

N-II: TRADUCTOR AUTOMÁTICO ESTADÍSTICO BASADO EN NGRAMAS

Marta R. Costa-jussà, Mireia Farrús, Marc Poch, Adolfo Hernández y José B. Mariño

Centro de Investigación TALP-UPC,
Campus Nord, 08034 Barcelona
{mruiz,mfarrus,mpoch,adolfohh,canton}@gps.tsc.upc.edu

RESUMEN

Esta comunicación describe el desarrollo del traductor estadístico N-II entre catalán y castellano. Para mejorar la calidad del sistema, se llevó a cabo un riguroso análisis lingüístico. Este ha permitido plantear soluciones estadísticas y basadas en reglas que afrontan con éxito los errores más comunes de la traducción puramente estadística.

1. INTRODUCCIÓN

La traducción de voz a voz se hace a partir de la concatenación de tres sistemas: reconocimiento de voz, traducción de texto y síntesis de voz. La web N-II¹ es un ejemplo del segundo sistema y proporciona una traducción automática entre castellano y catalán en ambas direcciones siguiendo una aproximación estadística. La traducción automática estadística se basa en el hecho de que cada oración e en una lengua destino es una posible traducción de una oración f en una lengua fuente. La principal diferencia entre dos posibles traducciones de una oración dada es la probabilidad asignada a cada una que se tiene que aprender de un texto bilingüe. Por lo tanto, la traducción de una oración fuente f se puede formular como la búsqueda de la oración destino e que maximiza la probabilidad de traducción $P(e|f)$.

Dentro de los sistemas estadísticos, el N-II utiliza una aproximación basada en un modelo de lenguaje de unidades bilingües. Este sistema ha participado en varias evaluaciones de prestigio internacional obteniendo resultados competitivos, a título de ejemplo ver [2].

Para estimar los parámetros del modelo, la aproximación estadística (y como tal, la basada en Ngramas) requiere corpus bilingües paralelos (formados por pares de oraciones que se traducen mutuamente). Concretamente, para entrenar el traductor N-II, hemos utilizado el corpus paralelo del diario *El Periódico* que contiene 1.7 millones de oraciones.

El resultado del traductor estadístico ha obtenido resultados BLEU superiores al 80 % cuando se testea con

un texto del mismo dominio que el utilizado en el entrenamiento. Sin embargo, las traducciones generan diversos errores que hay que tener en cuenta y rectificar para aumentar la calidad del sistema.

En esta comunicación se presenta un análisis preliminar de los errores encontrados más frecuentes, y se proponen diferentes técnicas para resolverlos, como la incorporación de reglas o información morfológica adicional. Finalmente, también se hace una breve relación de casos problemáticos que han quedado por resolver.

Así pues, la sección 2 presenta el sistema básico del N-II. La sección 3 presenta el análisis lingüístico de los errores y la sección 4 las soluciones aplicadas que se han dividido en dos tipos: las que utilizan reglas basadas en información gramatical, y las que optan por un procesamiento directo del texto. Finalmente, la sección 6 presenta las conclusiones y el trabajo futuro.

2. SISTEMA BÁSICO DE TRADUCCIÓN

El modelo de traducción puede entenderse como un modelo de lenguaje de unidades bilingües (llamadas tuplas). Dichas tuplas definen una segmentación monótona de los pares de oraciones utilizadas en el entrenamiento del sistema (f_1^J, e_1^J) , en K unidades (t_1, \dots, t_K) . En la extracción de las unidades bilingües, cada par de oraciones da lugar a una secuencia de tuplas que solo depende de los alineamientos internos entre las palabras de la oración.

El modelo de traducción se ha implementado utilizando un modelo de lenguaje (bilingüe) basado en n -gramas de tuplas [1]. En la traducción de una oración de entrada, el decodificador debe encontrar la secuencia de tuplas asociada a una segmentación de la oración de entrada que produzca probabilidad máxima. En general, tal probabilidad máxima se calcula como combinación lineal de modelos.

En la traducción del catalán-castellano, dado que son un par de lenguas muy paralelas, la utilización de un único modelo (el de traducción) ya permite obtener un traductor estadístico competente. Hay que tener en cuenta que este modelo de traducción incluye el modelo de lenguaje de destino. En caso de utilizar un corpus monolingüe adicional motivaría incorporar un modelo adicional de destino. El sistema de búsqueda utilizado se ha desarrollado en la

Este trabajo ha sido subvencionado por el Gobierno Español mediante el proyecto coordinado AVIVAVOZ (TEC2006-13694-C03) y el Govern de la Generalitat de Catalunya mediante el proyecto TecnoParla. Asimismo, agradecemos la colaboración de Yesika Laplaza y Carlos Alberto Henríquez y los comentarios de los revisores.

¹<http://www.n-ii.org/>

UPC (MARIE²).

3. ANÁLISIS LINGÜÍSTICO DE ERRORES

El traductor estadístico N-II presentaba, en un análisis preliminar, una serie de errores, en una o ambas direcciones de traducción, que describimos muy brevemente a continuación.

En primer lugar, los errores encontrados al traducir del castellano al catalán fueron los siguientes:

Obligación El traductor generaba la traducción literal de *tener que* como **tenir que*, en lugar de *haver de*.

Omisión de la preposición *de* La preposición *de* se omite al traducir el verbo *deber*.

Solo Es un término que corresponde a tres categorías gramaticales diferentes: adverbio, adjetivo y nombre. Según la categoría le corresponde una traducción diferente al catalán, y presenta una gran confusión en el traductor, especialmente entre el adjetivo y el adverbio.

Apóstrofe No se cumplen las reglas de apostrofación del catalán para los artículos *el* y *la* y la preposición *de* delante de vocales.

Ele geminada (*ll*) Aunque debería escribirse siempre con punto volado, es muy habitual encontrar la *ele* geminada con punto normal, hecho que causa traducciones incorrectas.

En segundo lugar, los errores encontrados al traducir del catalán al castellano se resumen en la siguiente lista:

Preposiciones *a* y *en* Estas preposiciones tienen usos muy delimitados que no se corresponden con una traducción literal correcta al castellano.

Poseivos Adjetivos y pronombres posesivos tienen la misma forma en catalán, hecho que crea ambigüedad a la hora de traducir al castellano, que utiliza formas diferentes.

Perquè Esta conjunción tiene traducciones distintas al castellano en según si introduce una oración subordinada de causa (*porque*) o de finalidad (*para que*).

Soler Las formas conjugadas *sol* y *sols* del verbo *soler* pueden confundirse con adjetivos.

Conjunciones y *y* *o* Estas dos conjunciones deben transformarse en *e* y *u* cuando preceden palabras que empiezan por *y* y *o*, respectivamente.

Omisión de la preposición *a* A diferencia del catalán, el castellano utiliza habitualmente la preposición *a* delante del objeto directo. Al no encontrarse en la lengua fuente, tampoco aparece en la lengua origen.

Finalmente, los errores encontrados en ambas direcciones fueron los siguientes:

Concordancia de género Una palabra femenina (masculina) en castellano se puede corresponder con una palabra masculina (femenina) en catalán (p.ej. *la señal* - *el senyal*).

²<http://gps-tsc.upc.es/veu/soft/soft/marie/>

Números Hay números que no aparecen en el corpus de entrenamiento, por lo que no se genera ninguna traducción.

Horas Las expresiones de las horas en castellano y en catalán son, formalmente, diferentes. Por consiguiente, la traducción, en muchos casos, no es literal. La diferencia principal es la utilización de los cuartos: mientras el castellano se expresa mediante los cuartos que *pasan* de una determinada hora, en catalán se habla de los cuartos que se *acercan* a la hora siguiente: *Las cuatro y cuarto* se traduce por *Un quart de cinc*.

Clíticos Con frecuencia, el traductor omite los pronombres personales adheridos al verbo. En otras ocasiones, aunque la traducción de los pronombres personales sea la correcta, el error se encuentra, a menudo, en una combinación incorrecta del pronombre con el verbo en cuestión.

Palabras desconocidas Hay palabras que el corpus contiene únicamente al inicio de oración; por consiguiente, estas palabras solo se encuentran en mayúscula, lo que implica que la misma palabra escrita en minúsculas aparezca como desconocida.

4. SOLUCIONES APLICADAS

Para la solución de algunos de los problemas descritos en la sección anterior se han aplicado dos tipos de técnicas: las técnicas basadas en la utilización de la categoría gramatical de las palabras, y las técnicas basadas en la corrección mediante un procesado directo del texto. Para la evaluación de las soluciones se ha realizado un análisis humano. En la Tabla 1 se muestran algunos ejemplos en los cuales dado un enunciado (O) se compara su correspondiente traducción antes (T1) y después (T2) de utilizar las técnicas descritas.

4.1. Reglas que utilizan la categoría gramatical

Las categorías gramaticales se han incorporado con éxito en traducción estadística para tratar problemas como el reordenamiento [4] y el análisis automático de los errores [5]. El objetivo es adjuntar la categoría gramatical (*tag*) correspondiente a la palabra a tratar, de manera que el modelo estadístico sea capaz de distinguir las palabras en función de su categoría y aprender el contexto.

4.1.1. Desambiguación de la homonimia

A menudo encontramos dos palabras iguales en la lengua origen que no lo son en la lengua destino y que causan traducciones incorrectas. Si las palabras son homónimas y se diferencian por su categoría gramatical, ésta podrá utilizarse para desambiguarlas.

En el caso del *solo*, se diseñan unas reglas que nos identifiquen, en los casos dudosos, si es un adverbio o un adjetivo. Se aplican las reglas en la lengua origen y se

adjunta el *tag* a la palabra. Así pues, una oración de la lengua fuente como *Venía solo*. se modifica a *Venía solo_<ADJ>*.), de manera que el modelo estadístico será capaz de distinguir ambos casos.

Un proceso similar se realiza para los posesivos del catalán: se han diseñado unas reglas que permiten etiquetar la palabra como adjetivo o pronombre posesivo y las etiquetas se incorporan posteriormente a la lengua origen. En el caso del *soler*, en lugar de generar unas reglas para detectar que *sol/sols* son verbos, se ha adjuntado el *tag* correspondiente que proporciona el Freeling [3].

4.1.2. Categorización

En una frase a traducir pueden aparecer números que no existen en el corpus de entrenamiento y, entonces, estas palabras son desconocidas y no se traducen. Para evitar este problema, hemos planteado unas reglas que detectan los números en la lengua origen, los codifican y los generan en la lengua destino. Para detectar los números hay que tener en cuenta su estructura (p.ej. palabra compuesta, con o sin guión) y, como dificultad añadida, se tiene que considerar que los números pueden tener género. Se define una codificación concreta para que en el momento de la generación se sea coherente con el número que se ha detectado. No se han categorizado los números: *un/una, dos/dues, nou* y *deu* por ser palabras que no son siempre números.

Por otro lado, la expresión de las horas en catalán y en castellano es distinta (como se ha explicado en 3) y el hecho de que el corpus contenga escasos ejemplos de horas puede generar errores en la traducción. Se ha optado por el mismo planteamiento que con los números: se detectan las horas, se codifican y se generan. Para detectar las horas primero hay que identificar su estructura, teniendo en cuenta que hay varias estructuras posibles para una misma hora. Asimismo, cara hora puede tener un contexto que se modifica en traducción. Por ejemplo: *Són dos quarts de dues* se traduce por *Es la una y media*. En este caso, también se modifica la forma verbal (pasa de 3a persona del plural a 3a persona del singular). Así pues, en este ejemplo, la frase entera se detecta toda como una hora. Se codifica de manera que se mantenga la información necesaria para poder generar la hora coherentemente.

4.1.3. Clíticos

En la lengua fuente, los clíticos se detectan y se separan del verbo utilizando el Freeling. Tras la traducción, deben juntarse de nuevo con el verbo. Este proceso de combinación se trata con unas reglas que tienen en cuenta dos factores; en primer lugar, las reglas de acentuación en castellano, ya que la posición de la sílaba tónica cambia al adherir un pronombre enclítico al verbo: *vende + lo* → *véndelo*. En segundo lugar, las reglas de combinación de los pronombres en catalán³ que, a diferencia del castellano,

se escriben con guión o apóstrofe según el caso, y no se alteran las reglas de acentuación: *seguir + lo* → *seguir-lo*; *compra + el* → *compra'l*; y *el + aixecava* → *l'aixecava*.

4.1.4. Apóstrofe

La apostrofación en catalán sigue, en general, una regla básica: se apostrofan los artículos *el, la* y la preposición *de* cuando preceden palabras que empiezan por vocal o *h muda*: *el arbre* → *l'arbre*; *la hora* → *l'hora*; y *de eines* → *d'eines*.

A esta regla se le aplican excepciones⁴:

- No se apostrofan delante de palabras que empiezan por *i* o *u* semiconsonánticas: *el uombat, la hiena, de iogurt*.
- No se apostrofa el artículo femenino delante de palabras que empiezan por *i* o *u* átonas (incluyendo la *h muda*): *la universitat, la Irene*.
- No se apostrofan ni el artículo femenino ni la preposición delante del prefijo negativo *a*: *la anormalitat, de asimètric*.
- No se apostrofan *la una* (hora), *la ira, la host* y los nombres de letra (*la e, la hac, la erra, etc.*).

4.1.5. Tratamiento de mayúsculas a inicio de oración

Para minorizar el problema genérico (al que se enfrenta cualquier traductor basado en corpus) de las palabras desconocidas, se ha propuesto una técnica que utiliza información morfológica. Concretamente, se trata de pasar a minúsculas todas aquellas palabras a inicio de oración a excepción de nombres propios, nombres comunes y adjetivos, ya que estas palabras son susceptibles de ser un nombre propio. De esta forma, las palabras que solo estaban en mayúscula en el corpus de entrenamiento y, por lo tanto, en minúscula eran desconocidas, tienen traducción.

4.1.6. Tratamiento de las concordancias

Afrontamos la concordancia de género mediante la utilización de un modelo de *tags* de la lengua destino. Esto permite beneficiar aquellas secuencias de palabras que mantienen coherencia en género, por ejemplo: será más probable una secuencia tal y como *pilota verda* que *pilota verde* porque el modelo de *tags* ha visto más veces un nombre femenino seguido de un adjetivo femenino que un nombre femenino seguido de un adjetivo masculino. El modelo de *tags* podrá ayudar en la medida que el modelo de traducción, es decir, las tuplas lo permitan. Por ejemplo, la traducción de *senyal blanc* continúa siendo *señal blanco* porque en el corpus no existe la tupla *blanc#blanca*.

Asimismo, como existe una tendencia a omitir palabras utilizamos una bonificación de palabras.

³www.cpnl.cat/media/upload/pdf/cnlortografia0305_editora_grup_30_19.pdf

⁴<http://www.uoc.edu/serveilinguistic/criteris/ortografia/apostrof.html>

posesivos	(O) Els meus amics no són els teus . (T1) Mis amigos no están *tus . (T2) Mis amigos no son los tuyos .
solo	(O) Era solo un niño. (T1) Era *sol un nen. (T2) Només era un nen.
sol/sols	(O) La Creu Roja sol disposar de quatre. (T1) La Cruz Roja *solo disponer de cuatro. (T2) La Cruz Roja suele disponer de cuatro.
ele	(O) S'ha reformat a Brussel.les .
geminada	(T1) Se ha reformado en *Bruselas. las . (T2) Se ha reformado en Bruselas .
obligación	(O) Nos lo tenemos que crear. (T1) Ens ho *tenim que creure. (T2) Ens ho hem de creure.
y/o	(O) Com a Blanes o Olot. (T1) Como Blanes *o Olot. (T2) Como Blanes u Olot.
clíticos	(O) No quiero verte más por aquí. (T1) No vull veure et més per aquí. (T2) No vull veure t més per aquí.
apóstrofe	(O) La acepta hasta el final. (T1) *La acepta fins al final. (T2) Lácepta fins al final.
números	(O) Ha aprovat l'alliberament de quatre-cents quaranta-un presoners. (T1) Ha aprobado la liberación de *quatre-cents quaranta-un prisioneros. (T2) Ha aprobado la liberación de cuatrocientos cuarenta y un prisioneros
horas	(O) Són tres quarts de vuit . (T1) Son *tres cuartos de ocho . (T2) Son las ocho menos cuarto .
mayúsculas	(O) No entenc per què no hi assisteixes . (T1) No entiendo por qué no *assisteixes . (T2) No entiendo por qué no asistes .

Tabla 1. Ejemplos de corrección de errores.

4.2. Procesado directo del texto

Algunos errores precisan de un preprocesado directo en el texto antes o después de realizar la traducción. La *ele* geminada se ha tratado antes de la traducción, normalizando la escritura del punto utilizado. En otros errores como la obligación *tener que* y las conjunciones *y/o* se han tratado como postprocesado después de realizar la traducción.

5. EVALUACIÓN

Para la evaluación de las técnicas que hemos incorporado usamos el test de *El Periódico* de 2000 oraciones y una sola referencia. La Tabla 2 compara la eficacia en BLEU de todas las mejoras de este artículo juntas frente a un sistema de referencia que no las incorporaba.

	es > ca	ca > es
Sistema de referencia	76.66	76.98
+ mejoras planteadas	82.28	81.74

Tabla 2. Resultados BLEU en ambas direcciones de traducción.

Hemos de tener en cuenta que lo que tiene más significancia estadística en este test es el tratamiento de los clíticos, puesto que en el sistema de referencia no se han tratado de ningún modo. Además, hemos planteado problemas que el usuario de un traductor podía encontrar y que no se suelen encontrar en textos periodísticos.

6. CONCLUSIONES Y TRABAJO FUTURO

En esta comunicación se ha descrito el traductor estadístico N-II. Se ha detallado el resultado de un análisis lingüístico de errores y se han propuesto soluciones basadas en reglas y de carácter estadístico así como pre y postprocesos de corrección de errores. La eficiencia de las soluciones aportadas se ha demostrado con ejemplos significativos y con una evaluación automática.

7. BIBLIOGRAFÍA

- [1] Mariño, J.B., Banchs, R.E., Crego, J.M., de Gispert, A., Lambert, P., Fonollosa, J.A.R. y Costa-jussà, M.R. *N-gram Based Machine Translation*. Computational Linguistics, 32:4:527–549, 2006.
- [2] Lambert, P., Costa-jussà, M.R., Crego, J.M., Khalilov, M., Mariño, J.B., Banchs, R.E., R. Fonollosa, J.A. y Schwenk, H. *The TALP Ngram-based SMT System for IWSLT 2007*. En Proc. of the International Workshop on Spoken Language Translation (IWSLT) pp 169-174, Trento, 2007.
- [3] Carreras, X., Chao, I., Padró, L., Padró, M., *FreeLing: An Open-Source Suite of Language Analyzers*, En Proc. of the Conference on Language Resources and Evaluation, LREC. Lisboa, 2004.
- [4] Crego, J.M. y Mariño, J.B. *Improving SMT by coupling reordering and decoding* En Machine Translation, 20:3:199–215, 2007.
- [5] Popović, M., de Gispert, A., Gupta, D., Lambert, P., Ney, H., Mariño, J.B. y Banchs, R. *Morpho-syntactic Information for Automatic Error Analysis of Statistical Machine Translation Output* En Proc. of the HLT/NAACL Workshop on Statistical Machine Translation, pages 1-6, New York, 2006.