TÉCNICAS ESTADÍSTICAS PARA EL FILTRADO DE UN CORPUS BILINGÜE EN TRADUCCIÓN AUTOMÁTICA

Enrique Montolar, Marta R. Costa-Jussà y José A. R. Fonollosa

Universitat Politècnica de Catalunya, UPC
Centro de Investigación TALP,
Campus Nord, 08034 Barcelona
enrique.montolar@gmail.com {mruiz,adrian}@gps.tsc.upc.edu

RESUMEN

Los sistemas de traducción automática estadística están basados en corpus bilingües. La calidad de estos es un factor determinante en la calidad de la traducción. En esta comunicación, presentamos dos filtrados estadísticos que permiten descartar las oraciones del corpus bilingüe que tienen menor probabilidad de ser paralelas. Como resultado, se obtiene una mejora superior a 1 punto BLEU y, al reducir el corpus de entrenamiento, disminuye el coste computacional.

1. INTRODUCCIÓN

La traducción automática estadística se basa en el hecho de que cada oración e en un lenguaje destino es una posible traducción de una oración f en 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 $P(e|f)$.

La aproximación estadística a la traducción automática es una aproximación basada en corpus. Concretamente, se requieren corpus paralelos a nivel de oración.

Un problema habitual de los corpus bilingües es la presencia de oraciones no paralelas que dan lugar a unidades de traducción erróneas. En esta comunicación se presentan dos métodos de filtrado del corpus con el objetivo de eliminar estas oraciones no paralelas.

La sección 2 presenta el sistema básico de traducción. La sección 3 presenta las técnica de filtrado estadístico basadas en el Modelo IBM1 y en el Position Error Rate (PER). La sección 4 muestra los experimentos y la sección 5 presenta las conclusiones.

Este trabajo ha sido subvencionado por el Gobierno Español mediante el proyecto coordinado AVIVAVOZ (TEC2006-13694-C03) y el Govern de la Generalitat de Catalunya mediante el proyecto TecnoParla.

2. SISTEMA DE TRADUCCIÓN

Como sistema de traducción estadística se ha utilizado un sistema basado en n-gramas [1].

El modelo de traducción puede entenderse como un modelo de lenguaje de unidades bilingües (llamadas tuplas). Dichas tuplas, definen una segmentación monótona de los pares de oraciones utilizadas en el entrenamiento del sistema (f_1^J, e_1^I), en K unidades $(t_1, ..., t_K)$.

En la extracción de las unidades bilingües, cada par de oraciones da lugar a una secuencia de tuplas que sólo depende de los alineamientos internos entre las palabras de la oración. La Figura 1 muestra un ejemplo de extracción de tuplas. El modelo de traducción se ha implementado utilizando un modelo de lenguaje (bilingüe) basado en n-gramas [1]:

$$p(e, f) = Pr(t_1^K) = \prod_{k=1}^{K} p(t_k \,|\, t_{k-2}, t_{k-1}) \qquad (1)$$

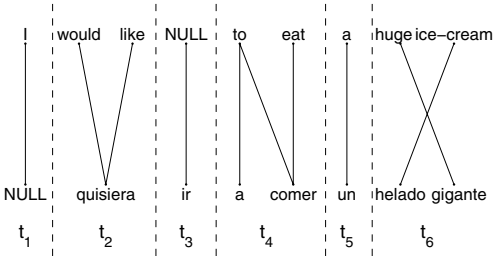

Figura 1. *Extracción de tuplas a partir de un par de oraciones alineadas palabra a palabra.*

En la traducción de una oración de entrada, el decodificador debe encontrar la secuencia de tuplas asociada a una segmentación de la oración de entrada que produzca probabilidad máxima. En general, tal probabilidad máxima se calcula como combinación lineal de modelos utilizados en el sistema de traducción.

El sistema de búsqueda utilizado se ha desarrollado en la UPC: MARIE [1].

[1] *http://gps-tsc.upc.es/veu/soft/soft/marie/*

3. FILTRADO ESTADÍSTICO DE UN CORPUS BILINGÜE

Dado un corpus bilingüe de entrenamiento, nuestro objetivo se centra en analizar dicho corpus y filtrarlo para intentar obtener un nuevo corpus que nos permita mejorar nuestro sistema de traducción. Concretamente, queremos reducir las tuplas erróneas y mejorar el vocabulario bilingüe. Para ello debemos detectar frases dentro del corpus que no sean paralelas. A continuación presentamos dos soluciones basadas en criterios diferentes para detectar y eliminar oraciones que no sean paralelas.

3.1. Planteamiento mediante modelo IBM1

El método propuesto para detectar y posteriormente descartar frases no paralelas, es decir para filtrar el corpus de entrenamiento, estará basado en el modelo de alineamiento IBM1. Recordemos que dicho modelo surge de la necesidad de establecer un alineamiento entre las palabras de un par de oraciones, dados dos textos paralelos a nivel de oración, que son traducciones mutuas del par de lenguas que nos ocupan en cada caso. Los modelos IBM calculan la probabilidad de que dos palabras estén alineadas entre ellas, es decir la probabilidad de que una palabra de la oración origen se corresponda con una palabra de la oración destino. Son 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.

Así pues podemos establecer la probabilidad de traducción de un par de frases en función de la probabilidad de traducción de las palabras que las componen. Analizando dicha probabilidad para cada par de frases de nuestro corpus, podemos buscar un umbral de probabilidad que nos indicará si las frase son paralelas. Es decir podremos determinar que la probabilidad de que una frase de un texto se corresponda a la frase alineada con esta del otro texto es tan baja, que las dos frases no se corresponden es decir no son paralelas.

La probabilidad de traducción asignada según el modelo IBM1 a cada oración se calcula mediante la expresión:

$$h_{LEX}(e,f) = log\frac{1}{(I+J)^J}\prod_{j=1}^{J}\sum_{i=0}^{I} p_{IBM_1(e_j^n|f_i^n)} \quad (2)$$

Donde e_j^n y f_i^n son la j^{esima} y i^{esima} palabras en la oraciones fuente y destino, con I número de palabras de la fuente y J número de palabras del destino. Así pues $p_{IBM_1(e_j^n|f_i^n)}$ serán las probabilidades de traducción en la dirección fuente-destino $p(e_k/f_k)$ asignada por el modelo IBM1.

Teniendo en cuenta que el modelo IBM-1 es asimétrico, es decir es diferente según el sentido de la traducción, debemos calcular también la probabilidad de traducción en el sentido inverso, este se calcula mediante la expresión:

$$h_{LEX}inv(f,e) = log\frac{1}{(J+I)^I}\prod_{i=1}^{I}\sum_{j=0}^{J} p_{IBM_1(f_i^n|e_j^n)} \quad (3)$$

Donde tenemos $p_{IBM_1(f_j^n|e_i^n)}$ serán las probabilidades de traducción en la dirección $p(f_k/e_k)$ asignada por el modelo IBM1.

Así pues, utilizando las probabilidades léxicas obtenidas del Modelo IBM1, calcularemos la probabilidad de que dos oraciones paralelas en el corpus bilingüe sean traducciones entre ellas.

Al calcular las probabilidades de frases paralelas, debemos tener en cuenta varios factores que influyen en el cálculo de la probabilidad, como la repetición de palabras con una gran probabilidad de coincidencia.

El hecho de que en un par de frases aparezcan las palabras más comunes de cada idioma, nos incrementaría mucho la probabilidad de que fueran paralelas, aunque realmente no tiene porque ser así. Dos frases no coincidentes en absoluto a nivel de significado, pueden estar compuestas de palabras con una gran probabilidad de coincidencia, de no tener en cuenta este factor, una frase de este tipo sería aceptada como válida en el sistema de traducción a pesar de ser errónea.

Para solventar este problema introduciremos el concepto de **stopwords**. Entendemos por **stopwords** aquellas palabras o signos de puntuación que son muy comunes en el texto, como ya hemos dicho su presencia influye aumentando considerablemente la probabilidad de que dos frases sean paralelas. Así pues elaboramos listas de las palabras más comunes para cada idioma y no las consideraremos cuando calculemos la probabilidad IBM1 entre frases.

La idea es poder realizar una comparativa de dos criterios de selección de frases, para así poder analizar que frases se han descartado según cada método y ver con qué método obtenemos mejores resultados.

3.2. Planteamiento mediante el PER

Análogamente al método basado en el modelo IBM1, utilizaremos otra herramienta para la selección de frases basada en el análisis del PER.

El sistema basado en PER consiste simplemente en utilizar el sistema de traducción inicial, con él traduciremos la parte fuente del corpus bilingüe y evaluaremos dicha traducción con la parte destino del corpus bilingüe.

Dada una herramienta que calcula el PER, eliminaremos los pares de líneas cuyo PER sea peor que el del resto. Determinamos el umbral PER entre frases aceptadas y eliminadas de tal manera que eliminemos el mismo número de frases que eliminábamos con el criterio de selección del IBM1.

A priori podemos suponer que el coste computacional de este sistema va a ser considerable, puesto que requiere de la traducción de todo el texto de entrenamiento, cuyo tamaño siempre es extenso, y la posterior evaluación de la traducción línea a línea que también resulta un proceso lento.

4. PROCESO EXPERIMENTAL

Utilizando el criterio IBM1 y el criterio PER haremos un análisis de las probabilidades de frases paralelas y una posterior eliminación de las frases que consideremos que son de peor calidad. Evaluaremos automáticamente mediante las medidas WER, PER, BLEU y NIST.

4.1. Datos

Realizamos el proceso de análisis para el corpus del proyecto TC-STAR [2] de castellano a inglés que utiliza los textos del *European Parliament Plenary Sessions* (EPPS). La Tabla 1 presenta las estadísticas del corpus de entrenamiento. Para evaluar se utilizó el test oficial correspondiente a la segunda evaluación del TC-STAR.

CORPUS	orac.	pal.	vocab.	Lmax	Lmean
train.eng	1,3M	37,0M	109,8k	100	27.3
train.spa		39,5M	147,6k	110	29.1

Tabla 1. *Estadísticas del corpus.*

4.2. Detalles del sistema de traducción

Alineamiento. Mediante la aplicación GIZA++ [2]. se realiza el alineamiento de los textos bilingües paralelos del material de entrenamiento, ejecutándose 4 iteraciones del modelo IBM1, 5 iteraciones del modelo HMM 3 iteraciones del modelo IBM4 y ninguna del modelo IBM3. Se obtiene el alineamiento en las dos direcciones de traducción: tomando alternativamente uno y otro idioma como lenguas fuente. A partir de estos dos alineamientos básicos, se obtienen los alineamientos unión e intersección de los mismos, definidos respectivamente,por los conjuntos unión e intersección de los enlaces establecidos en los alineamientos básicos.

Se eliminan los pares bilingües en el que una de las oraciones contenga más de 50 palabras o en el que el cociente entre el número de palabras de una y otra oración exceda 2.4 (fertilidad superior a 2.4)

Selección de tuplas. Una vez obtenido el alineamiento unión se procede a la segmentación en tuplas del material de entrenamiento. A efectos de simplificar el sistema de traducción, el vocabulario de tuplas se limita a aquellas que tengan una longitud máxima de 15 palabras tanto en el lenguaje fuente como en el lenguaje destino.

[2] *www.tcstar.org*

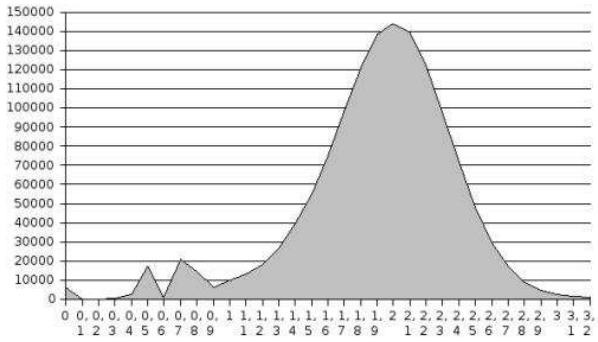

Figura 2. *Modelo IBM1 con Stop Words.*

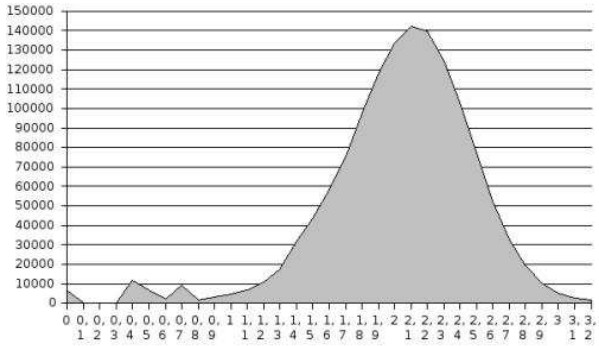

Figura 3. *Modelo IBM1 inverso con Stop Words.*

Estimación del modelo. Para estimar el modelo se utiliza la herramienta SRILM [3]. En este proceso se limita el vocabulario del modelo del lenguaje bilingüe a las tuplas seleccionadas, al que se añade una traducción (tupla) para todas aquellas palabras que no aparezcan 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“) son generadas a partir del alineamiento intersección. Como técnica de suavizado se utiliza el método de Kneser-Ney e interpolación lineal (Kneser and Ney, 1995). El modelo generado fue un trigrama (N=3) de tuplas.

4.3. Aplicación del filtrado estadístico

4.3.1. Distribución de probabilidad

A continuación presentamos los resultados de calcular el modelo IBM1 sobre cada una de las oraciones paralelas en el corpus. Las Figuras 2 y 3 muestran el número de oraciones (eje Y) con una determinada probabilidad de ser paralelas (la probabilidad está expresada en logaritmo negativo).

Hemos eliminado la aportación probabilística de las stopwords, en concreto seleccionamos las 30 palabras más comunes en cada idioma.

Corpus	Umbral dir	Umbral inv	Eliminad.
C1	2.51	2.62	12 %
C2	2.7	2.8	5 %

Tabla 2. *Umbrales propuestos y frases eliminadas.*

4.3.2. Umbrales propuestos para la selección de frases.

Dadas las Figuras anteriores, se trata de experimentalmente seleccionar un umbral de probabilidad. En nuestro caso, la probabilidad está expresada en logaritmo negativo, con lo cual, eliminaremos todas las frases que estén por encima del umbral escogido. El hecho de filtrar el corpus original nos da lugar a un nuevo corpus. La Tabla 2 muestra un par de corpus generados (C1 y C2) a partir de ciertos umbrales. Asimismo, se muestra el % de frases eliminadas en cada corpus.

Realizamos la evaluación para ver que resultados obtenemos utilizando los nuevos modelos provenientes de los corpus tratados (ver Tabla 3).

Tipo	Corpus Inicial	C1	C2
BLEU score	43.33	44.52	44.20
NIST score	9.6	9.76	9.72
PER score	31.15	30.98	31.15
WER score	41.41	40.65	40.83

Tabla 3. *Resultados Corpus entrenamiento inicial,corpus C1 y corpus C2*

4.3.3. Criterio selección OR.

Este criterio consiste en admitir una frase como válida si alguna de las probabilidades IBM1, ya sea la directa o la inversa, supera un determinado umbral.

Establecemos un umbral que nos permita eliminar el mismo número de oraciones que en el corpus C1. Ello lo conseguimos estableciendo un umbral tanto directo como inverso de 2.4. Construimos el modelo con el nuevo corpus filtrado (C3) y obtenemos los resultados que se muestran en la Tabla 4.

Tipo	C3	C4
BLEU score	44.50	44.37
NIST score	9.75	9.73
PER score	31.01	31.86
WER score	40.68	40.71

Tabla 4. *Resultados Corpus entrenamiento C3 y C4*

4.3.4. Resultados obtenidos con el sistema PER.

Vamos a utilizar la eliminación de frases mediante el PER con el fin de comprobar los resultados obtenidos mediante el modelo IBM1. Eliminamos el 12 % de las frases mediante este sistema y obtenemos un corpus al que denominaremos C4 que procedemos a evaluar como hemos hecho en cada caso (ver Tabla 4).

Podemos observar que los resultados, aunque mejoran los obtenidos en el sistema inicial, son inferiores a los obtenidos utilizando en el modelo creado a partir del corpus C1, que es el que proporciona mejores resultados.

4.3.5. Análisis de los resultados

En las frases descartadas, vemos la influencia de la longitud de las frases en la evaluación del modelo IBM1, en este caso la media de las palabras por frase del corpus original es de 20, mientras que en las frases descartadas es de 40.

Como ejemplos de frases descartadas podemos citar:

EN línea 70:*We see that the French Government has sent a mediator.*

ES línea 71:*Vemos que el Gobierno francés ha enviado a un mediador.*

donde hay un desplazamiento de línea de un texto a otro. En otros casos encontramos traducciones que en un contexto pueden tener cierto sentido, pero a nivel de corpus de entrenamiento no aportan una información adecuada, por ejemplo encontramos:

- *This is where the main obstacles lie.* se traduce como - *Las consideraciones van especialmente en esta dirección.*

- *To conclude , let me ask what lessons we should be learning.* se traduce como - *Permítanme que finalice mi intervención con una serie de propuestas.*

5. CONCLUSIONES

Se han presentado dos técnicas de filtrado estadístico de corpus bilingües. La técnica basada en el Modelo IBM1 obtiene resultados ligeramente superiores a la técnica basada en el PER. Incorporar la técnica de filtrado estadístico permite reducir el ruido del corpus de entrenamiento y como consecuencia, se reduce el coste computacional y se mejora la calidad del sistema de traducción. En los experimentos que hemos presentado se mejora en más de 1 punto BLEU.

6. BIBLIOGRAFÍA

[1] J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. R. Fonollosa, y M. R. Costa-Jussà, "N-gram-based machine translation," *Computational Linguistics, Association for Computacional Linguistics.*, vol. 32, no. 4, pp. 527–549, 2006.

[2] F.J. Och y H. Ney, "A systematic comparison of various statistical alignment models," vol. 29, no. 1, pp. 19–51, March 2003.

[3] A. Stolcke, "Srilm - an extensible language modeling toolkit," September 2002.