ESTIMACIÓN DE MODELOS DE TRADUCCIÓN DE SECUENCIAS DE PALABRAS A PARTIR DE CORPUS MUY GRANDES MEDIANTE THOT

Daniel Ortiz-Martínez[1] , *Ismael García-Varea*[2] , *Francisco Casacuberta*[1]

[1] Universidad Politécnica de Valencia

[2] Universidad de Castilla-la Mancha

RESUMEN

En el ámbito de la traducción automática estadística, los últimos tiempos se han caracterizado por la popularización de los modelos de traducción basados en secuencias de palabras, así como por la aparición de corpus bilingües más y más grandes como el bien conocido corpus EUROPARL. La coincidencia de estos dos acontecimientos ha planteado un problema importante debido a que los modelos de traducción basados en secuencias de palabras requieren un espacio de almacenamiento considerable cuando se estiman a partir de grandes corpus de entrenamiento. Para resolver este problema así como muchos otros relacionados con la estimación y la aplicación de modelos estadísticos de secuencias de palabras, se ha desarrollado la herramienta de libre uso denominada Thot, cuya funcionalidad básica se describe en este artículo. Adicionalmente, se incluyen experimentos de traducción para el corpus EUROPARL usando modelos de traducción generados con Thot y decodificadores basados en el estado del arte.

1. INTRODUCCIÓN

Desde comienzos de la decada de los 90, el interés en la disciplina de la traducción automática estadística (TAE) no ha hecho sino incrementarse, debido a los buenos resultados que ha obtenido al aplicarse en corpus de dominio restringido.

El proceso de traducción puede formularse desde un punto de vista estadístico de la siguiente forma: Sea $f_1^J = f_1 \dots f_J$ la frase origen que queremos traducir en su equivalente en el lenguaje destino $e_1^I = e_1 \dots e_I$. Se considera que cualquier posible frase de la lengua destino es traducción de la frase origen con una probabilidad a posteriori determinada $Pr(e_1^I|f_1^J)$. Según la regla de Bayes, la frase destino que buscamos $\hat{e}_1^I$ será aquella que maximiza[1] el producto del modelo del lenguaje destino $Pr(e_1^I)$ y el modelo de traducción $Pr(f_1^J|e_1^I)$. La ecuación que modela el proceso es la siguiente:

$$\hat{e}_1^I = \arg\max_{e_1^I} \{Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)\} \qquad (1)$$

En función de cómo se concibe la relación entre las palabras origen y destino, se han propuesto diferentes *modelos de traducción*; que intentan explicar la manera en que se genera la frase destino a partir de la frase origen. Esta relación se describe a través del concepto de *alineamiento*; dando lugar a diferentes *modelos estadísticos de alineamiento*. Los modelos de alineamiento IBM y HMM fueron propuestos en [1] y en [2] respectivamente. Estos modelos pertenecen a la categoría de los *modelos basados en palabras*, ya que asumen que en el proceso de traducción se establecen relaciones entre palabras individuales de las frases origen y destino. En los últimos tiempos, se ha demostrado que estos modelos no capturan adecuadamente la información de contexto a la hora de traducir, por lo que se han propuesto modelos que trabajan con grupos de palabras en lugar de palabras, los cuales han constituído una mejora con respecto a los modelos basados en palabras originales descritos en [1]. Se han propuesto diferentes modelos que trabajan con grupos de palabras, por ejemplo, los *modelos sintácticos* se describen en [3], los *alignment templates* se describen en [4]. Adicionalmente, en [5, 6, 7, 8] se ha descrito una reformulación del enfoque *alignment template* que ha dado lugar a los modelos llamados *phrase based*, a los cuales nos vamos a referir en castellano como *modelos de secuencias de palabras*.

En el transcurso de la investigación en traducción automática estadística, se han ido desarrollando herramientas software con objeto de ayudar a los investigadores a mejorar sus sistemas de traducción. Estas herramientas abarcan desde software de entrenamiento de modelos basados en palabra (como la herramienta Giza++) hasta algunos sistemas de traducción específicos, incluyendo el decodificador para modelos basados en secuencias denominado Pharaoh [9]. En el estado del arte actual, resulta muy necesaria una herramienta de entrenamiento de modelos de secuencias para ayudar en la mejora de los sistemas de traducción. En este artículo se presenta una herramienta de libre uso pensada para tal fin.

Este trabajo ha sido parcialmente subvencionado por el proyecto CICYT TIC2003-08681-C02-02, la *Agencia Valenciana de Ciencia y Tecnología* dentro del contrato GRUPOS03/031, la *Generalitat Valenciana*, y el proyecto HERMES II (Vicerrectorado de Investigación - UCLM-06/07)

[1] Obsérvese que la expresión debe ser maximizada también para I; sin embargo, para simplificar se asume que I es conocida.

2. MODELOS DE SECUENCIAS DE PALABRAS

Como ya hemos comentado, éstos modelos surgen como alternativa a los modelos de palabras (o más comúnmente conocidos como modelos de IBM [1]) para superar las limitaciones que presentan, de modo que en lugar de trabajar con diccionarios estadísticos de palabras ($Pr(f_j|e_i)$), trabajan con diccionarios estadísticos de secuencias ($Pr(\tilde{f}_k|\tilde{e}_k)$).

La traducción, usando modelos de secuencias de palabras, de una frase de entrada f_1^J en la frase destino equivalente e_1^I consiste, desde un punto de vista generativo, en escoger la forma en que dicha frase de entrada es segmentada en K secuencias $f_1^J = \tilde{f}_1^K$, seleccionar las secuencias en el lenguaje destino que traducen las secuencias origen y, por último, reordenar; con lo que terminamos obteniendo $e_1^I = \tilde{e}_1^K$. Podemos asumir que las relaciones entre las palabras origen y destino se sumarizan mediante una variable oculta $\tilde{\mathbf{a}} = \tilde{a}_1^K$, que contiene todas las decisiones que se hacen durante la historia generativa.

$$\begin{aligned} Pr(f_1^J|e_1^I) &= \sum_{\tilde{\mathbf{a}}} Pr(\tilde{\mathbf{a}}, \tilde{f}_1^J|\tilde{e}_1^I) \\ &= \sum_{\tilde{\mathbf{a}}} Pr(\tilde{\mathbf{a}}|\tilde{e}_1^I) Pr(\tilde{f}_1^J|\tilde{\mathbf{a}}, \tilde{e}_1^I) \end{aligned} \quad (2)$$

Se pueden hacer diferentes asunciones a partir de la ecuación anterior, aunque lo normal es que los modelos terminen convirtiéndose en diccionarios estadísticos de segmentos. Por ejemplo, en [5] se propone el siguiente modelo:

$$p_\theta(f_1^J, e_1^I) = \alpha(e_1^I) \sum_{\tilde{\mathbf{a}}} \prod_{k=1}^{K} p(\tilde{f}_k|\tilde{e}_{\tilde{a}_k}) \quad (3)$$

donde $\tilde{a}_k$ denota el índice de la secuencia origen $\tilde{e}$ que se alinea con la k'ésima secuencia destino $\tilde{f}_k$ y se asume que todas la posibles segmentaciones tienen la misma probabilidad.

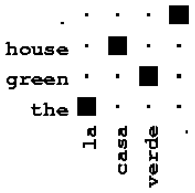
Estimación de parámetros del modelo

Existen principalmente tres maneras de obtener los pares de secuencias, tal y como se describe en [8]:

1. A partir de matrices de alineamiento a nivel de palabra.
2. Mediante criterios sintácticos (véase [3])
3. A partir de alineamientos a nivel de frase por medio de un modelo de probabilidad conjunta (véase [6]).

En este artículo nos vamos a centrar en el primer método, en el que nos basamos en matrices de alineamiento a nivel de palabra para extraer los pares de secuencias. Dichas matrices se obtendrán automáticamente como subproducto de la estimación de modelos IBM. Concretamente, dado un par de frases y su correspondiente matriz de alineamiento A, se extraerán aquellos pares de secuencias que sean *consistentes* con la matriz de alineamiento. La condición de consistencia viene dada por la ecuación (4) [4]. La Figura 1 muestra un ejemplo en el que se da un par de frases con su matriz de alineamiento, y el conjunto de todos los pares bilingües consistentes que se pueden extraer.

$$\begin{aligned} \mathcal{BP}(f_1^J, e_1^I, A) = \{(f_j^{j+m}, e_i^{i+n} : \forall (i', j') \in A : \\ j \leq j' \leq j+m \iff i \leq i' \leq i+n\} \end{aligned} \quad (4)$$



secuencia origen	secuencia destino
La	the
casa	house
verde	green
casa verde	green house
La casa verde	the green house
.	.
casa verde .	green house .
La casa verde .	the green house .

Figura 1. Conjunto de pares bilingües consistentes (derecha) para una matriz de alineamiento dada (izquierda).

Una vez que se han extraído los pares de secuencias, las probabilidades del modelo se calculan a través de las frecuencias relativas de los mismos.

Un importante inconveniente que poseen los modelos de secuencias es la gran cantidad de espacio en memoria requerida por sus parámetros, por lo que serán necesarias técnicas especiales para su estimación y manejo cuando se trabaja con corpus muy grandes.

3. DESCRIPCIÓN DE LA HERRAMIENTA

La herramienta Thot ha sido desarrollada utilizando el lenguaje de programación C++. Los principios de diseño que han dirigido el proceso de desarrollo han sido: eficiencia, extensibilidad, flexibilidad (trabaja con distintos y bien conocidos formatos de datos) y usabilidad .

En los apartados siguientes describiremos la funcionalidad básica ofrecida por Thot.

3.1. Operaciones entre matrices de alineamiento

Como se comentó en el apartado 2 es habitual la aplicación de operaciones entre matrices de alineamiento con objeto de mejorarlas.Thot incorpora las siguientes operaciones sobre matrices de alineamiento:

Unión : Obtiene la unión de dos matrices.

Intersección : Obtiene la intersección de dos matrices.

Suma : Obtiene la suma de dos o más matrices.

Simetrización : Obtiene una matriz de alineamiento a medio camino entre la unión y la intersección de dos matrices. Esta operación fue propuesta por primera vez en [4] y existen diferentes versiones.

El formato de fichero para los alineamientos es el generado por la herramienta Giza++. El formato de salida puede ser el formato de Giza++ así como otros dos: en forma de matriz bidimensional, o bien un formato intermedio que se puede convertir fácilmente en otros.

3.2. Modalidades de estimación de modelos de secuencias

Thot permite estimar modelos de secuencias a partir de matrices de alineamiento a nivel de palabra en formato Giza++, tal y como se explicó en la sección 2, estimación a la que nos referiremos de ahora en adelante como estimación *RF* por sus siglas en inglés (*relative frequency*).

La estrategia de estimación RF es heurística por dos razones. En primer lugar los pares bilingües son extraídos a partir de matrices de alineamiento a nivel de palabra, por lo que nos vemos obligados a aplicar un criterio heurístico de consistencia sobre los pares. En segundo lugar, los pares extraídos no son considerados como parte de segmentaciones bilingües completas a la hora de estimar las probabilidades. El primer problema no puede resolverse sin cambiar por completo la estrategia de estimación. Para el segundo, en cambio, pueden ofrecerse alternativas de solución.

Con este propósito, la herramienta implementa una nueva propuesta de estimación que hemos llamado estimación por *pseudo máxima verosimilitud* o estimación pML por sus siglas en inglés (*pseudo maximum-likelihood*)[2] que es diferente la la estrategia de estimación habitual. En concreto, dicha estimación consiste en un proceso de tres pasos que se repiten para cada par de frases del corpus con su correspondiente matriz de alineamiento (f_1^J, e_1^I, A):

1. Obtener el conjunto $\mathcal{BP}(f_1^J, e_1^I, A)$ de todos los pares de secuencias consistentes.

2. Obtener el conjunto $\mathcal{S}_{\mathcal{BP}(f_1^J, e_1^I, A)}$ de todas las segmentaciones bilingües posibles[3] del par (f_1^J, e_1^I) que pueden construirse utilizando pares bilingües consistentes.

3. Actualizar los contadores (en realidad contadores fraccionales) para cada par de secuencias distinto $(\tilde{f}, \tilde{e})$ en el conjunto $\mathcal{S}_{\mathcal{BP}(f_1^J, e_1^I, A)}$:

$$fracCount(\tilde{f}, \tilde{e}) += \frac{N(\tilde{f}, \tilde{e})}{|\mathcal{S}_{\mathcal{BP}(f_1^J, e_1^I, A)}|}$$

donde $N(\tilde{f}, \tilde{e})$ es la cantidad de veces que aparece el par $(\tilde{f}, \tilde{e})$ en $\mathcal{S}_{\mathcal{BP}(f_1^J, e_1^I, A)}$, y $|\cdot|$ denota la talla de un conjunto.

Finalmente, la probabilidad de cada par de secuencias $(\tilde{f}, \tilde{e})$ se calcula de la siguiente forma:

$$p(\tilde{f}|\tilde{e}) = \frac{fracCount(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}} fracCount(\tilde{f}, \tilde{e})}$$

Para clarificar lo anterior puede consultarse la Figura 2 que muestra el conjunto de posibles segmentaciones bilingües de un par de frases para la matriz de alineamiento que se puede ver en la Figura 1.

El paso 2 implica que si un par bilingüe no puede formar parte de una segmentación bilingüe para un par de frases dado, dicho par bilingüe no será extraído. Por esta razón, la estimación pML extrae una menor cantidad de pares que la estimación RF.

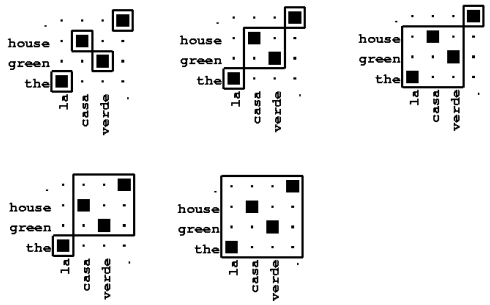


Figura 2. Posibles segmentaciones bilingües para una matriz de alineamiento a nivel de palabra dada.

Adicionalmente, la estimación pML permite obtener modelos de secuencias más completos, incluyendo or ejemplo, un submodelo para la talla de la segmentación K, funcionalidad que ha sido incluída en la herramienta. Por otro lado, y como principal desventaja, esta modalidad de estimación tiene un elevado coste computacional debido a la necesidad de obtener la segmentación bilingüe para cada par de frases.

3.3. Estimación a partir de corpus muy grandes

Uno de los mayores inconvenientes que presentan los modelos de traducción de secuencias reside en sus elevados requerimientos de memoria. Existen diversas alternativas para paliar este importante problema. Una de ellas consiste en limitar el tamaño máximo que pueden tener las secuencias, con lo que se corre el riesgo de obtener peores modelos (no obstante, existen resultados empíricos [8] que demuestran que dicha limitación puede imponerse sin que se produzca una disminución apreciable en la calidad de la traducción).

Otra posibilidad consiste, tanto para la estimación RF como pML, en quedarse con aquellos pares bilingües que satisfagan la restricción de monotonicidad. Todas estas variantes han sido implementadas por la herramienta, cuya salida puede obtenerse en un formato nativo propio, o en el formato de entrada esperado por el traductor de libre uso Pharaoh [9].

[2] Se ha elegido este nombre porque consideramos que la propuesta sería equivalente a la primera iteración del algoritmo EM, que debería utilizarse en una estimación más rigurosa

[3] Una segmentación bilingüe de talla K de un par de frases (f_1^J, e_1^I) se define como una tripleta $(\tilde{f}_1^K, \tilde{e}_1^K, \tilde{a}_1^K)$, donde $\tilde{a}_1^K$ es una correspondencia uno a uno entre las secuencias en las que se ha dividido ambas frases.

No obstante, las técnicas anteriores son insuficientes a la hora de manejar los enormes corpus que se vienen presentando actualmente. Una posible solución a este problema, que permite el entrenamiento de corpus de tamaño arbitrario, viene dada por el Algoritmo 1. Dicho algoritmo se apoya en el trabajo con cuentas en lugar de probabilidades. Consiste en entrenar el corpus dividiéndolo en trozos de tamaño fijo (tam_fragm) volcando cada uno de los submodelos que se obtienen en disco. Una vez volcados los modelos se juntan en un único fichero, se ordenan lexicográficamente y se fusionan las cuentas, dando lugar a un modelo idéntico al que se obtendría de haber entrenado el corpus completo de una sola vez.

Algoritmo 1 algoritmo_entrenamiento (fichAlin)

```
partir(fichAlin,tam_fragm)
para_todo f fragmento de fichAlin hacer
  entrena(f) >> fichCuentas
fin_para
ordenarCuentas(fichCuentas) > fichCuentasOrdenado
fundirCuentas(fichCuentasOrdenado) > modeloSecuencias
```

La única sobrecarga temporal que introduce el algoritmo se debe a la necesidad de ordenar el fichero de cuentas resultante. Creemos que el algoritmo es claramente susceptible de ser paralelizado, por lo que la mencionada sobrecarga temporal puede ser contrarrestada.

3.4. Segmentación de corpus bilingües

Dado un par de frases (f_1^J, e_1^I) y una matriz de alineamiento a nivel de palabra, la herramienta incorpora una funcionalidad que permite obtener la mejor segmentación bilingüe posible en K bisegmentos, e implícitamente el mejor alineamiento a nivel de secuencias $\tilde{a}_1^K$ (o alineamiento de Viterbi) entre ellas, de acuerdo con el siguiente algoritmo:

1. Para todo $K \in \{1 \cdots \text{mín}(J, I)\}$
   - *a*) Extraer todas las segmentaciones bilingües de talla K sujeto a las restricciones $A(f_1^J, e_1^I)$.
   - *b*) Computar y almacenar la probabilidad $p(\tilde{f}_1^K, \tilde{a}_1^K | \tilde{e}_1^K)$ de dichas segmentaciones.
2. Devolver la segmentación bilingüe $(\tilde{f}_1^K, \tilde{e}_1^K, \tilde{a}_1^K)$ de mayor probabilidad.

donde $p(\tilde{f}_1^K, \tilde{a}_1^K | \tilde{e}_1^K) = \prod_{k=1}^K p(\tilde{f}_{\tilde{a}_k} | \tilde{e}_k)$

4. EXPERIMENTOS

En esta sección se presentan resultados utilizando la funcionalidad principal que incorpora la herramienta Thot. En la experimentación se han utilizado tres corpus bien conocidos en el ámbito de la traducción automática estadística, a saber, EUTRANS-I, HANSARDS y EUROPARL cuyas principales características se muestran en la tabla 1.

4.1. Experimentos de generación de alineamientos

Para experimentar con la obtención de las mejores segmentaciones bilingües utilizamos un subconjunto del conjunto de test de EUTRANS-I consistente en 40 pares de frases aleatoriamente seleccionadas. Este corpus fue segmentado por lingüistas [10].

La Tabla 2 muestras las medidas estándar de calidad de segmentación *Recall*, *Precision*, y *F-measure* para tres técnicas diferentes de segmentación bilingüe, incluyendo la que proporciona la herramienta Thot. Las otras dos consisten en la técnica de alineamientos recursivos (RECalign) y los alineamientos de GIATI (GIATIalign) que se describen y utilizan en [10].

Tal y como muestra la Tabla 2, la calidad de segmentación bilingüe alcanzada por Thot mejora la que obtienen las otras dos técnicas.

Técnica	Recall	Precision	F-measure
RECalign	52.96	79.01	**63.41**
GIATIalign	39.99	85.52	**54.50**
Thot	72.58	65.49	**68.85**

Tabla 2. Calidad de segmentación bilingüe para 40 pares de frases extraídas del corpus EUTRANS-I.

4.2. Experimentos de traducción

Se llevaron a cabo experimentos de traducción utilizando la funcionalidad de la herramienta Thot y el traductor Pharaoh; a saber: operaciones entre alineamientos, estimación RF y pML y su aplicación en experimentos de calidad de traducción. Para los experimentos se han usado las medidas de error bien conocidas *Word Error Rate* (WER), *Position independent Error Rate* (PER) y *Bleu*.

Operaciones sobre matrices de alineamiento

Usando la funcionalidad que provee la herramienta, se estimó un modelo de secuencias RF con objeto de traducir el conjunto de test del corpus EUTRANS-I mediante el traductor Pharaoh. El modelo se estimó a partir de un conjunto de matrices de alineamiento a nivel de palabra que fueron obtenidos aplicando distintas operaciones sobre matrices. El tamaño máximo de segmento fue establecido en 6.

La Tabla 3 muestra las medidas WER, PER, Bleu y el número de secuencias extraídas según la operación aplicada (véase la sección 3.1). Como muestra la Tabla 3, la operación de simetrización produce los mejores resultados. Tal y como se esperaba, los peores resultados se obtienen cuando no se aplica ninguna operación. La operación de intersección extrae un mayor número de pares bilingües debido a que la matrices presentan una mayor cantidad de palabras no alineadas.

Estimación RF contra pML

Se llevaron a cabo experimentos con el corpus de

		EuTrans-I		Hansards		Europarl	
		Español	**Inglés**	**Francés**	**Inglés**	**Español**	**Inglés**
Entrenamiento	Frases	10 000		128 000		730 740	
	Palabras	97 131	99 292	2 062 403	1 929 186	15 725 136	15 222 505
	Vocabulario	686	513	37 542	29 414	113886	72 742
Test	Frases	2 996		500		3 064	
	Palabras	35 023	35 590	3 890	3 929	91 730	85 232
	Perplejidad (Trigramas)	–	3.62	–	79.4	–	115.0

Tabla 1. Estadísticas de los corpus EuTrans-I y Hansards

Op	WER	PER	Bleu	#Phrases
ninguna	8.0	6.7	0.894	29809
and	7.2	6.4	0.895	59715
or	8.0	5.6	0.900	14450
sum	8.0	5.7	0.900	14450
symmetr.	7.3	6.1	0.907	18686

Tabla 3. Efecto de las operaciones sobre matrices de alineamiento, tamaño máximo de secuencia=6, estimación no monótona RF, sobre el corpus EuTrans-I.

EuTrans-I aplicando las diferentes modalidades de estimación descritas en la sección 3.2.

La Tabla 4 muestra una comparativa entre las técnicas de estimación RF y pML en sus versiones monótona y no monótona. En la tabla se muestran: el número de pares bilingües extraídos (no se aplicaron operaciones sobre alineamientos y el tamaño máximo de secuencia se fijó en 6), el coste temporal del entrenamiento[4] y las medidas de error WER, PER y *Bleu* obtenidas cuando se traduce el corpus de test de la tarea EuTrans-I (de nuevo sin usar operaciones con matrices de alineamiento y con un tamaño máximo de secuencia igual a 6). Como era de esperar, la extracción monótona genera menos pares, y la estimación pML requiere mucho más tiempo que la RF, lo que se debe a lo comentado en la sección 3.2.

Estimación	Tiempo	WER	PER	Bleu	#Pares
RF mon	29	8.7	7.3	0.884	27429
RF	31	8.0	6.7	0.894	29809
pML mon	1969	8.6	7.2	0.890	25262
pML	2213	7.9	6.6	0.901	27788

Tabla 4. Comparación entre las técnicas de estimación RF y PML para el corpus de EuTrans-I.

En lo que respecta a las tasas de error obtenidas en experimentos de traducción, como puede observarse, la estimación pML obtiene resultados ligeramente mejores que la RF. Este resultado era de esperar, ya que se computó la log-verosimilitud según la ecuación 3 para ambas modalidades de estimación, y se observó que era mejor para la estimación pML, tanto para el corpus de entrenamiento como para el de test (ésto también es cierto si se sigue la aproximación del máximo).

Influencia del tamaño máximo de secuencia

Se llevaron a cabo experimentos para determinar el efecto del tamaño máximo de secuencia. La Tabla 5 muestra la influencia de este parámetro en la estimación RF. En la tabla se dan la medidas de error habituales, así como el tiempo de estimación y la cantidad de pares extraídos. Tal como muestra la Tabla 5, valores superiores a 4 no mejoran los resultados apreciablemente pero incrementan el tiempo de estimación. Se ha observado la misma situación para el caso de la estimación pML.

	Tiempo	WER	PER	Bleu	#pares
1	12.760	35.8	31.8	0.567	1656
3	17.370	9.8	8.4	0.867	10761
5	25.600	8.2	6.8	0.892	23871
7	38.320	8.0	6.7	0.894	33991

Tabla 5. Influencia del tamaño máximo de secuencia, estimación RF, para el corpus EuTrans-I.

Experimentos de calidad de la traducción

Finalmente, se realizaron experimentos de calidad de la traducción ajustando los parámetros de la herramienta Thot y los del traductor Pharaoh. En concreto, se estimó un modelo RF a partir de matrices de alineamiento simetrizadas. El tamaño máximo de secuencia se estableció en 6.

La Tabla 6 muestra las medidas de error WER, PER y *Bleu* para las tareas de EuTrans-I, Hansards y Europarl como resultado de traducir los conjuntos de test descritos en la tabla 1. Se comparó la calidad de la traducción obtenida por Pharaoh con la obtenida por otra herramienta de traducción: el *ISI ReWrite Decoder*, un traductor voraz de libre uso (ver [11]). En general, y como era de esperar, para todas las tareas estudiadas, los resultados obtenidos por Pharaoh fueron mucho mejores que los que se obtuvieron con el traductor voraz. Adicionalmente, los resultados obtenidos por Pharaoh podrían mejorarse ajustando los pesos que asigna el traductor a los modelos de traducción y lenguaje.

[4]Los resultados fueron obtenidos en un PC con procesador 1.6Ghz AMD Athlon y 512 MB de memoria utilizando Linux como sistema operativo. Todos los tiempos se dan en segundos.

Tarea	WER	PER	Bleu
EUTRANS-I	25.2/6.7	22.3/5.3	0.55/0.90
HANSARDS	57.0/52.8	52.0/48.1	0.22/0.31
EUROPARL	64.8/61.4	48.8/45.8	0.17/0.25

Tabla 6. Experimentos de calidad de la traducción para distintas tareas de traducción y los traductores *ISI ReWrite Decoder* y Pharaoh(izda. y dcha. respectivamente).

5. CONCLUSIONES

En este artículo hemos presentado las principales características de la herramienta Thot, que está disponible como software *open source* en *Sourceforge*: `http://thot.sourceforge.net/`.

El principal propósito del toolkit consiste en proporcionar una manera rápida y efectiva de entrenar modelos estadísticos de secuencias con objeto de ser utilizados en traducción automática u otras tareas relacionadas con el procesamiento del lenguaje natural. Se ha prestado especial atención al problema de la estimación a partir de corpus muy grandes, que tan importante viene siendo en los últimos tiempos.

Las principales características que ofrece la herramienta son:

- Diferentes formas de combinar alineamientos a nivel de palabra con vistas a la obtención de mejores matrices de alineamiento y mejores modelos de secuencias.
- Distintas modalidades de estimación de modelos de secuencias, de acuerdo con los enfoques descritos a lo largo del artículo, incluyendo un enfoque nuevo que hemos denominado estimación *pseudo_ML*.
- Estimación de modelos de frase a partir de corpus de tamaño arbitrario.
- Obtención de alineamientos a nivel de secuencia en base a un modelos de secuencias para un par de frases y su correspondiente matriz de alineamiento.

Creemos que la herramienta que se ha presentado (junto con otras herramientas de libre uso relacionadas con traducción automática estadística) pueden constituir un valioso recurso para la comunidad de traducción automática, pudiendo usarse para construir sistemas de traducción propios con bajos costes de desarrollo. Thot ha sido implementado siguiendo principios estándar de diseño como por ejemplo usabilidad y versatilidad en formatos. Estas características lo hacen interesante no sólo para expertos en el campo de la traducción automática sino para un público más general con conocimientos limitados sobre los fundamentos matemáticos de dicha disciplina.

Como trabajos futuros se plantea paralelizar diversas fases del entrenamiento y mejorar la documentación de la herramienta .

6. BIBLIOGRAFÍA

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, y R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[2] Hermann Ney, Sonja Nießen, Franz J. Och, Hassan Sawaf, Christoph Tillmann, y Stephan Vogel, "Algorithms for statistical translation of spoken language," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 1, pp. 24–36, Jan. 2000.

[3] Kenji Yamada y Kevin Knight, "A syntax-based statistical translation model," in *Proc. of the 39th Annual Meeting of ACL*, Toulouse, France, July 2001, pp. 523–530.

[4] Franz Joseph Och, *Statistical Machine Translation: From Single-Word Models to Alignment Templates*, Ph.D. thesis, Computer Science Department, RWTH Aachen, Germany, October 2002.

[5] J. Tomás y F. Casacuberta, "Monotone statistical translation using word groups," in *Procs. of the Machine Translation Summit VIII*, Santiago de Compostela, Spain, 2001, pp. 357–361.

[6] Daniel Marcu y William Wong, "A phrase-based, joint probability model for statistical machine translation," in *Proc. of the EMNLP*, Philadelphia, USA, July 2002, pp. 1408–1414.

[7] R. Zens, F.J. Och, y H.Ñey, "Phrase-based statistical machine translation," in *Advances in artificial intelligence. 25. Annual German Conference on AI*, vol. 2479 of *Lecture Notes in Computer Science*, pp. 18–32. Springer Verlag, September 2002.

[8] P. Koehn, F. J. Och, y D. Marcu, "Statistical phrase-based translation," in *Proceedings of the HLT/NAACL*, Edmonton, Canada, May 2003.

[9] Phillip Koehn, "Pharaoh: a beam search decoder for phrase-based statistical machine translation models. User manual and description," Technical report, USC Information Science Institute, Dec. 2003.

[10] F. Nevado, F. Casacuberta, y J. Landa, "Translation memories enrichment by statistical bilingual segmentation," in *Proc. of the Fourth Int. Conf. on LREC*, Lisbon, 2004.

[11] Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, y Kenji Yamada, "Fast decoding and optimal decoding for machine translation," in *Proc. of the 39th Annual Meeting of ACL*, Toulouse, France, July 2001, pp. 228–235.