Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 69-76 recibido 21-04-2005; aceptado 01-06-2005

Modelo estocástico de traducción basado en N-gramas de tuplas bilingües y combinación log-lineal de características

José B. Mariño **Rafael Banchs**
Josep Mª Crego **Adrià de Gispert**
Patrik Lambert **José A. R. Fonollosa**
Marta R. Costa-jussà
Centro de Investigación TALP
Universidad Politécnica de Cataluña
Campus Nord UPC. 08034-Barcelona.
{canton, rbanchs**,** jmcrego, agispert, lambert, adrian, mruiz}@gps.tsc.upc.edu

Resumen: En esta comunicación se presenta un sistema de traducción estocástica basado en el modelado mediante N-gramas de la probabilidad conjunta de textos bilingües. La unidad básica del modelo es la tupla, par de cadenas de palabras del lenguaje fuente (a traducir) y el lenguaje destino (traducción). La traducción se lleva a cabo mediante la maximización de una combinación lineal de los logaritmos de la probabilidad asignada a la traducción por el modelo de traducción y otras características, siguiendo la aproximación de entropía máxima. Las prestaciones del sistema de traducción son evaluadas con una tarea de traducción del habla: la traducción entre inglés y español (y viceversa) de transcripciones de intervenciones de los miembros del Parlamento Europeo. Los resultados alcanzados se encuentran al nivel del estado del arte.
Palabras clave: traducción automática del habla, traducción estocástica, N-gramas, modelo de lenguaje de entropía máxima.

Abstract: This communication introduces a stochastic machine translation system based on N-gram modelling of the joint probability of bilingual texts. The basic unit of this model is called a tuple and consists of a pair of both source (to be translated) language and target language (translation) word-strings. Translation is driven by a log-linear combination of the N-gram model probability and other features, according to the maximum entropy language modelling approach. The translation performance is evaluated by means of a speech-to-speech translation tasks: translation from Spanish to English (and viceversa) of European Parliament speeches. The system reaches a state-of-art performance.
Keywords: stochastic machine translation, speech-to-speech translation, N-gram model, maximum entropy language modelling.

1 Introducción

Recientemente los sistemas estocásticos de traducción han adquirido un notable protagonismo, gracias a los buenos resultados que han obtenido cuando se aplican a tareas de carácter limitado. Cuando se trata de traducir el habla, emerge otra razón importante para su popularidad: su capacidad para afrontar la traducción de oraciones no bien formadas desde el punto de vista gramatical. Esta agramaticalidad puede originarse en el carácter espontáneo del habla o en los errores de los sistemas de reconocimiento que actúan de intermediarios entre la señal de voz y el sistema de traducción. Por todo ello, hoy asistimos a un notable esfuerzo encaminado al desarrollo de sistemas estocásticos de traducción del habla capaces de abordar campos de aplicación no limitados, tanto en la talla del vocabulario como en su contenido semántico.

La aproximación estocástica considera que cualquier oración *f* de una lengua fuente (frase a traducir) puede ser traducida en cualquier otra *d* del lenguaje destino (en el que se desea la traducción) con probabilidad no nula. La traducción consiste precisamente en determinar la oración *d* con mayor probabilidad de constituir una traducción para la oración

ISSN: 1135-5948

original *f*. Las diferencias entre los distintos sistemas de traducción se originan en el modo que modelan la probabilidad de que *d* sea una traducción de *f*. Un rasgo común en esta aproximación es la necesidad de corpus bilingües paralelos (formados por pares de oraciones que se traducen mutuamente) a partir de los cuales estimar los parámetros del modelo.

El primer planteamiento (Brown et al., 1990) utilizó la palabra como la unidad básica del modelo de traducción. La probabilidad de traducción se establece en función de la probabilidad de traducción de las palabras, de un modelo estocástico de distorsión del orden de las palabras entre las dos lenguas y de la fertilidad de las palabras (la probabilidad de que una palabra de una lengua se traduzca en una, dos, tres, etc. palabras de la otra). Dentro de este planteamiento se establecieron diferentes modelos de complejidad creciente (llamados comúnmente modelos de IBM1, IBM2, etc.). Estos modelos son asimétricos, ya que para un par de lenguas dadas dependen del sentido de la traducción. El principal inconveniente de esta aproximación es la independencia del contexto de la probabilidad de traducción de las palabras y la dificultad algorítmica para estimar los modelos y realizar la traducción. Hoy en día, el principal fruto de este planteamiento y sus sucesivos refinamientos es su capacidad para establecer un alineamiento entre las palabras de un par de oraciones que son traducciones mutuas en el par de lenguas de interés. Es decir, como resultado del entrenamiento del modelo de traducción, se obtiene para cada par de frases del corpus de entrenamiento las palabras que se relacionan en la traducción o, dicho de otro modo, las palabras vinculadas (o enlazadas) entre sí de una y otra lengua (véase un ejemplo en la figura 1). GIZA++ (Och, 2003) es la herramienta distribuida gratuitamente de uso habitual a este fin.

Los sistemas estocásticos actuales de traducción utilizan como unidad básica del modelo secuencias de palabras (segmentos de oración) del par de lenguas que se encuentran vinculadas en la traducción. Este planteamiento permite adjudicar contexto a la traducción de las palabras. Estos segmentos son determinados tras un proceso de alineado de pares bilingües de oraciones pertenecientes a un corpus de entrenamiento. El modo en que se definen estos segmentos y se utilizan para modelar la probabilidad de traducción da origen a los diferentes sistemas.

Recientemente ha sido propuesto (Och y Ney, 2002) el uso del modelo de entropía máxima en la traducción estocástica. En esta propuesta, el modelo de traducción es una información más entre varias que pueden gobernar la traducción. Siguiendo la solución dual al problema de modelado (Berger et al, 1996), los logaritmos de las probabilidades asociadas a las diversas informaciones (características) son combinados linealmente para definir una función cuya maximización establece la traducción (modelo log-lineal). Esta estrategia es análoga a la combinación de los modelos fonético y de lenguaje empleada comúnmente en los sistemas de reconocimiento de voz. Los coeficientes de la combinación lineal son optimizados de acuerdo con algún criterio objetivo de la calidad de la traducción.

El sistema de traducción que se presenta en esta comunicación sigue el planteamiento de entropía máxima y es deudor del modelo de traducción basado en segmentos bilingües de palabras. En lo que sigue, se describe teóricamente el sistema de traducción (sección 2), se describe la tarea de traducción abordada en la sección 3, se recogen los detalles experimentales del entrenamiento del sistema y del proceso de traducción en la sección 4 y se ofrecen y discuten los resultados obtenidos en las secciones 5 y 6, respectivamente.

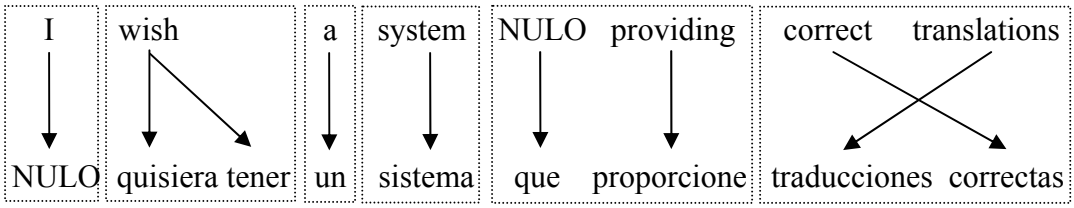

Figura 1: Par de oraciones bilingües en el que, mediante flechas, se indican las palabras vinculadas en la traducción. Mediante recuadros se muestran los pares bilingües de segmentos (tuplas) en los que se segmenta monótonamente el par de oraciones.

2 El sistema de traducción

2.1 Planteamiento

El problema de la traducción de una oración *f* del lenguaje original (o fuente) se convierte en la determinación de la oración *d* del lenguaje destino que maximiza la función

$$U = \sum_i \lambda_i h_i(d, f) \qquad (1)$$

formada por la combinación lineal de distintas características $h_i(d, f)$ relativas a pares bilingües de oraciones traducciones entre sí. Una buena traducción será deudora de una adecuada selección de características, que habitualmente se expresan mediante funciones logarítmicas. Por fuerza, la información de mayor relevancia en esta combinación es proporcionada por el modelo de traducción. En nuestro sistema, este modelo se expresa en función del concepto de tupla, que se describe a continuación.

2.2 La tupla

En la figura 1 se muestra el resultado del alineado de un par bilingüe de oraciones que expresan el mismo significado. Las flechas señalan las palabras vinculadas entre sí en la traducción. En el sentido de las flechas se indica el idioma inglés como fuente y el español como destino. A partir de este alineado pueden establecerse múltiples pares de secuencias de palabras de ambas lenguas que no se encuentran vinculadas a palabras fuera del par. Llamaremos a estos pares segmentos (de oración) bilingües. Por ejemplo:

(I wish a, quisiera tener un)

(a system, un sistema que)

Sin embargo no serían pares válidos

(I wish, quisiera)

(translations, traducciones correctas)

ya que el primero no incluye en la parte española todas las palabras que se relacionan con las palabras en la parte inglesa, y el segundo contiene en la parte española una palabra cuya traducción no se encuentra en la otra parte del par. Son diversos los sistemas de traducción que se basan en estos segmentos bilingües. En (Crego, Mariño y de Gispert, 2004) puede encontrarse una comparación entre algunos de ellos.

Nuestro sistema utiliza un subconjunto de estos segmentos bilingües, que llamamos tuplas, tal que satisface las dos condiciones siguientes:

- Proporciona una segmentación monótona del par bilingüe de oraciones. Es decir, en tuplas consecutivas los segmentos de una misma lengua son consecutivos en la oración monolingüe correspondiente.
- Cada tupla no puede ser descompuesta en segmentos bilingües más pequeños sin violar la condición anterior.

En la figura 1 se incluye la segmentación en tuplas: cada tupla se corresponde con un recuadro. Obsérvese que las palabras enlazadas a NULO generan una dificultad a la hora de establecer estos segmentos. En la figura 1, la secuencia de tuplas es adecuada para realizar una traducción del inglés al castellano, pero no al revés. Ello se debe a la tupla (I, NULO), que nos indica que el sujeto "I" no necesita ser traducido. Sin embargo, si se tradujese el español, NULO no es una palabra del castellano y no se encontraría presente en el texto a traducir. En realidad, esta es la razón para que una de las tuplas sea (providing, que proporcione), ya que de otro modo no se podría generar la palabra "que" al traducir el inglés. En la implementación actual de nuestro sistema no se ha hecho un estudio pormenorizado de este problema. De momento la solución adoptada es:

- Las palabras del idioma fuente enlazadas a NULO forman tupla.
- Las palabras del idioma destino enlazadas a NULO se incorporan a la tupla siguiente (esta opción proporciona mejores resultados que la contraria).

En un futuro se pretende abordar una asignación óptima de estas palabras, que considere la estadística de los contextos de la palabra enlazada a NULO o información lingüística.

2.3 El modelo de traducción

Si se realiza el alineado de todo el corpus bilingüe de entrenamiento y, posteriormente, su segmentación en tuplas, se obtiene un conjunto de secuencias de segmentos bilingües. Las propiedades estadísticas de estas secuencias pueden ser modeladas mediante cualquier técnica habitual en el modelado de lenguaje que considere a éste como una secuencia de unidades (típicamente, palabras o clases de palabras). En concreto, un *N*-grama es un modelo de amplia difusión. Mediante su uso, la probabilidad conjunta de un par bilingüe de oraciones (es decir, la probabilidad de que sean mutuas traducciones) puede expresarse

mediante la probabilidad de la secuencia de tuplas t^K en que puede segmentarse:

$$p(d,f) = \Pr\{ t^K \} = \prod_{k=1}^{K} p(t_k \mid t_{k-1}, ..., t_{k-N+1})$$

Este planteamiento es heredero de los sistemas de traducción del habla basados en autómatas de estados finitos (Vidal, 1997) (de Gispert y Mariño, 2002) y similar a (Picó et al., 2004).

2.4 Las características adicionales

Como ya se ha mencionado anteriormente, en la función que dirige la búsqueda de la mejor traducción se incluyen otras informaciones o características además del modelo de traducción:

$$h_1(d,f) = \log \prod_{k=1}^{K} p(t_k \mid t_{k-1}, ..., t_{k-N+1})$$

Actualmente, nuestro sistema incluye las siguientes características adicionales:

- Las probabilidades de traducción en cada dirección (de fuente a destino $p(d_k / f_k)$ y de destino a fuente $p(f_k / d_k)$) asignada por el modelo IBM1 a los segmentos de oración que constituyen cada tupla $t_k = (d_k, f_k)$. Ambas probabilidades se consideran informaciones independientes.

$$h_2(d,f) = \log \prod_{k=1}^{K} p(d_k / f_k)$$

$$h_3(d,f) = \log \prod_{k=1}^{K} p(f_k / d_k)$$

- La probabilidad de la oración generada para la lengua destino asignada por un *N*-grama en palabras:

$$h_4(d) = \log \prod_{i=1}^{I} p(d_i \mid d_{i-1}, ..., d_{i-N+1})$$

- Una penalización para las traducciones más cortas, que compense la tendencia a la generación de traducciones con el menor número de palabras:

$$h_5(d) = I$$

donde I es el número de palabras de la traducción hipotetizada.

3 Descripción de la tarea abordada

El sistema descrito ha sido aplicado a la traducción de intervenciones en las sesiones plenarias del Parlamento Europeo (EPPS). Las lenguas elegidas han sido inglés y español, realizándose traducciones en ambas direcciones.

En la tabla 1 se proporciona las principales estadísticas de los corpus de entrenamiento y test: número de oraciones (**orcn**), número total de palabras (**plbr**), talla de los correspondientes vocabularios (**vcblr**) y longitud media en palabras de las oraciones (**media**).

Entrenamiento				
Lng	**Orcn**	**Plbr**	**Vcblr**	**Media**
en	1.223 k	33.4 M	105 k	27.3
es		34.8 M	169 k	28.4
Test				
en	1094	26.8 k	3.9 k	24.5
es	840	22.7 k	4.0 k	27.0

Tabla 1: Estadísticas de los materiales de entrenamiento y test (M expresa millones y k miles de palabras).

El material de entrenamiento recoge las transcripciones de las sesiones desde abril de 1996 hasta septiembre de 2004. Este material es distribuido por el Parlamento Europeo a través de su página web[1]. En nuestra experimentación hemos hecho uso de la versión distribuida por RWTH de Aachen en el ámbito del proyecto TC-STAR[2].

El material de test corresponde al material utilizado en la primera evaluación realizada en el proyecto en marzo de 2005. Este material consiste en la transcripción de las sesiones del 15 al 18 de noviembre de 2004. Ha sido distribuido por ELDA[3].

En el caso del material de entrenamiento las oraciones son paralelas; es decir, el corpus está formado por parejas de frases que se traducen mutuamente. Puede observarse que el número total de palabras en el corpus de entrenamiento es muy parejo. No obstante, el número de palabras distintas (talla del vocabulario) es mucho mayor para el español. Esto puede explicarse por el carácter mucho más flexivo del español, con formas diversas para los adjetivos y, sobre todo, para los verbos.

El material de test es independiente para cada sentido de traducción. En este corpus se han encontrado 112 palabras inglesas y 46 españolas no presentes en el entrenamiento, que

[1] http://www.europarl.eu.int/
[2] http://www.tc-star.org/
[3] http://www.elda.org/

constituyen, respectivamente, el 0.4% y 0.2% del total de palabras del test. De estas palabras, son diferentes 81 y 40 en cada lengua. A efectos de evaluar la calidad de la traducción realizada se dispuso de 2 traducciones de referencia por cada frase a traducir.

4 Detalles experimentales

4.1 Preprocesado y alienamiento

Los textos de material de entrenamiento fueron tratados para individualizar todos los "tokens" (palabras, signos de puntuación, números, etc.). No se ha realizado categorización, de modo que nombres propios, números, fechas, etc. no reciben tratamiento especial. Se han eliminado los pares bilingües en el que una de las oraciones contenía más de 100 palabras o en el que el cociente entre el número de palabras de una y otra oración excedía 2.4 (fertilidad superior a 2.4).

Mediante la aplicación GIZA++ se realizó el alineamiento de los textos bilingües paralelos del material de entrenamiento, ejecutándose 5 iteraciones de los modelos IBM1 y HMM y 3 iteraciones de los modelos IBM3 e IBM4. Se obtuvo el alineamiento en las dos direcciones de traducción: tomando sucesivamente el inglés y el español como lenguas fuente. A partir de estos dos alineamientos básicos, se obtuvieron los alineamientos unión e intersección de los mismos, definidos, respectivamente, por los conjuntos unión e intersección de los enlaces establecidos en los alineamientos básicos. El primero proporciona la mejor cobertura ("recall") de los enlaces entre las palabras de ambas lenguas, que es importante para generar segmentos bilingües correctos. El segundo genera enlaces con alta precisión, que serán usados para la traducción de palabras.

4.2 Modelo de traducción

4.2.1 Selección de las tuplas

Una vez obtenido el alineamiento unión se procedió a la segmentación en tuplas del material de entrenamiento. En la tabla 2 se muestra la estadística de las tuplas: el total en ambas direcciones de traducción (que no coinciden debido a la presencia de los enlaces a NULO) y la talla del vocabulario de tuplas. En las figuras 2 y 3 se presentan los histogramas de las tuplas en función del número de apariciones en el entrenamiento y el número de traducciones diferentes que las tuplas ofrecen para una misma parte fuente. Como puede observarse la mayor parte de las tuplas aparecen muy pocas veces y abundan más las tuplas que ofrecen un número reducido de traducciones alternativas.

Número de tuplas en el entrenamiento			
sentido	**total**	**30**	**20**
es-> en	19.2 M	18.6 M	18.3 M
en -> es	18.6 M	17.7 M	17.5 M
Talla del vocabulario de tuplas			
es-> en	2.5 M	2.1 M	2.0 M
en -> es	2.5 M	2.0 M	1.9 M

Tabla 2: Estadística de las tuplas (M significa millón de tuplas).

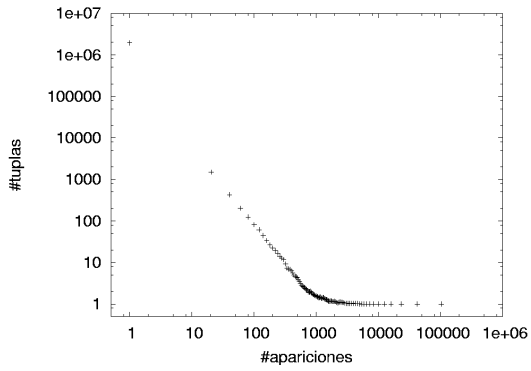

Figura 2: Histograma de tuplas en función del número de apariciones.

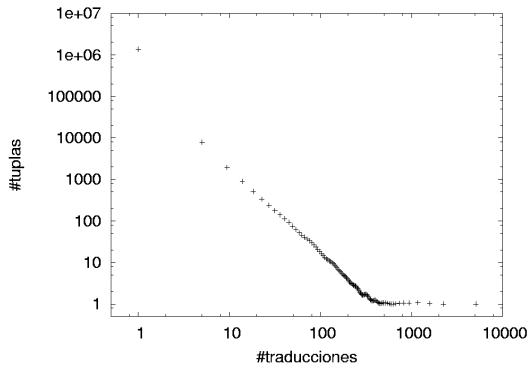

Figura 3: Histograma de tuplas en función del número de traducciones disponibles para la parte del idioma fuente.

A efectos de simplificar el sistema de traducción, el vocabulario de tuplas se limitó a aquellas que, para un mismo segmento fuente, contenían hasta 30 traducciones del inglés y

hasta 20 del español. Estos límites se determinaron experimentalmente de modo que la calidad de las traducciones no se viese afectada. Así, se redujo el vocabulario de tuplas en un 20% (véase la tabla 2).

4.2.2 Estimación del modelo

Para estimar el modelo se utilizó la herramienta SRILM (Stolcke, 2002) de libre distribución. En este proceso se limitó el vocabulario del modelo de lenguaje bilingüe a las tuplas seleccionadas conforme se ha explicado anteriormente, al que se añadió una traducción (tupla) para todas aquellas palabras que no aparecían solas en ninguna tupla (por lo que no se podrían traducir si en el test apareciesen en un contexto distinto a los existentes en el material de entrenamiento). Estas tuplas de traducción para las palabras "incrustadas" ("embedded") fueron generadas a partir del alineamiento intersección.

Como técnica de suavizado se utilizó el método de Kneser-Ney e interpolación lineal (Kneser and Ney, 1995).

El modelo generado fue un trigrama (N=3) de tuplas. En la tabla 3 se indica el número de unigramas, bigramas y trigramas contenidos en los modelos para cada sentido de traducción. La tabla 4 proporciona la perplejidad de los modelos de traducción en ambos sentidos evaluada en el material de entrenamiento.

	es->en	en->es
1-gramas	2.039.514	2.022.823
2-gramas	6.008.896	6.091.809
3-gramas	1.797.578	1.747.148

Tabla 3: Histograma de n-gramas en los modelos de traducción.

sentido	perplejidad
es->en	88.1
en->es	89.6

Tabla 4: Perplejidad en el material de entrenamiento de los modelos de traducción.

4.3 Características adicionales

La probabilidad de traducción asignada por el modelo IBM1 a los segmentos de oración que constituyen cada tupla se calcula mediante la expresión

$$p(d_k / f_k) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(d_k^i / f_k^j)$$

donde J e I son las longitudes en palabras de las partes fuente y destino, y f_k^j y d_k^i son, respectivamente, las palabras de cada una de dichas partes. Las probabilidades condicionales de traducción entre palabras se toman del alineamiento proporcionado por GIZA++ para el sentido de la traducción. Análogamente, se determina la probabilidad para la dirección contraria.

Para estimar el modelo para el lenguaje destino se volvió a hacer uso de la herramienta SRILM, con N=3 y la técnica de suavizado de Kneser-Ney. Como material de entrenamiento se hizo uso de los textos de la lengua correspondiente en el corpus bilingüe. En la tabla 5 se indica la perplejidad de estos modelos medida sobre el material de entrenamiento.

idioma	perplejidad
en	39.5
es	38.5

Tabla 5: Perplejidad en el material de entrenamiento de los modelos de lenguaje.

Los coeficientes λ_i de la combinación log-lineal (1) se optimizaron mediante el algoritmo Simplex (Press et al., 2002) para maximizar la medida de calidad BLEU (Papineni et al., 2002) de la traducción de 500 oraciones de un corpus de desarrollo que contenía 3 traducciones de referencia por cada texto origen. Este corpus fue extraído de las intervenciones en el plenario del Parlamento Europeo entre el 25 y el 28 de octubre de 2004. En la tabla 6 se proporcionan los valores de los coeficientes λ_i para las configuraciones de los sistemas de traducción que utilizan todas las características. Dados estos valores, puede decirse que todas las características tienen un grado significativo de influencia en la traducción, aunque con diferentes matices en función del sentido de la misma.

sentido	λ_1	λ_2	λ_3	λ_4	λ_5
es->en	1	0.48	0.13	0.48	0.28
en->es	1	0.23	0.18	0.80	0.75

Tabla 6: Coeficientes de la combinación log-lineal para los sistemas completos.

4.4 El algoritmo de traducción

La traducción del material de test fue llevada a cabo mediante la herramienta MARIE (Crego, Mariño y de Gispert, 2005), que maximiza la función U en (1) mediante un algoritmo de programación lineal de búsqueda en haz.

La búsqueda construye traducciones parciales (hipótesis), que se conservan en diferentes listas. Cada lista contiene aquellas hipótesis que han traducido las mismas palabras de la frase de entrada. Las hipótesis de cada lista se ordenan según la puntuación acumulada, lo que permite podar por separado en cada lista. Se mantienen las mejores hipótesis (poda por histograma) y aquéllas que tienen asignada una puntuación próxima a la mejor hipótesis de la lista (poda por umbral).

El algoritmo de búsqueda permite avanzar en la traducción cubriendo partes de la frase de origen de manera desordenada (distorsión), lo que da lugar a una traducción no monótona. Esta posibilidad no ha sido utilizada en los experimentos realizados en esta comunicación, dado que sólo es aconsejable en la traducción de pares de lenguas con necesidad de reordenamientos lejanos.

5 Resultados alcanzados

En las tablas 7 y 8 se muestran los resultados de la evaluación de las traducciones obtenidas para el material de test. La tabla 7 recoge el porcentaje de error en palabras (mWER) y el BLEU de la traducción al inglés de los textos en español. El mWER se determina a partir del error en la referencia de traducción para la que se produce menor error. El BLEU es una medida basada en el número de *N*-gramas (*N* de 1 a 4) correctos en la traducción en relación con los que contienen las referencias. Los algoritmos de evaluación utilizados fueron los oficiales del proyecto TC-STAR facilitados por ELDA, con distinción de mayúsculas y minúsculas. En la tabla se incluyen las evaluaciones para 4 configuraciones del sistema de traducción:

- Sistema con el modelo de traducción de tuplas únicamente (1).
- Sistema con el modelo de traducción de tuplas y las probabilidades de traducción de los segmentos de las tuplas (1, 2, 3).
- Sistema con el modelo de traducción de tuplas, modelo de lenguaje destino y la penalización para las traducciones cortas (1, 4, 5).
- Sistema completo con todas las informaciones (1, 2, 3, 4, 5).

En la tabla 8 se resumen las evaluaciones de las traducciones al español de los textos en inglés.

informaciones	mWER	BLEU
1	39.55	0.476
1, 2, 3	35.65	0.537
1, 4, 5	39.61	0.485
1, 2, 3, 4, 5	34.91	0.543

Tabla 7: Evaluación de la traducción en el sentido del español al inglés.

informaciones	mWER	BLEU
1	44.45	0.428
1, 2, 3	41.69	0.450
1, 4, 5	44.67	0.436
1, 2, 3, 4, 5	40.96	0.466

Tabla 8: Evaluación de la traducción en el sentido del inglés al español.

6 Discusión

En primer lugar debe señalarse que las evaluaciones obtenidas se comparan favorablemente con las alcanzadas por los sistemas que describen el estado actual del arte (TC-STAR, 2005).

Por otro lado, de la comparación de ambas tablas se desprende que la traducción al inglés es de mayor calidad que la traducción al español. Esto puede explicarse por el carácter más flexivo del español que se ha mencionado anteriormente. En ocasiones la traducción del lema es correcta pero no la instancia producida: error en número, género, tiempo verbal, persona, etc. (ver el ejemplo más adelante).

En cuanto a la aportación de las diversas informaciones a la calidad de las traducciones generadas puede establecerse:

- La limitada influencia del modelo del lenguaje destino y la penalización de las traducciones cortas.
- La importante contribución de la probabilidad de traducción de los componentes de las tuplas.

Se puede señalar incluso un incremento del mWER al incluir el modelo del idioma destino al modelo de traducción, aunque se observe una mejoría del BLEU. Este comportamiento puede comprenderse si se tiene en cuenta que la

optimización de los coeficientes λ_i de (1) se realiza en función del BLEU.

La influencia de la traducción de los componentes de la tupla sugiere el interés de explorar el uso de esta probabilidad como criterio para seleccionar las tuplas en el momento de la estimación del modelo de traducción.

En el siguiente ejemplo de traducción se muestran los tipos de error más frecuente:

Durante una semana ~~el~~ **las** ~~americano~~ fuerzas **americanas** de ocupación ~~comete~~ **han cometido** un crimen abominable en la ciudad de Faluya en Iraq.

Se han tachado las palabras que corresponden a errores en la traducción y se han añadido en negrita las correcciones. Las fuentes de error mostradas son:

- Falta de concordancia de género y número, que podrían subsanarse con el uso de información morfosintáctica.
- Orden equivocado entre nombre y adjetivo, que podría corregirse con la capacidad de reordenamiento del algoritmo de búsqueda.
- Defecto en la traducción de tiempos verbales y personas. Su enmienda puede ser obtenida mediante el uso de información lingüística (de Gispert, 2005).

7 Agradecimientos

Este trabajo ha sido financiado parcialmente por la CICYT a través del proyecto TIC2002-04447-C02 (ALIADO) y la Unión Europea mediante el proyecto FP6-506738 (TC-STAR).

Bibliografía

Berger, A., Della Pietra, S. y Della Pietra, V. 1996. A Maximun Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1): 39-72.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J.D., Mercer, D. y Rocín, P.S. 1990. A statistical approach to machine translation. *Computational Linguistics,* 16(2): 79-85.

Crego, J.M., Mariño, J.B. y de Gispert, A. 2004. Finite-state-based and Phrase-based Statistical Machine Translation. En *Proc. of the Int. Conf. on Spoken Language Processing*. Jeju, Corea.

Crego, J.M., Mariño, J.B. y de Gispert, A. 2005. Algoritmo de decodificación de traducción automática estocástica basada en N-gramas. *SEPLN'05*. Granada.

De Gispert, A. y Mariño, J.B. 2002. Using X-grams for speech-to-speech translation. En *Proc. of the Int. Conf. on Spoken Language Processing*, páginas 1885-1888. Denver, CO (USA).

De Gispert, A. 2005. Phrase linguistic classification and generalization for improving statistical machine translation. Aceptado en *ACL'05 Student Workshop*.

Kneser, R. y Ney, H. 1995. Improved backing-off for m-gram language modelling. En *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, páginas 49-52, Detroit. MI (USA).

Och, F.J. y Ney, H. 2002. Discriminative training and maximum entropy models for statistical machina translation. En *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, páginas 295-302.

Och, F.J. 2003. GIZA++. http://www-16.informatik.rwth-aachen.de/~och/software/giza++.html .

Papineni, K., Roukos, S., Ward, T. y Zhu, W-J. 2002. BLEU: a method for automatic evaluation of machine translation. En *Proc. of the 40th Ann. Conf. of the ACL*. Philadelphia, PA (USA).

Picó, D., Tomás, J. y Casacuberta, F. 2004. GIATI: a general methodology for finite-state translation using alignments. En *Proc. SSPR2004 and SPR2004*. Lisboa. Portugal.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. y Flannery, B.P. 2002. *Numerical Recipies in C++: the Art of Scientific Computing*. Cambridge University Press.

Stolcke, A. 2002. SRILM: an extensible language modelling toolkit. En *Proc. of the Int. Conf. on Spoken Language Processing*, páginas 901-904. Denver, CO (USA).

TC-STAR. 2005. Deliverable D5: SLT progress report. http://www.tc-star.org/documents/deliverable/Deliv_D5_Total_21May05.pdf.

Vidal, E. 1997. Finite-State Speech-to-Speech Translation. En Proc. of 1997 IEEE Internacional Conference on Acoustics, Speech and Signal Processing, páginas: 111-114. Munich, Germany.