

Creación de Recursos Lingüísticos para la Traducción Automática

Victoria Arranz, Núria Castell y Jesús Giménez

Centre de Recerca TALP
Universitat Politècnica de Catalunya
{varranz,castell, jgimenez}@lsi.upc.es

Resumen

En los últimos años la construcción de recursos lingüísticos se ha convertido en una tarea cada vez más importante dadas las múltiples áreas de la Ingeniería Lingüística para las que son necesarios. En este artículo nos centramos en la creación de recursos lingüísticos orientados a la traducción automática oral basada en técnicas estadísticas. Estos recursos están siendo desarrollados tanto para el castellano como para el catalán y el inglés. Uno de los objetivos ha sido que estos puedan servir como referentes para la creación de recursos similares en otras lenguas. Otro de los objetivos ha sido la creación de recursos flexibles que puedan incorporar nueva información necesaria para otras aplicaciones de ingeniería lingüística. Para alcanzar estos objetivos, se ha desarrollado una Definición de Tipo de Documento (DTD) basada en el lenguaje *Extensible Markup Language* (XML) y que permite la representación de corpora orales y textuales, monolingües y multilingües, generales y de dominio específico.

1.- Introducción

En los últimos años la construcción de recursos lingüísticos se ha convertido en una tarea cada vez más importante dadas las múltiples áreas de la Ingeniería Lingüística para las que son necesarios. Congresos como el LREC (*International Conference on Language Resources and Evaluation*)¹ demuestran el gran interés que existe por los recursos lingüísticos, su evolución y la necesidad de establecer criterios de evaluación de su calidad. Asimismo, aspectos como la creación de recursos para lenguas minoritarias también han ganado protagonismo en las últimas ediciones de este congreso. En este artículo nos centramos en la creación de recursos lingüísticos orientados a la traducción automática oral basada en técnicas estadísticas.

La traducción automática basada en el enfoque estadístico está demostrando buenos resultados, especialmente en dominios restringidos. Últimamente se está incorporando cierto conocimiento lingüístico a estos sistemas con el fin de mejorar los resultados de traducción (cf. [1], [2] y [3]). Se cree que a medida que se vaya utilizando esta información lingüística, contribuirá más a la mejora de los resultados. Por ejemplo, [4] presentan una mejora inicial en la

traducción estadística al integrar información como la categorización del tiempo y las expresiones de fechas, información que nosotros también hemos considerado e incluido en la Definición de Tipo de Documento (DTD) que se presenta en este trabajo.

La traducción automática estadística requiere corpora multilingües y paralelos para ser utilizados como entrenamiento de los sistemas automáticos de traducción. Adicionalmente pueden estar alineados (a nivel de frase, sintagma o palabra), posiblemente etiquetados con información lingüística, y ser generados a partir de transcripciones de grabaciones (cuando se trata de corpora orales). Estos recursos son costosos de crear y, especialmente, cuando se quiere trabajar con habla espontánea. Aunque es cierto que existen algunos recursos disponibles a través de organizaciones como el ELRA² y el LDC³, sólo algunas lenguas tienen una presencia importante en estos repositorios.

El centro TALP está trabajando en el ámbito de la traducción automática desde diferentes vertientes. En este trabajo nos centraremos en el desarrollo de recursos lingüísticos en el marco del proyecto europeo LC-STAR⁴, donde trabajamos tanto con el castellano como con el catalán y el inglés, y del proyecto español ALIADO⁵, donde trabajamos con el castellano y el inglés. Estos recursos se harán públicos a través de la organización ELRA, que se encargará de su distribución una vez hayan sido finalizados y validados.

2.- Corpora Desarrollados

En el marco de este proyecto se han creado dos tipos de recursos útiles para la traducción automática estadística. Estos recursos se centran en el dominio turístico. Por un lado, se han creado un corpus trilingüe (castellano, catalán e inglés americano) de diálogos y ocho corpora bilingües más pequeños. Todos los corpora bilingües tienen el inglés americano como una de las lenguas y la segunda lengua es una de las siguientes: alemán,

¹ <http://www.lrec-conf.org>

² <http://www.elra.info>

³ <http://www.ldc.upenn.edu>

⁴ <http://www.lc-star.com>

⁵ <http://gps-tsc.upc.es/veu/aliado/>

castellano, catalán, esloveno, finlandés, hebreo, italiano, ruso. En nuestro caso, hemos trabajado en las parejas inglés/castellano e inglés/catalán.

A su vez, también se ha llevado a cabo una tarea de diseño y especificación de cómo deben ser estos tipos de recursos (qué información deben contener, qué formato utilizar para su almacenado e intercambio, etc.). En los próximos meses se harán públicos los informes del proyecto donde se describen estas especificaciones.

2.1. Corpora Bilingües

Los corpora bilingües desarrollados consisten en una compilación de frases cortas o fragmentos de frases en inglés americano que han sido traducidos, en nuestro caso, al castellano y al catalán. Estos corpora tienen 10.500 entradas, están almacenados en un formato basado en XML e incorporan información lingüística a nivel de palabra (etiqueta morfosintáctica y lema).

Las frases utilizadas como base de estos corpora han sido extraídas de las siguientes fuentes:

- Los corpora de diálogos que describiremos a continuación.
- Un corpus de textos de información turística obtenidos en páginas web.
- Ejemplos creados manualmente a partir de libros de frases típicas (*phrasal books*) que utilizan los turistas para viajar a países extranjeros.

El criterio seguido para extraer las frases de los diálogos y de los textos ha sido el de alcanzar la cobertura de las palabras más frecuentes de ambos corpora.

La Figura 1 presenta un ejemplo de la parte creada a partir de los *phrasal books* una vez las frases han sido traducidas y previo a su almacenamiento en la DTD en formato XML (la DTD con formato XML se puede observar en la Sección 3):

10527	across from	enfrente de
10528	additional fee	suplemento
10529	adjoining rooms	habitaciones contiguas
10530	again and again	una y otra vez
10531	against the stream	contracorriente
10532	air-tight	hermético
10533	airline company	compañía aérea
10534	airport taxes	tasas de aeropuerto
10535	aisle seat	asiento de pasillo
10536	alarm clock	despertador
10537	all day along	todo el día

Figura 1: Ejemplos de los *phrasal books*

2.2. Corpus Trilingüe

El corpus trilingüe se ha creado a partir de dos fuentes [5]. Una primera parte se ha creado a partir del corpus VerbMobil⁶, que se desarrolló en Alemania en el marco de un proyecto nacional de traducción automática. La segunda parte del corpus (TALP-tourism) se ha creado a partir de cero, comenzando por la grabación de diálogos en castellano y catalán.

El corpus VerbMobil es un conjunto de diálogos relacionados con la concertación de reuniones, encuentros, etc. (*appointments*). Hemos seleccionado el material disponible en inglés americano y lo hemos traducido al castellano y al catalán. El corpus tiene marcas de fenómenos de habla espontánea como, por ejemplo, ruidos, repeticiones, interrupciones, etc. Se han elaborado unos criterios de traducción para tener en cuenta todos estos fenómenos y el objetivo de disponer de un recurso para sistemas de traducción estadística. El corpus seleccionado y traducido tiene más de 200.000 palabras, con un vocabulario de unas 3.300 palabras.

El corpus TALP-tourism se ha creado a partir de cero tomando como modelo el corpus VerbMobil, pero centrándose en el dominio turístico. Se han diseñado una serie de escenarios (situaciones típicas con las que se puede encontrar un turista) y se han grabado más de 350 diálogos (unas 600.000 palabras) utilizando locutores voluntarios. Esta base de datos oral creada contiene un vocabulario de unas 10.000 palabras en catalán y unas 11.000 en castellano. La Tabla 1 que se presenta a continuación muestra las cifras de la base de datos oral grabada.

	Castellano	Catalán
Tiempo de grabación	31h:7m:32s	23h:43m:55s
Nº de locutores	77	56
Nº de diálogos	217	172
Nº de turnos	10.998	9.321
Nº de oraciones	24.372	19.113
Nº de tokens	349.970	277.777
Nº de tokens diferentes	11.714	10.057

Tabla 1: Base de datos oral

Los escenarios básicos eran: hotel, agencia de viajes, oficina de información turística y compañía aérea o de trenes. Dentro de cada escenario había unos subescenarios más concretos que permitían ubicarse a los locutores como, por ejemplo, hacer una reserva de hotel o un billete de avión. Además, se disponía de unas plantillas que proporcionaban a los locutores descripciones ejemplo sobre las que elaborar sus propias

⁶ <http://verbmobil.dfki.de/verbmobil>

conversaciones. Un ejemplo de plantilla sería el siguiente:

Hotel Reservation

Speaker 0 actúa como un turista reservando alojamiento en un hotel determinado para una fecha determinada, un número de personas y bajo ciertas condiciones específicas.

Speaker 1 actúa como un empleado de hotel proporcionando la información requerida.

Estas grabaciones se han realizado en castellano y catalán. Una vez transcritos los diálogos, el texto en castellano se ha traducido al catalán y al inglés americano, mientras que el texto catalán se ha traducido al castellano y al inglés americano. Los criterios de traducción han sido los mismos que los utilizados para la traducción del corpus VerbMobil. La Figura 2 muestra un ejemplo del contenido de este corpus trilingüe tras la fase de traducción y antes de su almacenamiento en la DTD en formato XML. El código en negrita marca el número de la intervención dentro del diálogo, así como la lengua en que se encuentra:

00001_EN: the hotel La_Habana_Neptuno in Cuba is located in Havana . it 's right on the beach , thirty-five kilometers from the nearest airport .

00001_ES: el hotel Habana_Neptuno_de_Cuba está situado en La_Habana . está justo en la playa , a treinta y cinco kilómetros del aeropuerto más próximo.

00001_CA: l' hotel l' Havana_Neptuno_de_Cuba està situat a L'_Havana . està just a la platja , a trenta-cinc quilòmetres de l' aeroport més pròxim .

00002_EN: can you continue ?

00002_ES: ¿ puede continuar ?

00002_CA: pot continuar ?

00003_EN: it has five_hundred_sixty units . and each room has a bathroom , telephone , and a satellite T-V , among other things .

00003_ES: tiene quinientas sesenta unidades . y cada habitación dispone de baño , teléfono y tiene televisión por satélite , entre otros .

00003_CA: en té cinc-centes_seixanta unitats . i cada habitació disposa de bany , telèfon i té televisió per satèl·lit , entre altres .

00004_EN: great , great .

00004_ES: muy bien , muy bien .

00004_CA: molt bé , molt bé .

Figura 2: Ejemplo del contenido del corpus trilingüe

La Tabla 2 presenta las medidas de perplejidad del corpus trilingüe al completo.

Lenguas	Perplejidad
Castellano	23,721
Catalán	24,9373
Inglés	19,5766

Tabla 2: Medidas de perplejidad

3.- Enriquecimiento de los Recursos

El resultado de la tarea descrita en la sección anterior es la existencia de un corpus trilingüe que puede ser utilizado como tal o que se puede explotar como corpora bilingües diferentes. En ambos casos se puede incluir información lingüística utilizando los analizadores, etiquetadores y lematizadores que tenemos disponibles para las tres lenguas. En el caso de los corpora bilingües, se puede incorporar información de alineamiento entre ambas lenguas. El alineamiento a nivel de turnos de diálogo es simple y no presenta problemas, pero el alineamiento a nivel de sintagmas, locuciones o palabras es una tarea compleja. Cuanta más información tengamos sobre alineamiento entre dos lenguas, mejor funcionarán los sistemas estadísticos de traducción automática. En el marco del proyecto LC-STAR también se ha llevado a cabo una tarea de diseño de formatos [6] lo suficientemente flexibles como para permitir representar tanto los corpora bilingües, de menor tamaño, como los corpora trilingües, más grandes. El formato se basa en el XML [7] y está pensado para representar corpora de muchos tipos diferentes: mono/bi/multi-lingües, textuales u orales, de dominio restringido o generales, etc.

3.1. Diseño de la DTD: Formato y Contenido

Previo al diseño de nuestra DTD realizamos un estudio de los diferentes enfoques que se han seguido para el almacenamiento e intercambio de recursos lingüísticos de este tipo. Uno de estos es el descrito en [8], que consiste en el *Corpus Encoding Standard* (CES), pero para XML: el XCES.

Tomando estos enfoques como referencia y con el objetivo de especificar un formato adecuado para los algoritmos de la Traducción Automática estadística, hemos construido una DTD basada en XML. Las implicaciones de utilizar un formato adaptado son las siguientes:

- Las etiquetas son menos redundantes.
- Los elementos tienen exactamente el significado que necesitan.
- Los usuarios del corpus marcado pueden leerlo fácilmente.

A pesar de que la utilización de XML representa un consumo de espacio de disco y memoria considerable, así como un esfuerzo extra para portar todos los datos a XML, también nos aporta unos beneficios importantes, como por ejemplo el hecho de disponer de documentos accesibles por humanos y máquinas, poder acceder a los datos de una manera transparente y disfrutar de usabilidad y portabilidad, de flexibilidad y escalabilidad, y de efectividad y robusteza.

Con nuestra DTD, el corpus es un repositorio de documentos, es decir, una colección de *documents*. Un *document* es un concepto general que puede representar cualquier tipo de componente de un corpus, como por ejemplo un ítem de noticias (en un corpus de noticias) o un diálogo (en un corpus de diálogos). Cada *document* se puede dividir en *sections*, como son un párrafo (en un corpus de noticias) o una intervención (en un corpus de diálogos). Cada *section*, a su vez, consiste en uno o más *segments* (que pueden ser o un pasaje o una oración), y cada *segment* se puede dividir en *lsegments* (un segmento por lengua). De esta manera se garantiza el alineamiento a nivel de segmento. Este es el requisito mínimo para convertir estos datos multilingües en un corpus valioso. La Figura 3 nos proporciona un ejemplo de la estructura más sencilla para segmentos alineados, donde los segmentos están representados por cadenas de texto.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE lstar SYSTEM "lstar.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 3: Sección trilingüe alineada a nivel de segmento

La figura 4 muestra un ejemplo de una estructura más compleja, con segmentos que se descomponen en palabras que portan información morfológica. Esta estructura es susceptible de incorporar alineamientos de nivel más bajo, ya sea a nivel de palabra o de grupos

sintagmáticos. La **L** se refiere al lema y la **P** a la categoría morfológica (*Part-of-Speech – POS*). Para el etiquetado de los corpora hemos utilizado dos herramientas, el etiquetador *SVMTool* [9] para el inglés, y el analizador morfológico y desambiguador *FreeLing* [10] para el castellano y el catalán.

```
<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 4: Segmento trilingüe

4.-Beneficios para la Traducción Estadística

La investigación en traducción automática estadística necesita pasar por la costosa tarea de preparar datos de entrenamiento y de evaluación, la cual es muy compleja si tenemos en cuenta la naturaleza multilingüe de estos datos.

A esto debemos sumar el hecho de que estos recursos se enriquezcan y contengan información lingüística, la cual ayuda a mejorar los resultados de traducción (ver sección 1) y se espera que aún contribuya más a la mejora de estos sistemas a medida que se disponga de información como *named entities*, *semantic tags*, etc.

Sin embargo, para que esta información resulte de gran utilidad a los diferentes investigadores en traducción estadística, un factor a tener en cuenta es el formato en el que se va a almacenar. Es importante que se trate de

un formato estándar y flexible, y es por esto que el presente trabajo ha optado por un esquema basado en XML, el cual nos permite insertar una cantidad considerable de información y de una naturaleza compleja. Por ejemplo, nos permite llevar a cabo la integración de diferentes tipos de información como pueden ser etiquetas POS o alineamientos a nivel de palabra o de estructuras más complejas.

Para resumir, la propuesta de DTD que aquí se presenta proporciona una estructura rica y flexible que nos permite incorporar la información necesaria en el momento que se requiera dentro de un esquema estándar y portable y asimismo reutilizable.

5.- Conclusiones

La creación de recursos lingüísticos es una tarea extremadamente importante en el ámbito de la ingeniería lingüística. Por esta razón, es una tarea que debe realizarse a pesar de su alto coste de desarrollo debido a la necesaria intervención humana para obtener unos recursos de calidad. Su alto coste hace que resulte difícil disponer de este tipo de recursos para lenguas minoritarias. No es comparable la cantidad de recursos lingüísticos disponibles para el inglés con los disponibles para el catalán, por ejemplo. Pero es necesario asumir los costes para poder avanzar en el tratamiento automático de las lenguas minoritarias.

Aplicaciones como la traducción automática son un claro ejemplo de la necesidad de estos recursos. En nuestro centro TALP se han desarrollado, y se desarrollan, recursos textuales tanto para el castellano como para el catalán. Para la elaboración de los recursos para traducción, aquí descritos, se ha realizado una tarea de especificación en cuanto a contenido y formato de los recursos. Uno de los objetivos ha sido que estos puedan servir como referentes para la creación de recursos similares en otras lenguas. Otro de los objetivos ha sido la creación de recursos flexibles que puedan incorporar nueva información necesaria para otras aplicaciones de ingeniería lingüística. Para alcanzar estos objetivos, se ha desarrollado una DTD basada en el lenguaje XML y que permite la representación de corpora orales y textuales, monolingües y multilingües, generales y de dominio específico,...Está pendiente la incorporación de un mecanismo que permita la representación del alineado de corpora a nivel de palabra o grupos de palabras.

6.- Agradecimientos

Esta investigación está financiada por el programa de Tecnologías de la Sociedad de la Información de la Unión Europea mediante el proyecto LC-STAR (*Lexica and Corpora for Speech to Speech Translation Components*) (IST-2001-32216) y por el Ministerio de

Ciencia y Tecnología mediante el proyecto ALIADO (Tecnologías del Habla y el Lenguaje para un Asistente Personal) (TIC2002-04447-C02). Nuestro agradecimiento a David Conejero por su ayuda.

7.- Referencias

- [1] Ueffing, N. y Ney, H., "Using POS Information for Statistical Machine Translation into Morphologically Rich Languages", EACL-2003. Budapest, Hungría, 2003.
- [2] Toutanova, K., Tolga Ilhan, H y Manning, C.D., "Extensions to HMM-based Statistical Word Alignment Models", EMNLP-2002, Filadelfia, Estados Unidos, 2002.
- [3] Och, F.J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z. y Radev, D. Final Report of Johns Hopkins 2003 Summer Workshop on Syntax for Statistical Machine Translation, Estados Unidos, 2003.
- [4] de Gispert, A. y Mariño, J.B., "Experiments in Word-Ordering and Morphological Preprocessing for Transducer-based Statistical Machine Translation", ASRU-2003, St. Thomas, Estados Unidos, 2003.
- [5] Arranz, V., Castell, N. y Giménez, J. "Development of Language Resources for Speech-to-Speech Translation", RANLP-2003, Borovets, Bulgaria.
- [6] Arranz, V. Castell, N., Crego, J.M., Giménez, J., de Gispert, A. y Lambert, P. "Bilingual Connections for Trilingual Corpora: an XML Approach", 4th International Conference on Language Resources and Evaluation (LREC-2004), Lisboa, Portugal. 2004.
- [7] Bray, T., Paoli, J., Sperberg-McQueen, C.M. and Maler, E. (eds), "Extensible Markup Language (XML) 1.0" (2ª Edición) Recomendación W3C, 2000, también disponible en <http://www.w3.org/TR/REC-xml/>.
- [8] Ide, N., Bonhomme, P. and Romary, L., "XCES: An XML-based Encoding Standard for Linguistic Corpora", LREC-2000, Atenas, Grecia, 2000.
- [9] Jiménez, J. and Márquez, L. "SVMTool: A General POS Tagger Generator Based on Support Vector Machines". 4th International Conference on Language Resources and Evaluation (LREC-2004), Lisboa, Portugal. 2004.
- [10] Carreras, X., Chao, I., Padró, L., Padró, M. FreeLing: "An Open-Source Suite of Language Analyzers". 4th International Conference on Language Resources and Evaluation (LREC-2004). Lisboa, Portugal, 2004.