Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 77-84 recibido 27-04-2005; aceptado 01-06-2005

Traducción automática estadística basada en n-gramas

Antonio Oliver, Toni Badia, Gemma Boleda, Maite Melero
Universitat Pompeu Fabra
Passeig de Circumval·lació 8, 08003 Barcelona
{antonio.oliver,toni.badia,gemma.boleda,maite.melero}@upf.edu

Resumen: En este artículo presentamos un sistema experimental de traducción automática de tipo estadístico basado en n-gramas. El sistema utiliza un corpus paralelo y fue concebido inicialmente como una extensión de un sistema de Traducción Asistida (TAO). Los buenos resultados obtenidos para el par de lenguas catalán-castellano nos han impulsado a explorar su utilización como sistema de traducción automática, comparándolo con una memoria de traducción comercial y con un sistema de traducción automática convencional.
Palabras clave: traducción automática estadística

Abstract: In this paper we present an experimental statistical machine translation system based on n-grams extracted form a parallel corpus. The system was conceived as an extension of a CAT system. However, its good performance in the language pair Catalan-Spanish has driven us to explore its use as a machine translation system, comparing it to a commercial translation memory and a conventional machine translation system.
Keywords: statistical machine translation

1 *Introducción*

En este artículo describimos un sistema experimental de traducción automática de tipo estadístico basado en n-gramas. La metodología se diseñó originalmente como una extensión de una aplicación basada en memorias de traducción. Sin embargo los resultados obtenidos en la traducción de textos entre catalán y castellano son cercanos a los sistemas de traducción automática comerciales, y claramente superiores a las memorias de traducción. Ello nos ha impulsado a seguir desarrollándolo incorporando diversas mejoras y optimizaciones. El prototipo inicial se ha incorporado al módulo de traducción de un sistema multilingüe de subtitulación automática desarrollado en el contexto del proyecto europeo e-Title.[1]

El resto del artículo se estructura de la manera siguiente. En el apartado 2 se describe el prototipo, que en esta fase de desarrollo utiliza texto no procesado lingüísticamente. En el apartado 3 se presenta una evaluación de este prototipo, utilizando una medida automática (BLEU) y comparando sus resultados con un sistema de traducción automática convencional (Internostrum) y una memoria de traducción comercial (DéjàVu). En el apartado 4 se describen dos direcciones en las que está previsto modificar el prototipo inicial. En la primera de ellas, se pretende optimizar sus resultados lematizando y etiquetando morfosintácticamente el corpus de aprendizaje, así como el texto a traducir. En la segunda, se sustituye el corpus paralelo como principal fuente de conocimiento por un diccionario bilingüe y un corpus monolingüe de la lengua de destino. Esta modificación se realizará en el contexto del proyecto europeo de Traducción Automática Metis-II.[2]

2 *Características y funcionamiento del prototipo*

El sistema está formado por los siguientes componentes:

- Un corpus bilingüe paralelo, previamente alineado.
- Un pequeño diccionario bilingüe.
- El algoritmo de traducción, implementado en Perl.

[1] European Multilingual Transcription and Subtitling Services for Digital Media Content (22160Y3C3DMAL2). http://www.etitle.co.uk/

[2] IST-FP6-003768 http://www.ilsp.gr/metis2

ISSN: 1135-5948

2.1 Corpus paralelo de entrenamiento

El corpus paralelo catalán-castellano con el que se ha entrenado el sistema consiste en 263.721 segmentos, que en total suponen unos 7 millones de palabras por cada lengua. El corpus procede de dos fuentes distintas:

- la versión bilingüe del diario "El Periódico de Catalunya";
- la versión bilingüe del "Diari Oficial de la Generalitat de Catalunya".

Los textos se han obtenido mediante descarga automática de los archivos html. El formato de los archivos nos ha permitido alinearlos mediante una estrategia muy sencilla basada principalmente en marcas html. La alineación se ha realizado a nivel de segmento, mediante unas reglas de segmentación sencillas e iguales para ambos idiomas, que se han aplicado sobre los párrafos detectados mediante las marcas de formato. Para poder aumentar el corpus paralelo en un futuro, se están explorando también diversos algoritmos más complejos de alineación automática de documentos (ver apartado 4).

2.2 Diccionario bilingüe

Para mejorar el rendimiento del sistema se ha compilado a mano un pequeño diccionario bilingüe, de sólo 52 entradas, que continene palabras o expresiones multipalabra muy frecuentes con sus respectivas traducciones. El diccionario que utilizamos en este experimento contiene principalmente preposiciones, artículos, conjunciones y posesivos. El uso de este diccionari permite solucionar casos como *el seu* $\rightarrow$ *suyo* y *als* $\rightarrow$ *a los*.

2.3 El algoritmo de traducción

El sistema aquí presentado se enmarca dentro de la Traducción Automática Estadística (SMT), porque se basa en un corpus paralelo y no utiliza información de tipo lingüístico. Sin embargo, a diferencia de los algoritmos tradicionales de traducción estadística, en los que se calcula un modelo de traducción y un modelo de la lengua destino [Brown et al.1993], nuestro sistema únicamente calcula un modelo de traducción. Así, se podría considerar una generalización de los algoritmos de extracción de lexicones bilingües [Hiemstra1998].

El modelo de traducción se calcula para todos los n-gramas (hasta un orden n determinado) que aparezcan un mínimo número de veces en el corpus paralelo. Esto lo acerca también a los sistemas de Traducción Automática Basada en Ejemplos (EBMT), especialmente cuando se aplica el algoritmo sobre un corpus etiquetado (ver apartado 4). Sin embargo, dichos sistemas suelen utilizar una maquinaria lingüística más potente para las dos lenguas implicadas en la traducción [Turcato y Popowich2001].

El algoritmo de traducción sigue los siguientes pasos:

- Se calcula cada n-grama, empezando por el valor más alto de n (por ejemplo, 4).
  - Si este n-grama aparece en el diccionario bilingüe, utilizamos esta traducción.
  - Si no, verificamos si aparece al menos 5 veces en el corpus[3] y calculamos la traducción más probable. Para ello se utiliza la siguiente metodología:
    - Por cada segmento donde aparece el n-grama original se considera que su posible traducción es cualquier n_2-grama (con $n-1 \leq n_2 \leq n+1$) del correspondiente segmento traducido.[4]
    - Se cuenta el número de veces que aparece cada candidato y se asigna una pseudoprobabilidad a cada candidato basada en este contaje. El candidato con la mayor probabilidad (si es que existe alguno) es la traducción que finalmente se utilizará. [5]
- Cuando se finaliza con todos los n-gramas con un valor dado de n, se proce-

[3] Por el momento se ha fijado provisionalmente en 5 el número mínimo de veces que debe aparecer un determinado n-grama. Tenemos previsto estudiar la influencia de este parámetro de manera que se pueda establecer un valor óptimo.

[4] Por tanto, se parte de la base de que la secuencia de palabras traducida tendrá una longitud similar a la de la secuencia original: el mismo número de palabras, una palabra más, o una palabra menos.

[5] El algoritmo, así, no sigue una estrategia direccional (p. ej. de izquierda a derecha), sino que elige en cada paso la traducción más probable para determinados n-gramas, independientemente de su posición. Los n-gramas que se han fijado no se vuelven a evaluar.

de a traducir los fragmentos no traducidos de la frase de entrada, recalculando los n-gramas de los fragmentos no traducidos con un valor de n decrementado en una unidad.

A continuación ilustraremos el funcionamiento del algoritmo mediante la traducción de una frase de ejemplo: *El president del Govern presentarà avui la seva dimissió.*

- $n = 4$: El único 4-grama que aparece más de 5 veces en el corpus es *El president del Govern*, y el algoritmo le asigna la traducción *El presidente del Gobierno.* Queda pues, por traducir *presentarà avui la seva dimissió*
- $n = 3$: El sistema encuentra los siguientes trigramas, *avui la seva* que traduce por *hoy su* y *la seva dimissió* que traduce por *su dimisión.* Estos dos trigramas entran en conflicto, ya que se superponen. Dado que la pseudoprobabilidad asiganada al primero es mayor, se utilizará únicamente este. Una vez substituido en la frase nos quedan por traducir dos fragmentos, *presentarà* y *dimissió*
- $n = 2$: No existe ningún bigrama adecuado, ya que lo que nos queda por traducir son dos unigramas
- $n = 1$: El sistema encuentra la traducción de *presentarà* y de *dimissió* en el corpus de entrenamiento, y los substituye directamente en la frase.

La traducción resultante es *El presidente del Gobierno presentará hoy su dimisión.*

Esta metodología básica de traducción funciona bien, como demostramos en el siguiente apartado, para la traducción entre catalán y castellano, donde son poco frecuentes los cambios de orden a larga distancia.

3 Evaluación del prototipo

3.1 Metodología

Para evaluar el prototipo, se ha comparado los resultados obtenidos por este, sobre un corpus de prueba, con los obtenidos por un sistema de memoria de traducción comercial y los obtenidos por un sistema de traducción automática convencional. Como sistema de memoria de traducción, se optó por DéjàVu,[6] uno de los sistemas más utilizados del mercado. En cuanto al sistema de TA se escogió Internostrum [Canals-Marote et al.2001], tanto por su disponibilidad como porque su calidad no se aleja demasiado de la de otros sistemas más complejos y costosos [Tomás, Mas, y Casacuberta2003, Guillén y Iturraspe-Bellver2001]. Nuestro sistema debería ser mejor que una memoria de traducción al uso, ya que permite un mejor aprovechamiento de la información por segmentos (en este caso, n-gramas). En cambio, no esperamos que, en la fase actual de desarrollo, alcance el nivel de aceptabilidad de un sistema de traducción basado en reglas como Internostrum.[7]

El corpus de prueba consiste en un ejemplar de *El Periódico de Catalunya* (32.077 palabras, 1.192 segmentos) y uno del *Diari Oficial de la Generalitat de Catalunya* (6.574 palabras, 517 segmentos), ambos del 15 abril de 2005. Ninguno de estos dos ejemplares pertenece al corpus de entrenamiento.

Para la comparación con DéjàVu, se ha traducido (o más precisamente, pretraducido, ya que no se revisa la traducción) el mismo corpus de prueba, utilizando una memoria de traducción formada por el corpus de entrenamiento. Las opciones seleccionadas para llevar a cabo la pretraducción son: *Assemble from portion* y *Insert source text for failed portion*, es decir ensamblar fragmentos, e insertar texto original en caso de fragmentos no hallados en la memoria.

Para la traducción con Internostrum, se ha utilizado la versión en línea que dicho sistema ofrece en su página web.[8]

Los resultados de cada uno de los tres sistemas han sido evaluados de forma automática utilizando la métrica BLEU [Papineni et al.2002], una de las metodologías de evaluación automática más conocidas y que obtiene mejor correlación con las evaluaciones humanas. La evaluación se ha llevado a cabo mediante el programa MTEval del NIST.[9]

BLEU utiliza un corpus de referencia (constituido por traducciones humanas de buena calidad) y una métrica numérica que

[6] http://www.atril.com

[7] Internostrum utiliza una estrategia de transferencia morfológica avanzada basada en estados finitos.

[8] http://www.internostrum.com

[9] National Institute of Standards and Technology http://www.nist.gov. El programa está disponible en http://www.nist.gov/speech/tests/mt/mt2001/resource/.

mide la proximidad de la frase traducida automáticamente con las de referencia, mediante la comparación de los n-gramas que las componen, sin tener en cuenta el orden. Puesto que normalmente existe más de una traducción posible para cada frase, este método requiere idealmente usar más de una frase de referencia para cada frase traducida. Sin embargo aquí lo hemos aplicado con una sola referencia por traducción (es decir, la versión bilingüe del documento original).[10] La métrica va desde 0, que indica gran distancia entre traducción y referencia, hasta 1, que indica identidad absoluta.

3.2 Resultados

En la Tabla 1 se pueden observar los resultados de la evaluación para cada sistema, sobre los ejemplares de *El Periódico* y del *Diari Oficial de la Generalitat de Catalunya* (*DOGC*) evaluados.

	DOGC	Periódico	*Media*
DéjàVu	0,61	0,23	*0,42*
n-gramas	0,74	0,57	*0,65*
Internostrum	0,81	0,77	*0,79*

Tabla 1: Evaluación de los sistemas de TA catalán-español

El patrón de los resultados es consistente con lo que se esperaba: según la media aritmética del coeficiente BLEU en los dos documentos, Internostrum es el mejor sistema (0,79), seguido del prototipo de sistema basado en n-gramas presentado en este artículo (0,65), y en último lugar aparece la aplicación de memorias de traducción (0,42).

Aunque los tres sistemas tienen un BLEU superior para el DOGC, debido seguramente a la mayor repetitividad y la menor variabilidad sintáctica de este tipo de texto respecto al texto periodístico, es la memoria de traducción la que evidencia una mayor disparidad entre los dos tipos de texto: 0,23 para el texto periodístico y 0,61 para el DOGC, es decir, una diferencia de 0,38.[11]

El hecho de que el sistema basado en n-gramas tenga un grado de corrección más similar en los dos tipos de texto (la diferencia entre los dos coeficientes es de 0,17) indica que es de aplicación más general que las memorias de traducción, y que sus resultados son más estables e independientes del grado de repetición del texto. Los resultados de este sistema, especialmente sobre el corpus de prueba procedente de *El Periódico*, claramente superiores a los de DéjàVu, confirman nuestra predicción inicial de que su mayor capacidad de generalización le conferiría ventajas sobre las memorias de traducción tradicionales. Estas, aun usando la opción *Assemble from portions*, requieren la presencia de fragmentos coincidentes más largos en la memoria.

Con respecto al sistema de traducción automática Internostrum, los resultados del sistema basado n-gramas son más bajos, sobre todo en un género más libre como el periodístico. Sin embargo, en el *DOGC* los resultados son parecidos (0,74 vs. 0,81), e incluso en *El Periódico* son notablemente cercanos (0,57 vs. 0,77), sobre todo si se tiene en cuenta la gran diferencia en tiempo de desarrollo de los dos tipos de sistema.

3.3 Análisis de los resultados

Más allá de los resultados en términos absolutos de corrección, nos interesaba comparar los resultados obtenidos por cada uno de los tres sistemas, así como una evaluación somera de los tipos de errores cometidos por cada uno. En primer lugar queríamos evaluar si los tres sistemas cometían errores en fragmentos distintos o en los mismos fragmentos, es decir, si coincidían en el nivel de dificultad que suponen los distintos fragmentos. Para ello, realizamos un análisis de la correlación entre los valores BLEU de los tres sistemas (tomando como referencia el coeficiente de cada segmento, información provista por el programa MTEval). Dado que los coeficientes no presentaban una distribución normal, utilizamos el test correlación de rango de Spearman (intervalo de confianza: 95%, alternativa: correlación positiva). El resultado está reflejado en la Tabla 2. [12]

Como vemos, los coeficientes de correlación ρ no son altos, sobre todo en el caso del DOGC. Sin embargo, en todos los casos

[10] Cuantas más oraciones de referencia por frase traducida se utilizan, mayor suele ser la puntuación obtenida, por lo tanto esperamos que los resultados de nuestra comparativas sean todos uniformemente más bajos de lo que serían evaluados usando más traducciones de referencia.

[11] La repetitividad en el DOGC es extremadamente elevada; hay párrafos enteros con fórmulas jurídicas que se repiten con mínimas variaciones, como fechas o nombres de lugares.

[12] En esta tabla, DV = DéjàVu, n = n-gramas, IN = Internostrum.

	DOGC		Periódico	
	ρ	p	ρ	p
DV vs. n	0,20	<0,001	0,34	<0,001
n vs. IN	0,32	<0,001	0,33	<0,001
IN vs. DV	0,14	<0,001	0,20	<0,001

Tabla 2: Test de correlación entre coeficientes BLEU de distintos sistemas

el test de correlación da resultados altamente significativos ($p < 0{,}001$), lo que indica que la correlación positiva detectada es estadísticamente relevante. Así, pues, los tres sistemas coinciden bastante en los segmentos que traducen bien y los que traducen mal, o, dicho de otro modo, en el nivel de dificultad de los fragmentos.

Para analizar de manera cualitativa el tipo de errores que comete nuestro prototipo en comparación con un sistema de traducción automática convencional, se extrajeron los segmentos del corpus de evaluación que cumplían una de las dos condiciones siguientes:

- el prototipo de sistema de n-gramas tiene un BLEU superior al de Internostrum en más de 0.1. Esto corresponde a 79 fragmentos del DOGC y 25 de El Periódico.
- Internostrum tiene un BLEU superior al del prototipo en más de 0.3. Son 84 fragmentos del DOGC y 383 de El Periódico.

Un análisis manual de los datos reveló que el sistema de n-gramas tiene mejor actuación básicamente en dos aspectos. El primero son los casos en que hay terminología (en un sentido amplio) propia del campo o de la institución de la que proviene el corpus, como en los ejemplos consignados en la Tabla 3.

Como cabe esperar, un sistema entrenado sobre un corpus paralelo se adapta mejor al uso léxico de cada campo o incluso corpus específico que un sistema que utiliza un diccionario bilingüe genérico. Lo que no es tan esperable a priori es que el sistema de n-gramas también resuelva mejor fenómenos correspondientes a propiedades léxicas generales de la lengua en cuestión. Este es el segundo aspecto en que nuestro sistema aventaja a Internostrum en muchos casos.

En efecto, el sistema de n-gramas traduce la subcategorización verbal de manera más correcta: por ejemplo traduce 'publicat a' por *publicado en* (Internostrum: *publicado a*). También trata mejor otros casos de régimen preposicional no estrictamente relacionados con la subcategorización, como muchos casos de *para* vs. *por* (en catalán, ambos son 'per'). Éste es el caso de la frase 'el Govern nord-americà usa el pretext de la seguretat nacional **per** augmentar el volum d'informació confidencial' (*el Gobierno estadounidense usa el pretexto de la seguridad nacional* **para** *aumentar el volumen de información confidencial*), traducido por Internostrum como *por aumentar* y por el sistema de n-gramas como *para aumentar*.

Por otro lado, el sistema de n-gramas, al carecer de mecanismos lingüísticos de control global de la gramaticalidad y coherencia de las oraciones, comete errores que son imposibles en Internostrum, como por ejemplo repetición de palabras (p.e. 'se especifican en en el, que pertenecen ...')[13] o problemas de orden de palabras o de coaparición (p.e. 'pero vecinos los edificios más afectados ...').[14]

Finalmente, cabe destacar que en muchos fragmentos el sistema de n-gramas deja una gran parte de las palabras en catalán, ya que no tiene suficiente evidencia en el corpus de entrenamiento.

4 Extensiones y mejoras del prototipo inicial

4.1 Optimización del algoritmo básico

Aún sin modificar la filosofía inicial de utilizar como fuente de conocimiento lingüístico principal un corpus paralelo sin etiquetado lingüístico, todavía quedan algunas líneas de mejora por explorar. Entre estas líneas podemos destacar:

- **Precalculado del modelo de traducción.** En el prototipo no se calcula ningún modelo previo, sino que para cada n-grama se calcula su posible traducción y su probabilidad, cada vez que se ejecuta. Para mejorar el rendimiento, los diferentes n-gramas están indexados respecto a los segmentos que aparecen. Tenemos previsto realizar un cálculo previo de las traducciones de todos los n-gramas que aparecen en el corpus de entrenamiento.

[13] Traducción de referencia: *se especifican en el anexo, que pertenecen ...*

[14] Traducción de referencia: *pero vecinos* **de** *los edificios más afectados ...*

Original	Referencia	n-gramas	Internostrum
plec (DOGC)	pliego	pliego	pliegue
errada (DOGC)	errata	errata	error
licitació (DOGC)	licitación	licitación	puja
referèndum (Periódico)	referendo	referendo	referéndum
trets (Periódico)	tiros	tiros	disparos

Tabla 3: Ejemplos de terminología en que el prototipo de sistema de n-gramas consigue mejores resultados que Internostrum

De esta manera se conseguirá mejorar la eficiencia del sistema tanto en velocidad, como en calidad de las traducciones. Al realizar el cálculo de las traducciones, se empezará por los n-gramas de mayor frecuencia, para los cuales existen más frases de ejemplo. Una vez realizadas estas traducciones se evita que las traducciones ya asignadas sean reasignadas a otro n-grama. De esta manera se puede prescindir del pequeño diccionario bilingüe.

- **Nuevo algoritmo de creación de la traducción resultante.** En el algoritmo presentado, la traducción resultante se va creando a medida que se van calculando las traducciones de los n-gramas desde el orden mayor de n hasta $n = 1$. Esto hace que no se exploren todas las posibles combinaciones de n-gramas que pueden formar la traducción. Si disponemos del modelo de traducción precalculado, podemos explorar todas las posibles combinaciones de una manera eficiente. De esta manera podremos escoger la traducción resultante que presente una mayor probabilidad global.

- **Utilización de fragmentos superpuestos.** El sistema propuesto utiliza n-gramas traducidos siempre y cuando los respectivos originales no se superpongan entre ellos. En caso de superposición, se utiliza el n-grama que presenta una probabilidad mayor. Como se apunta en [Brown et al.2003], la utilización de n-gramas traducidos que se superponen pueden suponer una mejora en el resultado de la traducción, ya que el hecho de que dos n-gramas adyacentes se superpongan hace que aumente la probabilidad que la combinación sea una traducción precisa.

- **Utilización de diccionarios bilingües.** El algoritmo presentado no puede ofrecer ninguna traducción de una palabra que no aparezca en el corpus de entrenamiento. Todos aquellos unigramas que queden por traducir se dejan como el original, lo que representa uno de los mayores problemas de nuestro sistema, tal como se menciona en el apartado de evaluación. En caso de disponer de un diccionario bilingüe de formas lo suficientemente grande, estos casos podrían consultarse, lo que supondría una mejora relativamente poco costosa.

- **Ampliación del corpus paralelo.** El aumento del tamaño y la variedad del corpus paralelo implicaría mejores resultados de traducción. Para ello se ha adaptado el algoritmo de alineación automática de documentos de Moore [Moore2002]. Este algoritmo permite obtener las alineaciones 1 : 1 de un conjunto de documentos y su correspondiente traducción.

4.2 Utilización de información lingüística básica

El sistema descrito en el apartado 2.3 se basa únicamente en propiedades estadísticas del lenguaje sin recurrir a información lingüística de ningún tipo, si exceptuamos el pequeño diccionario bilingüe.

Existen distintas posibilidades de enriquecer lingüísticamente el sistema utilizando recursos básicos de procesamiento del lenguaje.

Mediante la lematización y etiquetado del corpus de la lengua origen, en nuestro caso del catalán, así como de la frase de entrada, podemos realizar una búsqueda más amplia de n-gramas coincidentes. [15]

Así, distinguimos entre:

- **coincidencia perfecta**: los caracteres que componen la cadena de los n-gramas comparados son idénticos (no tiene en cuenta flexión morfológica, etc.)

[15] Para lematizar y etiquetar el corpus en catalán está previsto utilizar CatCG [Alsina et al.2002]

- **coincidencia optimizada**: coincide el lema (pero no la forma) o coincide la etiqueta morfosintáctica (pero no el lema) de alguno de los componentes de los n-gramas comparados.

Como ilustración de esto, volvamos a la frase de ejemplo del apartado 2.3. Supongamos que en lugar de encontrar en nuestro corpus el 4-grama *El president del Govern*, tuviéramos por un lado el 4-grama *El secretari del Govern*, y por otro el unigrama *president*. Puesto que *president* y *secretari* comparten categoría sintáctica, así como género y número, obtendríamos el 4-grama *El N+masc+sing del Govern*, y en la obtención de la cadena final sustituiríamos la etiqueta por la traducción del unigrama *president*.

Si, por otro lado, en lugar de aparecer la cadena exacta *presentarà* apareciera *presenta*, la coincidencia de lema nos permitiría escoger el unigrama adecuado. En la construcción de la secuencia final deberíamos proceder a la generación morfológica de la palabra.

La información de categoría sintáctica nos permitirá también dar mayor peso (o aumentar la probabilidad) de ciertas combinaciones de n-gramas por encima de otras. Por ejemplo, el bigrama Det+N es preferible al bigrama N+Det.

Por otro lado, el etiquetado, no sólo del corpus de la lengua origen sino también del de la lengua destino, permitiría optimizar la creación de los modelos de traducción, puesto que se consideraría más probable que una palabra se tradujera por otra de la misma categoría.

4.3 Uso de corpus monolingüe en lugar de corpus paralelo

Un inconveniente del sistema experimental descrito aquí (inconveniente común a todos los sistemas de traducción automática estadística actuales) reside en la necesidad de que existan corpus paralelos adecuados de los que poder extraer las traducciones. Por razones obvias, los corpus paralelos son mucho más difíciles de conseguir que los corpus monolingües. Partiendo de esta premisa, el proyecto europeo Metis [Dologlou et al.2003] se propone construir un sistema de traducción automática basado únicamente en un diccionario bilingüe y un corpus monolingüe de la lengua destino. La primera fase del proyecto (Metis I) demostró la viabilidad de la metodología para frases completas. En la actual fase del proyecto (Metis II) se pretende ampliar la cobertura del sistema utilizando unidades menores que la frase, ya sean fragmentos lingüísticos coherentes (*chunks*) o n-gramas.[16]

Como una continuación natural del sistema de traducción basado en n-gramas bilingües, nuestro grupo estudia la adaptación de la metodología aquí presentada al marco propuesto por Metis. Para ello, sustituimos el corpus paralelo por un diccionario bilingüe de amplia cobertura léxica, en el que cada palabra origen puede tener más de una traducción. Puesto que el diccionario traduce lemas y no formas flexionadas, hay que lematizar también la frase de entrada, así como el corpus monolingüe que se consulta para construir la frase traducida. Este corpus se utiliza como fuente de información para la desambiguación léxica y la generación del orden de palabras.

Tras lematizar el corpus monolingüe, extraemos los n-gramas correspondientes. Cada lema de la lengua origen se compara con la parte izquierda del diccionario bilingüe obteniéndose así una serie de traducciones candidatas. Se procede a traducir, empezando por el valor más alto de n (por ejemplo, 4), de izquierda a derecha. Para tratar el problema del orden de palabras, los n-gramas traducidos se consideran conjuntos (o *bolsas*) y no secuencias ordenadas de palabras.[17] Basándonos en la frecuencia de cada n-grama traducido en el corpus monolingüe, se asigna una probabilidad a cada candidato. En caso de que no se encuentren traducciones para un n-grama determinado, se recurre a las etiquetas morfosintácticas en vez de a los lemas.

Cuando se han hecho todos los cálculos para un determinado valor de n, se traducen todos los fragmentos de la frase (empezando por el n-grama con mayor probabilidad) y se recalculan los n-gramas de los fragmentos no traducidos, para empezar de nuevo el proceso con n=n-1. Esta modificación del prototipo inicial está en fase de implementación.

[16] Metis II tiene como objetivo la traducción de cuatro lenguas europeas (español, griego, holandés y alemán) al inglés, y utiliza el British National Corpus como corpus de la lengua destino.

[17] En el caso de lenguas tan próximas como el castellano y el catalán, este paso puede no ser necesario.

5 Conclusiones

En este artículo hemos presentado un sistema experimental de traducción automática estadística basado en n-gramas que, a partir únicamente de un corpus paralelo, es capaz de traducir oraciones nuevas con un nivel de aceptabilidad notablemente superior al de una memoria de traducción comercial que utiliza el mismo corpus paralelo. En la comparación con un sistema de traducción automática convencional, aunque sus resultados son esperablemente inferiores, el análisis de los errores desvela algunos puntos fuertes del sistema basado en n-gramas, notablemente la selección léxica contextual, tanto en el ámbito terminológico, como de elementos gramaticales (preposiciones, etc.).

Por tanto, el prototipo de traductor en el actual estadio de desarrollo da un resultado a medio camino entre las memorias de traducción y los traductores automáticos completos. Ello lo convertiría en una herramienta útil como complemento a una aplicación de traducción asistida. Además, y como mínimo para lenguas próximas como el catalán y el castellano, se puede convertir en un traductor automático eficiente sobre todo en entornos de creación de documentación repetitiva y del que se disponga de corpus paralelos suficientemente grandes.

Finalmente, hemos descrito dos prometedoras direcciones de desarrollo del sistema. La primera consistiría en ampliar su cobertura utilizando información lingüística en los corpus de entrenamiento. La segunda permitiría entrenar el sistema usando un corpus monolingüe y un diccionario bilingüe, en lugar de un corpus paralelo, que es un recurso mucho más dificil de obtener.

Referencias

[Alsina et al.2002] Alsina, À., T. Badia, G. Boleda, S. Bott, À. Gil, M. Quixal, y O. Valentín. 2002. CATCG: a general purpose parsing tool applied. En *Proceedings of the Third LREC.*

[Brown et al.1993] Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, y R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter stimation. *Computational Linguistics*, 19(2):263–311.

[Brown et al.2003] Brown, R., R. Hutchinson, P. Bennett, J.G. Carbonell, y Jansen P. 2003. Reducing boundary friction using translation-fragment overlap. En *Proceedings of the IX MT Summit*, páginas 311–318.

[Canals-Marote et al.2001] Canals-Marote, R., A. Esteve-Guillén, A. Garrido-Alenda, M.I. Guardiola-Savall, A. Iturraspe-Bellver, S. Montserrat-Buendía, S. Ortiz-Rojas, H. Pastor-Pina, P.M. Pérez-Antón, y M.L. Forcada. 2001. The spanish-catalan machine translation system internostrum (r). En *Proceedings of the VIII MT Summit.*

[Dologlou et al.2003] Dologlou, Y., S. Markantonatou, G. Tambouratzis, O. Yannoutsou, A. Fourla, y N. Ioannou. 2003. Using monolingual corpora for statistical machine translation: The metis system. En *Proceedings of the EAMT-CLAW 03: Controlled Language Translation.*

[Guillén y Iturraspe-Bellver2001] Guillén, E. y A. Iturraspe-Bellver. 2001. Avaluació de sistemes de traducció automàtica del parell castellà-català. En *Actas del XVI encuentro de la Asociación de Jóvenes Lingüistas.*

[Hiemstra1998] Hiemstra, D. 1998. Multilingual domain modeling in twenty-one: automatic creation of a bi-directional translation lexicon from a parallel corpus. En *Proceedings of the 8th CLIN meeting*, páginas 41–58.

[Moore2002] Moore, R. 2002. Fast and accurate sentence alignement of bilingual corpora. En *Machine Translation: from Research to Real Users: Proceedings of the 5th Conference of the AMTA*, páginas 135–244.

[Papineni et al.2002] Papineni, K., S. Roukos, T. Ward, y W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. En *Proceedings of the 40th ACL*, páginas 311–318.

[Tomás, Mas, y Casacuberta2003] Tomás, J., J.A. Mas, y F. Casacuberta. 2003. A quantitative method for machine translation evaluation. En *Proceedings of the of EACL 2003 workshop on Evaluation Initiatives in NLP.*

[Turcato y Popowich2001] Turcato, D. y F. Popowich. 2001. What is example-based machine translation? En *Proceedings of the Workshop on EBMT, hosted by MT-Summit VIII.*