per a la creació de recursos similars en altres llengües com de base per a la incorporació de més informació necessària per altres aplicacions. L'ús d'una DTD basada en el llenguatge XML contribueix de manera essencial a l'assoliment d'aquest objectiu.

AGRAÏMENTS

Aquesta recerca ha estat finançada pel programa de Tecnologies de la Societat de la Informació de la Unió Europea mitjançant el projecte LC-STAR (*Lexica and Corpora for Speech to Speech Translation Components*) (IST-2001-32216) i pel Ministerio de Ciencia y Tecnología mitjançant el projecte ALIADO (TIC2002-04447-C02).

REFERÈNCIES

- [Arranz *et al.*, 2003] Arranz, V., Castell, N. i Giménez, J. "Development of Language Resources for Speech-to-Speech Translation", RANLP-2003, Borovets, Bulgaria.
- [Arranz *et al.*, 2004] Arranz, V., Castell, N., Crego, J.M., Giménez, J., de Gispert, A. i Lambert, P. "Bilingual Connections for Trilingual Corpora: an XML Approach", LREC-2004, Lisboa, Portugal.
- [Bray *et al.*, 2000] Bray, T., Paoli, J., Sperberg-McQueen, C.M. i Maler, E. (eds) "Extensible Markup Language (XML) 1.0" (2a. Edició) Recomanació W3C, disponible a <http://www.w3.org/TR/REC-xml/>.
- [Brill, 1993] Brill, E. "A Corpus-Based Approach to Language Learning". Tesi Doctoral, Department of Computer and Information Science, University of Pennsylvania, Estats Units.
- [Carmona *et al.*, 1998] Carmona, J., Cervell, S., Màrquez, L., Martí, M.A., Padró, L., Placer, R., Rodríguez, H., Taulé, M. i Turmo, J. "An Environment for Morphosyntactic Processing of Unrestricted Spanish Text", LREC-1998, Granada.
- [de Gispert i Mariño, 2003] de Gispert, A. i Mariño, J.B. "Experiments in Word-Ordering and Morphological Preprocessing for Transducer-based Statistical machine Translation", ASRU-2003, St. Thomas, Estats Units.
- [Ide *et al.*, 2000] Ide, N., Bonhomme, P. i Romary, L. "XCES: An XML-based Encoding Standard for Linguistic Corpora", LREC-2000, Atenes, Grècia.
- [Padró, 1997] Padró, L. "A Hybrid Environment For Syntax-Semantic Tagging", Tesi Doctoral, Universitat Politècnica de Catalunya, Barcelona.
- [Toutanova *et al.*, 2002] Toutanova, K., Tolga Ilhan, H i Manning, C.D. "Extensions to HMM-based Statistical Word Alignment Models", EMNLP-2002, Filadelfia, Estats Units.
- [Ueffing i Ney 2003] Ueffing, N. i Ney, H. "Using POS Information for Statistical Machine Translation into Morphologically Rich Languages". EACL-2003. Budapest, Hungria.