Integración de reordenamientos en el algoritmo de decodificación en traducción automática estocástica

Josep M. Crego y José B. Mariño

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

Resumen: En esta comunicación se presenta un marco de trabajo para introducir la capacidad de reordenamiento de palabras en traducción automática (TA). Los reordenamientos producidos en la oración fuente se integran en el algoritmo de decodificación, lo que permite construir un grafo de búsqueda de dimensiones reducidas. A partir de un grafo de búsqueda monótono (sin reordenamientos), se utilizan patrones de reordenamiento (patrones de reescritura motivados lingüísticamente) para añadir arcos que introducen permutaciones de las palabras fuente. Los patrones se aprenden de manera automática a partir del conjunto de entrenamiento, utilizando los alineamientos de palabras (entre las oraciones fuente y destino) y las etiquetas morfo-sintácticas (POS) de las oraciones fuente. Una vez completado el grafo de búsqueda, el algoritmo de decodificación lo atraviesa asignando una probabilidad (coste) a cada hipótesis, ayudándose por un modelo de lenguaje N-grama aprendido de las etiquetas POS del idioma origen después de ser reordenadas (además de por un conjunto de modelos típico en traducción automática). El método propuesto se evalúa en una tarea de traducción del español al inglés y viceversa, utilizando el corpus del Parlamento Europeo, donde pueden observarse mejoras tanto en calidad de la traducción (con medidas subjetivas y automáticas) como en eficiencia computacional.
Palabras clave: traducción automática estocástica, etiquetado POS, algoritmos de decodificación, reordenamiento

Abstract: This paper presents a reordering framework for statistical machine translation (SMT) where source-side reorderings are integrated into SMT decoding, allowing for a highly constrained reordered search graph. The monotone search is extended by means of a set of reordering patterns (linguistically motivated rewrite patterns). Patterns are automatically learnt in training from word-to-word alignments and source-side Part-Of-Speech (POS) tags. Traversing the extended search graph, the decoder evaluates every hypothesis making use of a group of widely used SMT models and helped by an additional Ngram language model of source-side POS tags. Experiments are reported on the Euparl task (Spanish-to-English and English-to-Spanish). Results are presented regarding translation accuracy (using human and automatic evaluations) and computational efficiency, showing significant improvements in translation quality for both translation directions at a very low computational cost.
Keywords: stochastic machine translation, POS tagging, decoding algorithms, reordering

1. *Introducción*

En traducción automática estocástica (TAE), el uso de estrategias de reordenamiento permite importantes mejoras en la calidad de la traducción, especialmente cuando se traduce entre pares de lenguas con disparidad en el orden de sus palabras. Por otra parte, cuando se permiten todas las permutaciones de palabras de la oración origen, el problema algorítmico se convierte en NP-completo (Knight, 1999), mientras que existen algoritmos de búsqueda de complejidad polinomial bajo condiciones monótonas (sin reordenamientos).

Los primeros sistemas de TAE en introducir habilidades de reordenamiento se basaban en aprovechar la fuerza bruta de los ordenadores. El algoritmo de búsqueda atravesaba un grafo donde todas las permutacines de palabras de la oración fuente estaban permitidas. Este enfoque resulta computacionalmente muy ineficiente, incluso cuando se aplica a oraciones fuente de tamaño reducido. Para que el proceso de búsqueda sea factible, debe reducirse el conjunto de permutaciones. Inicialmente se aplicaron restricciones basadas en la distancia de reordenamiento: **ITG** (Wu, 1996), **IBM** (Berger, Della Pietra, y Della Pietra, 1996), **Local** (Kanthak et al., 2005), etc. La utilización de tales restricciones implica la necesidad de acuerdo entre eficiencia computacional y calidad final de la traducción.

En combinación con las restricciones anteriores, se utiliza un modelo basado en la distancia de reordenamiento, que penaliza aquellos desplazamientos en que la distancia de reordenamiento es mayor, de manera que sólo se permitiran los que consigan una buena puntuación (coste) bajo el conjunto restante de modelos. Evidentemente, tal modelo no pretende explicar ninguna propiedad del lenguaje. Recientemente se han introducido modelos de reordenamiento basados en los desplazamientos de palabras aprendidos en el conjunto de entrenamiento, (Koehn et al., 2005), (Kumar y Byrne, 2005).

La principal crítica hacia el enfoque por fuerza bruta consiste en que no hace ningun uso de información linguística, cuando en la teoría, el orden relativo de las palabras entre diferentes pares de lenguas ha sido extensamente descrito.

Recientemente, con el objetivo de abordar el problema del reordenamiento, algunos sistemas de TAE introducen información linguística de diversa índole:

- Uso de unidades de traducción jerárquicas que conllevan reordenamientos de manera implícita (Chiang, 2005), o a través de unidades de traducción que contienen posiciones a cubrir (por nuevas unidades) (Simard et al., 2005).

- Monotonización del orden de las palabras entre los dos idiomas implicados en la traducción (tanto en el conjunto de entrenamiento como en el de test). Reordenamientos de la oración fuente para conseguir el mismo orden de las palabras de la oración destino (Collins, Koehn, y Kucerova, 2005), (Xia y McCord, 2004).

En el trabajo de (Xia y McCord, 2004), se aplican reordenamientos a partir de un conjunto de patrones aprendidos automaticamente utilizando información léxica, sintáctica y morfológica (palabras, árboles sintácticos y etiquetas POS). Previo a la decodificación, se aplican los patrones de reordenamiento que alteran el orden de las palabras de la oración fuente. La decodificación se efectua bajo condiciones monótonas.

En el presente trabajo, se utiliza una estrategia parecida para aprender los patrones de reordenamiento. El objetivo final es doble: inicialmente pretendemos utilizar información linguística para encontrar el conjunto de reordenamientos que facilitan la traducción (consiguiendo cierto poder de generalización a partir de utilizar etiquetas POS), además de decidir la mejor permutación de palabras de la oración fuente con el máximo de información disponible, es decir, en el proceso de búsqueda cuando disponemos de todos los modelos.

En (Matusov, Kanthak, y Ney, 2005) se presenta un trabajo similar donde las permutaciones admitidas (en el grafo de búsqueda) consisten en aquellas que no dividen alguna secuencia de palabras vista (consecutiva) en el conjunto de entrenamiento. Es decir, sólo se permitirá reordenar una palabra (o secuencia de palabras) que no haya sido vista siguiendo a su predecesora en alguna de las oraciones del conjunto de entrenamiento.

En lo que sigue, se describe teóricamente el sistema de traducción (sección 2), se presenta el marco de trabajo propuesto que facilita el reordenamiento de palabras en TAE (sección 3), dándose detalles del método utilizado para extraer patrones y de cómo estos se constituyen en arcos del grafo de búsqueda. Se recogen los detalles experimentales del entrenamiento del sistema y del proceso de traducción (sección 4) así como se ofrecen y discuten los resultados obtenidos. Finalmente se plantean las conclusiones (sección 5).

2. Systema de TAE basado en N-gramas

Nuestro sistema de TAE se sitúa bajo el enfoque de máxima entropia (Berger, Della Pietra, y Della Pietra, 1996), donde podemos definir la hipótesis de traducción d dada una oración fuente f, como la oración destino que maximiza una combinación lineal de varias funciones características (expresadas mediante funciones logarítmicas):

$$\hat{t}_1^I = \arg\max_{t_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(s_1^J, t_1^I) \right\} \qquad (1)$$

donde λ_m corresponde a los coeficientes que ponderan cada una de las funciones características $h_m(s,t)$.

El sistema descrito en este artículo implementa una combinación lineal de un modelo de traducción con **cinco** modelos adicionales. En nuestro sistema, el modelo de traducción está expresado en función del concepto de *tupla*.

Dado un par de oraciones alineadas a nivel de palabra, las tuplas definen una segmentación única y monotónica del par de oraciones, permitiendo la estimación de un lenguaje basado en N-gramas (lenguaje bilingüe), que tiene en cuenta la historia en el proceso de traducción.

El modelo de tuplas (modelo de traducción), consiste en un modelo de lenguaje formado a partir de unidades bilingües (tuplas). El modelo queda descrito por la siguiente equación:

$$\hat{d}_1^I = \arg\max_{d_1^I} \{p(f_1^J, d_1^I)\} = \cdots = \qquad (2)$$

$$\arg\max_{d_1^I} \{\prod_{i=1}^{K} p((f,d)_i | (f,d)_{i-N+1}, ..., (f,d)_{i-1})\} \qquad (3)$$

donde $(f,d)_i$ hace referencia a la i^{esima} tupla de un par de oraciones bilingüe, que ha sido segmentado en K unidades.

Las demás funciones características:

- un modelo de lenguaje destino (N-grama),
- un modelo que penaliza las traducciones más cortas,
- un modelo de traducción fuente-a-destino,
- un modelo de traducción destino-a-fuente,
- un modelo de lenguaje del destino usando etiquetas morfo-sintácticas (POS).

En (Mariño et al., 2005) puede encontrarse una descripción de las cuatro primeras funciones características.

La quinta función característica consiste en un modelo de lenguaje N-grama, estimado sobre las etiquetas POS de las oraciones destino del conjunto de entrenamiento. La equación que expresa la función característica es la siguiente:

$$p_{LM}(d_k) \approx \prod_{n=1}^{k} p(pos_n | pos_{n-2}, pos_{n-1}) \qquad (4)$$

3. *Integración de reordenamientos en la búsqueda global*

En esta sección se detallan los métodos que se han utilizado para extraer los patrones de reordenamiento y para ampliar el grafo de búsqueda (inicialmente monótono) con los arcos que representan las permutaciones de palabras de las oraciones fuente.

3.1. Patrones a partir de etiquetas morfo-sintácticas (POS)

En la extracción de patrones se han utilizado los alineamientos (la unión de los alineamientos en ambos sentidos) y las etiquetas POS de las oraciones fuente. En general, el procedimiento consiste en identificar todos los cruces producidos en los alineamientos.

La ecuación siguiente formaliza el conjunto de cruces dado un par de oraciones alineadas a nivel de palabras:

$$\{(j_1, j_2) \,/\, (j_1 < j_2) \wedge (a[j_1] > a[j_2])\} \qquad (5)$$

donde $a[j]$ hace referencia a la máxima posición (en la parte destino) a la que la palabra fuente j esta alineada, y j_1, j_2 varían entre $[1, J]$.

Una vez que un cruce ha sido detectado, se utilizan sus alineamientos y etiquetas POS (de la parte fuente de la secuencia de palabras que generan el cruce) para dar cuenta de una nueva instancia de patrón. La parte derecha del patrón (reordenamiento), se calcula utilizando el órden original de las palabras destino a las que las palabras fuente del patrón están alineadas. En la figura 1 puede verse un ejemplo clarificador.

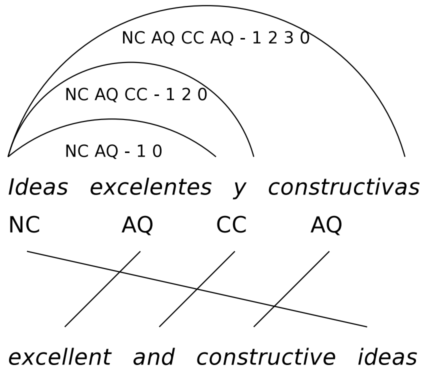

Figura 1: *Extracción de patrones usando alineamientos a nivel de palabras. Se consigue capacidad de generalización a través de las etiquetas POS. En el ejemplo, se han extraido tres instancias de patrones. Las etiquetas **NC**,**AQ** y **CC** corresponden respectivamente a: nombre común, adjetivo calificativo y conjunción coordinada.*

3.2. Amplicación del grafo monótono

Utilizando los patrones de reordenamiento anteriores, se añaden nuevos arcos (uno por patrón) al grafo de búsqueda (inicialmente) monótono. El procedimiento consiste en identificar primero las secuencias de palabras de la oración fuente (sus etiquetas POS) que encajan con alguno de los patrones. Una vez identificados, se añade un nuevo arco al grafo de búsqueda indicando el reordenamiento a evaluar. Una única excepción a este procedimiento consiste en no añadir el arco si existe una unidad de traducción (tupla) con las mismas palabras que generan el cruce (dado que probablemente el reordenamiento ya esté interiorizado en la unidad de traducción). En la figura 2 puede verse un ejemplo.

Una vez que el grafo de búsqueda se ha construido (ampliado), el algoritmo de decodificación lo atraviesa buscando la mejor traducción. Así pues, la hipótesis ganadora resultará de utilizar toda la información disponible (todos los modelos).

En la sección 4.2 se dan algunos detalles adicionales.

4. Experimentos

4.1. Corpus

El conjunto de datos EPPS corresponde a las transcripciones de sesiones del Parlamento Europeo, accesibles a tra-

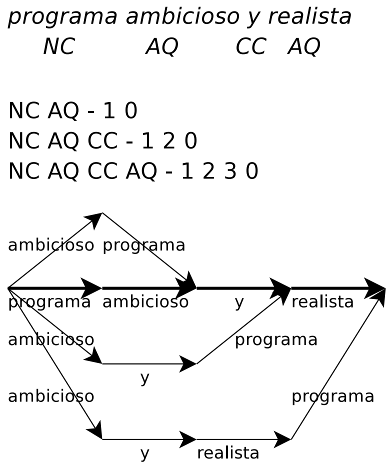

Figura 2: *Utilizando los patrones que se han identificado para la oración fuente, se han añadido tres arcos al grafo de búsqueda (en negrita).*

ves de la página web del Parlamento (`http://www.euro parl.eu.int/`). Para los experimentos aquí presentados, se ha utilizado la versión disponible por la Universidad de Aachen RWTH, a través del consorcio TC-STAR[1].

En el cuadro 1 se presentan algunas estadísticas de los conjuntos de entrenamiento, desarrollo y test en que se dividió el corpus, para cada idioma: inglés y español. En concreto consisten en el número total de oraciones, número de palabras, vocabulario de palabras y el vocabulario de etiquetas POS.

	orcn	plbr	vcb	POS vcb
Entr.				
Inglés	1.28 M	34.9 M	106 k	44
Español		36.6 M	153 k	328
Des.				
Inglés	735	18,764	3,193	41
Español	430	15,332	3,217	181
Test				
Inglés	1,094	26,917	3,958	42
Español	840	22,774	4,081	196

Cuadro 1: *Estadísticas del corpus paralelo (inglés-español) del TC-Star.*

[1]TC-STAR (Technology and Corpora for Speech to Speech Translation), que es un proyecto Europeo enmarcado en el *Sixth Framework Programme*. Para mayor información, consultar la página web del consorcio `http: //www.tc-star.org/`

4.2. Detalles del sistema

A partir del conjunto de entrenamiento alineado a nivel de oraciones, se ha alineado a nivel de palabras usando GIZA++ (en concreto se utiliza la unión de los alineamientos en ambas direcciones) (Och, 2003).

Posteriormente, a partir de los alineamientos anteriores (unión), se han extraido las unidades de traducción (tuplas), el vocabulario de las cuales se ha podado, quedándose con las $N = 30$ mejores traducciones para cada parte fuente (de cada tupla) en el caso inglés a español, y $N = 20$ en el caso español a inglés.

La parte inglesa del corpus ha sido etiquetada (etiquetas POS) usando la herramienta TNT (Brants, 2000), para la parte española se ha usado Freeling (Carreras et al., 2004). En el caso de las etiquetas del español, sólo los dos primeros carácteres han sido tenidos en cuenta (correspondientes a los carácteres que indican la categoría morfosintáctica de la palabra), con la intención de aumentar el poder de generalización y para obtener un nivel equivalente al inglés en riqueza de etiquetas.

Para aprender los modelos de lenguaje N-grama se ha utilizado el SRI Language modelling toolkit (Stolcke, 2002), usando respectivamente 4, 5 y 5 como órdenes de los modelos de traducción, destino y destino etiquetado.

En la optimización del sistema (búsqueda de los coeficientes a utilizar para cada modelo en la combinación lineal de la ecuación 1) se ha utilizado una implementación propia del método downhill simplex (Nelder y Mead, 1965).

El algoritmo de decodificación trabaja siempre realizando una poda (por histograma), quedándose con las $b = 25$ mejores hipótesis.

4.3. Patrones usados

Usando el método descrito en la sección 3, se extraen una gran cantidad de patrones, muchos de los cuales sólo aparecen a partir de errores en los alineamientos a nivel de palabras. Además, trabajar con una gran cantidad de patrones es poco deseable al disminuir la velocidad del algoritmo de búsqueda. Para evitar esa situación se ha realizado un filtro de patrones, utilizando las siguientes restricciones:

- Limitar a 4 la diferencia entre palabras de cada instancia de patrón.
- Limitar a 8 el número máximo de palabras fuente de cada patrón.
- Eliminar todos aquellos patrones que no superen un número mínimo de instancias ($N > 1000$).
- Se ha utilizado una probabilidad p para medir la bondad de cada patrón, consistente en el número de apariciones del patrón (en el conjunto de entrenamiento), dividido por el número de apariciones de la parte fuente del patrón. Sólo se han considerado aquellos patrones con un mínimo de probabilidad ($p > 0{,}2$).

La lista resultante de patrones (después del filtrado) puede consultarse en el cuadro 3 (29 patrones han sido considerados en la dirección inglés a español). En el cuadro también puede verse el número de apariciones de cada patrón en los conjuntos de entrenamiento, desarrollo y test, así como un ejemplo de cada patrón. A pesar del filtrado de patrones, aún existen instancias incorrectas. Algunas debidas a malos alineamientos, otras debidas al método de extracción utilizado. Por ejemplo, el patrón **(NC RG - 1 0)** (la etiqueta **RG** corresponde a la categoría de adverbio general), no debería tenerse en cuenta al ser un cruece interno. Es decir, un patrón correcto puede contener varios cruces, siendo sólo el exterior el que debería ser tenido en cuenta, en este caso el patrón probablemente sería **(NC RG AQ - 1 2 0)**.

En cualquier caso, en este trabajo no se pretende encontrar un método para realizar decisiones de reordenamiento perfectas previas al proceso de búsqueda, sino reducir el número total de permutaciones que se consideran en un grafo enteramente reordenado. Evidentemente cuanto mejor sea la lista de permutaciones presentadas al algoritmo de búsqueda, mayor será su eficiencia.

La lista actual de patrones es útil incluso con algunos (a priori) erróneos, dado que nos permite evaluar la habilidad del algoritmo de búsqueda (en conjunto con la totalidad de modelos del sistema) para descartar los malos. Debería ser capaz de distinguir entre los unos y los otros realizando sólo, las permutaciones bien consideradas por el conjunto de modelos.

Usando la lista de patrones de el cuadro 3, y las etiquetas POS de la parte fuente del conjunto de entrenamiento, se procedió a reordenar las unidades

El conjunto de entrenamiento fuente, tras ser etiquetado (POS), ha sido reordenado utilizando los patrones anteriores (sólo las etiquetas, no las palabras) como fuente para aprender un modelo de lenguage N-grama. (de orden 5). Este modelo pretende ser utilizado como modelo de reordenamiento, indicando al algoritmo de búsqueda qué permutaciones debe realizar y cuáles descartar.

4.4. Resultados

Se han considerado tres configuraciones diferentes:

- **base**: Sin reordenamientos, usando una búsqueda monótona.
- **rgraph**: Permitiendo reordenamientos usando los patrones descritos en este artículo (por lo demás se utiliza la misma configuración anterior).
- **pos**: Añadiendo a la configuración anterior el modelo de lenguaje de etiquetas POS del fuente reordenadas.

Las medidas automáticas de evaluación utilizadas (BLEU, NIST, mWER and PER) son las mismas que se han utilizado en la evaluación del proyecto TC-STAR, distribuidas por ELDA(http://www.elda.org/). En todas las evaluaciones se han utilizado dos referencias.

En el cuadro 2 pueden verse los resultados obtenidos por cada configuración, la segunda columna indica la medida (BLEU) obtenida sobre el conjunto de desarrollo tras la optimización. El resto de columnas indican los resultados sobre el conjunto de test.

Conf	bleu'	bleu	nist	mwer	per
español a inglés					
base	52,9	55,2	10,69	34,40	25,32
rgraph	53,3	55,6	10,70	34,23	25,50
pos	53,9	56,4	10,75	33,75	25,41
inglés a español					
base	48,1	48,0	9,84	41,18	31,11
rgraph	49,0	48,5	9,81	41,15	31,87
pos	49,1	48,9	9,91	40,29	31,27

Cuadro 2: *Resultados obtenidos por el sistema en su configuración base, usando los reordenamientos, y ayudado por el modelo adicional. Los intérvalos de confianza para el BLEU son de* ±1,12 *y* ±1,62 *(español a inglés y inglés a español respectivamente) para un nivel de confianza del* 95 %.

Tal y como puede verse, en la tarea de español a inglés, los resultados en los conjuntos de desarrollo están muy correlacionados si comparamos las diferentes configuraciones en todas las medidas objetivas excepto en PER. La explicación se encuentra en que PER no tiene en cuenta los errores derivados de reordenamientos, que són las diferencias básicas entre las diferentes configuraciones. El resto de medidas muestran una mejora en la calidad de la traducción en la configuración *rgraph* así como en la configuración *pos*.

En el caso de la tarea de inglés a español, comparando las configuraciones *base* y *rgraph*, sólo BLEU muestra una mejora clara, el resto de medidas apenas muestran diferencias. Para la configuracione *pos*, la mejora es clara bajo todas las medidas.

En el cuadro 3 también puede verse una evaluación subjetiva para el conjunto de test (columnas quinta y sexta). En la quinta columna se indica el número de reordenamientos realizados por el algoritmo, en la columna sexta se indica el número de errores subjetivos detectados para las secuencias de palabras en que el algoritmo debía tomar una decisión (reordenar o no).

Respecto a la evaluación subjetiva, nos hemos concentrado en los arcos adicionales (secuencias de palabras de la oración fuente para las que se ofrecía la posibilidad al algoritmo de búsqueda de reordenar), y hemos considerado erróneas tanto los errores por reordenamiento como por no reordenamiento. En cualquier caso, en la evaluación subjetiva no se consideraban las malas traducciones, sino el orden de las palabras en el idioma destino. Por ejemplo, dada la oración fuente **'programa ambicioso y realista'**, si el decoder decide reordenar usando el patrón **(NC AQ CC AQ - 1 2 3 0)**, llegando a la traducción **'ambitious and unrealistic programme'**, se ha considerado como acierto, a pesar de ser semánticamente una mala traducción (el orden de las palabras destino es correcto).

En la tarea de español a inglés, el decodificador decidió ordenar un total de 214 secuencias del total de 379 opciones (arcos adicionales) ($\sim 56\,\%$). Del total de opciones 42 fueron erróneas (tanto por reordenar cuando no debía como por no reordenar cuando debía) ($\sim 11\,\%$). Se observaron resultados muy parecidos en la tarea de inglés a español.

A partir de la evaluación humana, podemos dividir los patrones en 2 grandes grupos. Primero (en cursiva), aquellos para los que se han tomado muy pocas veces la decisión

Patrón	entr.	des.	test	giro	error	Ejemplo
NC RG AQ CC AQ - 1 2 3 4 0	1,406	1	1	1	0	ideas muy sencillas y elementales
NC AQ CC AQ - 1 2 3 0	27,119	13	**23**	**17**	**2**	programa ambicioso y realista
NC AQ RG AQ - 2 3 1 0	1,971	0	4	1	0	control fronterizo más estricto
NC CC NC AQ - 3 0 1 2	3,355	6	**12**	**6**	**3**	mezquitas y centros islámicos
NC RG AQ CC - 1 2 3 0	2,226	3	2	0	0	ideas muy sencillas y
AQ RG AQ - 1 2 0	2,777	21	*7*	*2*	*1*	europea más sólida
NC AQ AQ - 2 1 0	35,661	11	**24**	**18**	**3**	decisiones políticas delicadas
NC RG AQ - 1 2 0	32,887	0	**35**	**26**	**1**	ideas muy sencillas
NC RG RG - 1 2 0	1,473	0	3	3	2	texto mucho más
NC AQ - 1 0	877,580	113	**142**	**110**	**16**	preguntas serias
NC RG - 1 0	54,968	27	*47*	*7*	*7*	actividades aparentemente
AQ AQ - 1 0	46,509	14	*40*	*4*	*2*	medioambientales europeas
RN VM - 1 0	45,777	4	2	1	1	no promuevan
RG VA - 1 0	9,824	0	2	1	0	ahora habíamos
AQ RG - 1 0	8,701	11	*21*	*4*	*2*	suficiente todavía
RG VS - 1 0	5,043	1	1	1	0	supuestamente somos
VM PP - 1 0	4,769	6	**13**	**12**	**2**	estar ustedes
Total (17)	1,162,046	231	379	214	42	

Cuadro 3: *Lista de patrones extraidos del conjunto de entrenamiento para la dirección de español a inglés.*

de reordenar (tenidos en cuenta muy poco por el decodificador). Segundo (en negrita), aquellos para los que la decisión de reordenar (giro) fué tomada aproximadamente el mismo número de veces que la decisión de no reordenar.

El primer grupo nos sirve para identificar aquellos patrones que podrían ser filtrados de la lista de patrones, ya que prácticamente nunca se utilizan y la decisión de no utilizarlos es acertada (evaluación subjetiva). El segundo grupo nos indica que el grupo de modelos utilizado por el algoritmo de búsqueda es hábil en la elección de cuándo utilizar un patrón (en algunos casos una regla demasiado general).

Finalmente, la figura 3 indica el número de hipótesis expandidas por el algoritmo de búsqueda bajo tres condiciones diferentes: búsqueda monótona, usando los patrones para ampliar el grafo monótono y usando todas las permutaciones posibles de palabras de las oraciones fuente restringidas a una distancia máxima de $m = 3$ palábras y un número máximo de $j = 3$ reordenamientos. Se ha utilizado el conjunto de test de español a inglés en todos los casos.

Resulta remarcable cómo el espacio de búsqueda con reordenamientos (usando patrones) es prácticamente igual que el espacio de búsqueda monótono.

5. Conclusiones y trabajo futuro

En este trabajo se ha presentado un marco de trabajo en el que se integran reordenamientos extraídos con motivación lingüística en el proceso de búsqueda en traducción automática. Los patrones se extraen automáticamente de los alineamientos a nivel de palabras utilizando información morfo-sintáctica (POS).

El marco de trabajo ha sido probado para un sistema basado en N-gramas (aplicable también a un sistema basado en secuéncias de palabras *'phrases'*), consiguiendo mejoras en la calidad de la traducción a un precio muy bajo en eficiencia computacional. También se ha realizado una evaluación subjetiva que ha corroborado los resultados mostrando la robustez del sistema ante patrones de reordenamiento poco exactos.

Se prevé trabajar en la identificación de patrones, a través de mejorar el filtrado de

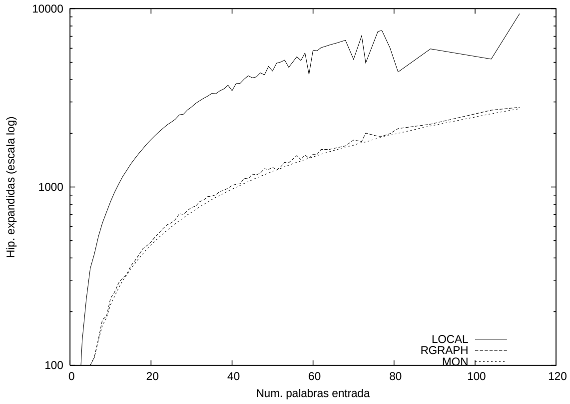

Figura 3: *Número de hipótesis en el grafo de búsqueda bajo diferentes condiciones de reordenamiento. Monótono (MON), usando los patrones de reordenamiento (RGRAPH) y usando todas las permutaciones bajo restricciones basadas en la distancia de reordenamiento (LOCAL).*

estos utilizando información adicional (palabras, árboles sintácticos, etc.). También pretendemos exportar esta idea a nuevos pares de lenguas, donde las necesidades de reordenamiento sean mayores (tales como chino-inglés o árabe-inglés).

6. Agradecimientos

Esta comunicación ha sido parcialmente subvencionada por el gobierno español, TIC-2002-04447-C02 (proyecto Aliado), la Unión Europea, FP6-506738 (proyecto TC-STAR) y la Universidad Politècnica de Catalunya (beca UPC-RECERCA).

Bibliografía

Berger, A., S. Della Pietra, y V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72, March.

Brants, T. 2000. TnT – a statistical part-of-speech tagger. En *Proc. of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.

Carreras, X., I. Chao, L. Padró, y M. Padró. 2004. Freeling: An open-source suite of language analyzers. *4th Int. Conf. on Language Resources and Evaluation, LREC'04*, May.

Chiang, D. 2005. A hierarchical phrase-based model for statistical machine translation. *43rd Annual Meeting of the Association for Computational Linguistics*, páginas 263–270, June.

Collins, Michael, Philipp Koehn, y Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. En *43rd Annual Meeting of the Association for Computational Linguistics*, páginas 531–540, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Kanthak, S., D. Vilar, E. Matusov, R. Zens, y H.Ñey. 2005. Novel reordering approaches in phrase-based statistical machine translation. *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, páginas 167–174, June.

Knight, K. 1999. Decoding complexity in word replacement translation models. *Computational Linguistics*, 26(2):607–615.

Koehn, P., A. Axelrod, A. Birch, C. Callison-Burch, M. Osborne, y D. Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'05*, October.

Kumar, S. y W. Byrne. 2005. Local phrase reordering models for statistical machine translation. *Proc. of the Human Language Technology Conference, HLT-EMNLP'2005*, October 6-8.

Mariño, J.B., R Banchs, J.M. Crego, A. de Gispert, P. Lambert, M. R. Costa-jussà, y J.A.R. Fonollosa. 2005. Bilingual n–gram statistical machine translation. *Proc. of the MT Summit X*, September.

Matusov, E., S. Kanthak, y H.Ñey. 2005. Efficient statistical machine translation with constrained reordering. *Proceedings of EAMT 2005 (10th Annual Conference of the European Association for Machine Translation)*, páginas 181–188, May.

Nelder, J.A. y R. Mead. 1965. A simplex method for function minimization. *The Computer Journal*, 7:308–313.

Och, F.J. 2003. Giza++ software. http://www-i6.informatik.rwth-aachen.de/˜och/ software/giza++.html. Informe técnico, RWTH Aachen University.

Simard, M., N. Cancedda, B. Cavestro, M. Dymetman, E. Gaussier, C. Goutte, K. Yamada, P. Langlais, y A. Mauser. 2005. Translating with non-contiguous phrases. *Proc. of the Human Language Technology Conference, HLT-EMNLP'2005*, página 8, October 6-8.

Stolcke, A. 2002. Srilm - an extensible language modeling toolkit. *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, September.

Wu, D. 1996. A polynomial-time algorithm for statistical machine translation. *34th Annual Meeting of the Association for Computational Linguistics*, páginas 152–158, June.

Xia, F. y M. McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. *Proc. of the 20th Int. Conf. on Computational Linguistics, COLING'04*, August 22-29.