# Técnicas mejoradas para la traducción basada en frases*

**Marta Ruiz Costa-jussà  José A. R. Fonollosa**
Centro de Investigación TALP
Universidad Politècnica de Catalunya (UPC)
Campus Nord UPC. 08034-Barcelona
{mruiz,adrian}@gps.tsc.upc.edu

**Resumen:** Actualmente, los métodos estadísticos de traducción automática suelen basarse en grupos de palabras, comúnmente llamados frases. En este artículo, estudiamos diferentes mejoras del sistema estándard de traducción basado en frases. Por un lado, describimos una modificación del método de extracción de frases. Por otro lado, proponemos características adicionales que conducen a una clara mejora en la calidad de la traducción. Finalmente, presentamos resultados con las tareas del EuroParl en ambas direcciones de traducción.
**Palabras clave:** traducción estadística, frases, modelado de máxima entropía.

**Abstract:** Nowadays, most of the statistical translation systems are based in phrases (i.e. groups of words). Here, we study different improvements to the standard phrase-based translation system. We describe a modified method for the phrase extraction which deals with larger phrases while keeping a reasonable number of phrases. We also propose additional features which lead to a clear improvement in the performance of the translation. We present results with the EuroParl task in both directions of translation between English and Spanish.
**Keywords:** stochastic machine translation, phrases, maximum entropy modelling.

## 1. *Introducción*

La traducción estadística se basa en que cada oración $e$ en un lenguaje destino es una posible traducción de una oración $f$ de un lenguaje fuente. La principal diferencia entre dos posibles traducciones de una oración dada es la probabilidad asignada a cada una, que se tiene que aprender de un texto bilingüe. Por lo tanto la traducción de una oración fuente $f$ se puede formular como la búsqueda de la oración destino $e$ que maximiza la probabilidad de traducción $P(e|f)$,

$$\tilde{e} = \operatorname*{argmax}_{e} P(e|f) \tag{1}$$

Aplicando la regla de Bayes y teniendo en cuenta que $P(f)$ no depende de $e$,

$$\tilde{e} = \operatorname*{argmax}_{e} P(f|e)P(e) \tag{2}$$

A esta aproximación se la conoce como la aproximación del modelo de canal ruidoso (Brown et al., 1990), donde: $P(e)$ representa la probabilidad de obtener la cadena de salida, y $P(f|e)$ es la probabilidad de obtener $f$ habiendo observado $e$. En la práctica se obtiene una estimación de $P(e)$ mediante un modelo de lenguaje y $P(f|e)$ mediante un modelo de traducción (García-Varea, 2003).

## 2. *Sistema básico de traducción*

Presentamos un sistema básico de traducción que se utilizará de referencia y se fundamenta en las siguientes estructuras:

**El modelo de traducción.** Se basa en frases bilingües. Cada frase se constituye de dos frases monolingües una de las cuales se supone la traducción de la otra. Una frase monolingüe es una sequencia de palabras. Por lo tanto, la principal idea de una traducción basada en frases radica en segmentar la oración fuente en frases, entonces traducir cada frase y finalmente hacer la composición de la oración destino a partir de estas frases traducidas (Zens, Och, y Ney, 2004). Antes, el sistema tiene que aprender de alguna manera un diccionario de frases.

Empezamos por alinear el corpus de entrenamiento utilizando GIZA++ (Och, 2003). Se entrena en las dos direcciones de tra-

* Este trabajo ha sido promovido parcialmente por el Gobierno Español TIC-2002-04447-C02 (Proyecto Aliado) y por la Unión Europea FP6-506738 (TC-STAR project). Los autores también quieren agradecer a José B. Mariño, Adrià de Gispert, Josep M. Crego, Patrik Lambert y Rafael E. Banchs (miembros del TALP) su contribución a este trabajo.

ducción. Tomamos la unión de ambos alineamientos para obtener una matriz simetrizada de alineamientos. Esta matriz es el punto de partida para la extracción de frases (Brown et al., 1993).

Básicamente, el criterio de extracción (Och y Ney, 2004) de frases se basa en:

- Las palabras son consecutivas en ambas frases monolingües.
- Ninguna palabra en cualesquiera de las frases monolingües está alineada con una palabra fuera del conjunto de la frase bilingüe.
- Como mínimo una palabra de la frase fuente está alineada con una palabra de la frase destino.

**El modelo de lenguaje destino.** Este modelo se combina con la probabilidad de traducción tal y como se muestra en la ecuación (2). Pretende dar coherencia al texto destino que se obtiene con la concatenación de frases.

**El decodificador.** Finalmente, el decodificador combina la información que se obtiene de los modelos de traducción y lenguaje. Realiza el proceso de búsqueda de maximización de la ecuación (2).

## *3. Extracción de frases*

**Motivación.** Definiremos la longitud de una frase monolingüe como el número de palabras que contiene. La longitud de una frase bilingüe será el máximo de las longitudes de cada frase monolingüe.

Al trabajar con corpus de tamaño considerable (ver sección de estadísticas del corpus), no resulta práctico hacer una tabla que contenga todas las *frases* de 4 palabras o más. De hecho, en (Koehn, Och, y Marcu, 2003) se limita la frase a 3 palabras, solamente para mantener la tabla de frases a un tamaño razonable. De lo contrario, se pasa a un coste computacional excesivo para la poca mejora obtenida.

**Variación de la longitud.** En nuestro sistema, consideramos dos límites de longitud. Primero, extraemos todas las frases de longitud 3 o inferior. Entonces, añadimos también las frases hasta longitud 5 si no se pueden generar con frases más pequeñas.

Básicamente, seleccionamos frases adicionales con palabras fuente que de otra manera se habrían perdido debido a los alineamientos cruzados o demasiado largos. Por ejemplo, dada la siguiente oración.

*Cuando el Parlamento Europeo , que tan frecuentemente insiste en los derechos de los trabajadores y en la debida protección social , (...)*

*NULL ( ) When ( 1 ) the ( 2 ) European ( 4 ) Parliament ( 3 4 ) , ( 5 ) that ( 6 ) so ( 7 ) frequently ( 8 ) insists ( 9 ) on ( 10 ) workers ( 11 15 ) ' ( 14 ) rights ( 12 ) and ( 16 ) proper ( 19 ) social ( 21 ) protection ( 20 ) , ( 22 ) (...)*

donde el número entre paréntesis indica la palabra o palabras alineadas. Nuestro algoritmo extrae la frase siguiente:

*los derechos de los trabajadores # workers ' rights*

En este caso, la frase tiene longitud 5, con lo cual, se habría perdido si hubiéramos utilizado longitud máxima 3. Y en caso, que utilizáramos longitud máxima 5, el número total de frases extraídas habría crecido hasta ser difícil de manejar.

## *4. Asignación de probabilidades*

Podemos generalizar la ecuación (2) del modelo de canal ruidoso, mediante el denominado modelo de entropía máxima.

$$\tilde{e} = \underset{e}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e,f) \right\} \quad (3)$$

$$h_m(e,f) = \log P(e,f;m) \quad (4)$$

donde $M$ es el número de modelos y $\lambda_m$ el peso que se asigna a cada modelo. Este modelo se deriva de la aproximación que encontramos en (Berger, Della Pietra, y Della Pietra, 1996) (Och y Ney, 2004). En este caso, el modelo de traducción y el modelo de lenguaje del destino son una información más entre varias que pueden gobernar la traducción. Las informaciones en la ecuación (3) las denominamos por $h_m(e,f)$. Los logaritmos de las *probabilidades* asociadas a las diversas informaciones (o características) se combinan linealmente (y con diferentes pesos $\lambda_m$) para definir una función cuya maximización establece la traducción. Este planteamiento tiene la ventaja que se pueden integrar fácilmente diferentes informaciones en el sistema completo. Por lo tanto, podemos plantear diferentes maneras de calcular las probabilidades de las frases.

### 4.1. La probabilidad condicional $P(f|e)$

Dadas unas frases determinadas, estimamos su probabilidad por frecuencia relativa de aparición. Así pues, tenemos:

$$P(f|e) = \frac{N(f,e)}{N(e)} \tag{5}$$

donde $N(f,e)$ significa el número de veces que aparece la frase $f$ traducida por la frase $e$. Dada una oración, si una frase $e$ tiene $N > 1$ posibles traducciones, entonces cada una de ellas contribuye al conteo ($N(f,e)$ y $N(e)$) con *1/N* (Zens, Och, y Ney, 2004).

Hay que tener en cuenta que, como no hay suavizado, esta frecuencia tiende a estar sobreestimada. Esto es especialmente perjudicial en una frases donde la parte fuente aparece muchas veces pero la parte destino aparece en raras ocasiones.

### 4.2. La probabilidad a posteriori $P(e|f)$

Como información adicional se ha considerada la probabilidad a posteriori,

$$P(f|e) = \frac{N'(f,e)}{N(f)} \tag{6}$$

donde $N'(f,e)$ significa el número de veces que aparece la frase $f$ traducida por la frase $e$. En este caso, dada una oración, si una frase $f$ tiene $N > 1$ posibles traducciones, entonces cada una de ellas contribuye al conteo ($N'(f,e)$ y $N(f)$) con *1/N*.

La probabilidad condicional y la probabilidad a posteriori son informaciones del modelo de la ecuación (3). Ambas probabilidades debidamente combinadas permitirían ayudar en los casos en los cuales las frases monolingües componentes de una frase tengan una frecuencia desequilibrada. Es decir, si la frase fuente tiene una frecuencia muy alta de aparición, y si la frase destino tiene una frecuencia de aparición baja, entonces la $P(f|e)$ tiende a estar sobrestimada. Sin embargo, la $P(e|f)$ permitiría compensar esta estimación errónea. Por ejemplo, la frase bilingüe que traduce *you* por *la que no* tendría que tener una probabilidad muy baja, porque es una mala traducción. Sin embargo, la probabilidad condicional estimada por frecuencia relativa proporciona,

$$P(f|e) = P(\text{you}|\text{la que no}) = 0{,}23$$

es decir, una probabilidad sobrestimada. Además, el modelo de lenguaje no penaliza *la que no* porque aunque como frase no aparece mucho (porque su alineamiento suele ser complejo), sí que aparece frecuentemente como trigrama. En cambio, si tomamos la probabilidad a posteriori, entonces tenemos una probabilidad mejor estimada.

$$P(e|f) = P(\text{la que no}|\text{you}) = 2 * 10^{-5}$$

### 4.3. IBM modelo 1

Tal como hemos mencionado en el apartado 4.1, utilizamos la frecuencia relativa para estimar las probabilidades de las frases de traducción. Sin embargo, la mayoría de las frases largas se ven sólo una vez en el corpus de entrenamiento. Y, la frecuencia relativa tiende a sobrestimar su probabilidad. Para solventar este problema, hemos propuesto el uso de la probabilidad a posteriori (en el párrafo anterior). Pero además, ya en la bibliografía, se había propuesto usar para suavizar las probabilidades de traducción el modelo 1 de IBM (Brown et al., 1990) (Och et al., 2004).

Mediante este modelo, la probabilidad de una frase se calcula de la siguiente manera.

$$P(f|e;\text{IBM}) = \frac{1}{(J+1)^I} \prod_{i=1}^{I} \sum_{j=0}^{J} p(f_j|e_i) \tag{7}$$

Las probabilidades $p(f_j|e_i)$ del modelo 1 de IBM se obtienen durante el alineamiento con el programa GIZA++. A los pares de palabras que no aparecen en la tabla generada se les asigna una probabilidad de $10^{-40}$.

Utilizamos también la $P(f_i|e_i)$, que nos ofrece el IBM$^{-1}$. Es decir, calculamos,

$$P(e|f;\text{IBM}^{-1}) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(e_i|f_j) \tag{8}$$

| | *Castellano* |
|---|---|
| Frases | 666.034 |
| Palabras | 43.390.713 |
| Vocabulario | 425.668 |

Cuadro 1: *Estadísticas del Parlamento Español*

### 4.4. Penalización por número de palabras y penalización por número de frases

Se trata de utilizar dos informaciones simples muy utilizadas en la bibliografía (Zens, Och, y Ney, 2004) (Koehn, 2003). Una penalización negativa de palabras beneficia salidas largas, es decir, compensa la tendencia a la generación de traducciones con el menor número de palabras.

La penalización de frases es un coste constante que se añade a cada una. Por lo tanto, valores positivos para la penalización de frases favorecen salidas con menos frases, mientras que valores negativos favorecen salidas con más frases.

## 5. *Modelo de lenguaje*

Un buen modelado del lenguaje destino resulta importante para la traducción basada en frases. El modelo de lenguaje $P(e)$ debe ser capaz de dar una idea de lo correcta que es una frase $e$ generada en el lenguaje destino.

Habitualmente, se utilizan los modelos $n$-gramas para generar el modelo de lenguaje mediante diversos métodos de suavizado. En nuestro caso se estimaron trigramas con suavizado de Kneser-Ney (Chen y Goodman, 1996) e interpolación mediante el software SRILM (Stolcke, 2002).

## 6. *Estadísticas del corpus*

Utilizamos la tarea del EuroParl en inglés y en español. Este corpus proviene del Parlamento Europeo (Koehn, 2002). Las estadísticas del corpus de entrenamiento que se ha utilizado se muestran en el cuadro 2. 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 la RWTH de Aachen en el ámbito del proyecto TC-STAR.[2]

| | *Castellano* | *Inglés* |
|---|---|---|
| Entrenamiento Frases | 1.223 M | 1.223 M |
| Palabras | 34,8 M | 33,4 M |
| Vocabulario | 169 k | 105 k |
| Desarrollo Frases | 1.008 | 1.008 |
| Palabras | 25,7 k | 26 k |
| Vocabulario | 3.937 | 3.208 |
| Test Frases | 840 | 1.094 |
| Palabras | 22,7 k | 26,8 k |
| Vocabulario | 4 k | 3,9 k |

Cuadro 2: *Estadísticas de los materiales de entrenamiento, desarrollo y test (k indica miles y M indica millones)*

Se han definido dos conjuntos de test para cada idioma: un conjunto de desarrollo y un conjunto de test. Para el corpus de desarrollo tenemos 3 traducciones de referencia, lo cual permite utilizar corpus más pequeños sin penalizar el margen de confianza. Y para el corpus propiamente de test tenemos 2 traducciones de referencia. El material de test consiste en la transcripción de las sesiones del 15 al 18 de noviembre de 2004.

Para mejorar la calidad del modelo de lenguaje del castellano (que en traducción sólo será útil para la dirección en→es), introducimos como material adicional el corpus del Parlamento Español.[3] Se han recogido las estadísticas en el Cuadro 1. Este corpus tiene proximidad en la forma y en la temática al EuroParl y nos permite reducir la perplejidad en un 10 %.

## 7. *Experimentos*

Para decodificar utilizamos el decodificador MARIE (Crego, Mariño, y de Gispert, 2005).

Las medidas de evaluación que utilizamos son las siguientes.

- *Word Error Rate* o tasa de error por palabra (WER): el WER se calcula como el mínimo número de sustituciones, inserciones y borrados que se tienen que realizar para convertir la oración generada en la oración de referencia.
- BLEU: esta medida se encarga de calcular la precisión de los $n$-gramas (con

[1] http://www.europarl.eu.int

[2] http://www.tc-star.org

[3] http://www.tc-star.org

| Dirección | $\lambda_{\text{LM}}$ | $\lambda_{p(f\|e)}$ | $\lambda_{p(e\|f)}$ | $\lambda_{\text{IBM}}$ | $\lambda_{\text{IBM}^{-1}}$ | *WER* | *BLEU* |
|---|---|---|---|---|---|---|---|
| en→es | 1 | 1 | 0 | 0 | 0 | 46,60 | 39,42 |
| en→es | 1 | 1 | 1 | 0 | 0 | 43,61 | 40,62 |
| en→es | 1 | 1 | 0 | 0,3 | 0,2 | 43,42 | 42,28 |
| es→en | 1 | 1 | 0 | 0 | 0 | 41,09 | 44,16 |
| es→en | 1 | 1 | 1 | 0 | 0 | 39,18 | 46,30 |
| es→en | 1 | 1 | 0 | 0,3 | 0,2 | 38,02 | 49,19 |

Cuadro 3: *Tarea EuroParl. Resultados para: el sistema básico; el sistema básico más la probabilidad a a posteriori; y el sistema básico más* IBM *y* $\text{IBM}^{-1}$. *Las medidas WER y BLEU se expresan en porcentaje.*

| Carac. Frases | $\lambda_{\text{LM}}$ | $\lambda_{p(f\|e)}$ | $\lambda_{p(e\|f)}$ | $\lambda_{\text{IBM}}$ | $\lambda_{\text{IBM}^{-1}}$ | $\lambda_{PP}$ | $\lambda_{WP}$ | *WER* | *BLEU* | *# frases* |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 46,60 | 39,42 | 67,7M |
| 3 | 0,7 | 0,48 | 0,52 | 0,62 | 0,08 | 1,27 | -0,08 | 41,64 | 44,35 | 67,7M |
| 3+5long | 0,7 | 0,62 | 0,33 | 0,67 | 0,18 | 1,49 | -0,01 | 41,72 | 44,42 | 68M |

Cuadro 4: *Tarea del EuroParl. Resultados optimizados en la dirección en→es. Las puntuaciones WER y BLEU se muestran en porcentajes. # indica número, M indica millones, 3+5long indica una longitud de 3 más las frases seleccionadas de longitud 5 según el criterio en la sección 3.*

$n = 1..,4$) con respecto a la traducción de referencia. Un mayor BLEU suele indicar una mejor traducción (Papineni et al., 2002)

**Optimización.** Los pesos asignados a cada una de las informaciones utilizadas de acuerdo con la ecuación (3) pueden optimizarse de diversas maneras para mejorar la traducción. En nuestro caso hemos utilizado el algoritmo de optimización denominado SIMPLEX (Nelder y Mead, 1965). Y optimizamos con el corpus de desarrollo y respecto a la medida BLEU. Las mejores combinaciones las presentamos en los cuadros 4 y 5, expuestas de manera que se ve la influencia de la variación de longitud del apartado 3. Como podíamos pensar, no todas las mejoras son acumulativas.

**Resultados.** El cuadro 3 muestra los resultados del sistema básico junto con los resultados de la incorporación de varias características. Vemos que la probabilidad a posteriori mejora en 2.12 de BLEU en la dirección es→en. En la misma dirección, el efecto del IBM combinado con $\text{IBM}^{-1}$ mejora en 3.02 de BLEU. En la otra dirección las mejoras son similares. Podemos observar que los pesos no están optimizados, por lo tanto, las mejoras puede que sean más importantes.

Los cuadros 4 y 5 muestran los resultados del sistema básico junto con los resultados de todas las características optimizadas. La influencia de la incorporación de frases más largas es pequeña (0.15 de BLEU, en el mejor caso) pero también implica poco incremento del coste computacional.

Veáse que los resultados son notablemente mejores en la dirección de es→en. Esto se debe a la simplicidad del inglés respecto al castellano. Por ejemplo, el inglés tiene menos vocabulario y la conjugación verbal es más homogénea con lo cual se generan menos errores.

## *8. Conclusiones*

En primer lugar, hemos presentado un nuevo método para extraer frases más largas manteniendo razonable la cantidad de frases.

Además, hemos propuesto características adicionales que conducen a una clara mejora en la calidad de la traducción: la $P(e|f)$; el modelo IBM y $\text{IBM}^{-1}$; y la penalización de frase y de palabra.

Estas características en combinación con las informaciones del modelo de canal ruidoso han dado lugar a una mejora muy significativa en la calidad de la traducción.

La evaluación se ha hecho con la tarea del EuroParl. Con las diferentes características añadidas hemos obtenido una mejora absoluta de 10 puntos de BLEU en la dirección del es→en.

| Carac. Frases | $\lambda_{\mathrm{LM}}$ | $\lambda_{p(f\|e)}$ | $\lambda_{p(e\|f)}$ | $\lambda_{\mathrm{IBM}}$ | $\lambda_{\mathrm{IBM}^{-1}}$ | $\lambda_{PP}$ | $\lambda_{WP}$ | *WER* | *BLEU* | *# frases* |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 41,09 | 44,16 | 67,7M |
| 3 | 0,7 | 0,49 | 0,63 | 0,66 | 0,20 | 2,20 | -0,40 | 35,07 | 54,08 | 67,7M |
| 3+5long | 0,7 | 0,62 | 0,36 | 0,57 | 0,20 | 2,42 | -0,36 | 35,10 | 54,23 | 68M |

Cuadro 5: *Tarea del EuroParl. Resultados optimizados en la dirección es→en. Las puntuaciones WER y BLEU se muestran en porcentajes. # indica número, M indica millones, 3+5long indica una longitud de 3 más las frases seleccionadas de longitud 5 según el criterio en la sección 3.*

### *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.

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

Brown, P., S. Della Pietra, V. Della Pietra, y R. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311.

Chen, S. y J. Goodman. 1996. An Empirical Study of Smoothing techniques for Language Modeling. En *Proceedings of 34th ACL*, páginas 310–318, San Francisco, July.

Crego, J.M., J. Mariño, y A. de Gispert. 2005. An Ngram-based statistical machine translation decoder. *Submitted to Interspeech 2005*, April.

García-Varea, I. 2003. *Traducción Automática estadística: Modelos de Traducción basados en Máxima Entropía y Algoritmos de Búsqueda* . UPV, Diciembre.

Koehn, P. 2002. Europarl: A Multilingual Corpus for Evaluation of Machine Translation. En *Draft, Unpublished*.

Koehn, P. 2003. A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. *Technical Manual of the Pharaoh decoder*.

Koehn, P., F. J. Och, y D. Marcu. 2003. Statistical phrase-based translation. En *Proceedings of the Human Language Technology Conference (HLT-NAACL)*, páginas 127–133, May.

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

Och, F. J. y H. Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational linguistics*, 30:417–449, December.

Och, F.J. 2003. Giza++ software, http://www-i6.informatik.rwth-aachen.de/ och/software/giza++.html/.

Och, F.J., D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, y D. Radev. 2004. A smorgasbord of features for statistical machine translation. *Proc. of the Human Language Technology Conference, HLT-NAACL'2004*, páginas 161–168, May.

Papineni, K.A., S. Roukos, R.T. Ward, y W-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *40th Annual Meeting of the Association for Computational Linguistics*, páginas 311–318, July.

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

Zens, R., F.J. Och, y H. Ney. 2004. Improvements in phrase-based statistical machine translation. *Proc. of the Human Language Technology Conference, HLT-NAACL'2004*, páginas 257–264, May.