Técnicas para el filtrado y la categorización de un corpus bilingüe en traducción automática estadística

Enrique Montolar

17 de septiembre de 2008

Resumen

Este proyecto se desarrolla en el marco de la Traducción Automática Estadística (TAE). Tomamos como punto de partida el sistema de traducción estadística de la UPC.

Este sistema utiliza secuencias de palabras, denominadas tuplas, como unidades básicas del modelo de traducción, estas tuplas son unidades bilingües que se extraen de un corpus (recopilación de textos) paralelo a nivel de oración.

Nuestro trabajo se centrará en el análisis de los corpus de entrenamiento y posible modificación de los mismos, para conseguir aumentar la eficiencia de nuestro sistema de traducción y mejorar la calidad de las traducciones realizadas.

Para ello desarrollaremos dos técnicas, el filtrado estadístico del corpus y la categorización de palabras.

La primera técnica pretende establecer unos criterios que determinen la calidad de las oraciones bilingües que componen el corpus de entrenamiento, para ello utilizaremos diferentes herramientas de filtrado estadístico, y estableceremos cual es la más conveniente en cada caso.

La segunda técnica se basa en el hecho de que en cualquier idioma, existen cierto tipo de palabras o construcciones que poseen unas características particulares, tienen una estructura específica que nos permite tener una información adicional. Nuestro objetivo será la utilización de esta información para la traducción de este tipo de palabras, para ello será necesario detectar estas palabras y traducirlas independientemente. En concreto realizaremos la categorización tanto de números como de páginas web.

Hemos realizado experimentos con diferentes corpus: chino-inglés, castellano-ingles y catalán-castellano.

El filtrado estadístico ha permitido reducir el tamaño del modelo de traducción, de manera que reduce el coste computacional, y también aumentar ligeramente la calidad de la traducción.

La categorización ha permitido reducir las palabras desconocidas en el caso del chino-inglés, y en general, mejorarla traducción de números y páginas web.

Introducción

Motivación y contexto del proyecto

Las primeras palabras de esta introducción van dirigidas a justificar la necesidad de investigar en el ámbito de la traducción automática. Para ello partimos de los siguientes hechos: por un lado, de la internacionalización de las lenguas, por otro, de la globalización a que están siendo conducidas nuestras sociedades.Con ello se entenderá el papel que puede -y debe- desempeñar la traducción automática para superar barreras lingüísticas. Y al mismo tiempo, se admite que la actual sociedad de la información nos lleva hacia la existencia de una "aldea global", basada en el respeto a la diversidad lingüística.

Uno de los factores de la globalización es la comercialización de productos en mercados lo suficientemente amplios como para que los beneficios lleguen a compensar las inversiones realizadas en su desarrollo.

En un mundo en el que la información juega un importante papel en la sociedad, los procesos en los que intervienen mecanismos de traducción se han hecho imprescindibles en la cadena comunicativa que demandan los usuarios. Dado que la informática nos ofrece hoy día herramientas que ayudan al usuario en todos los ámbitos de la vida, automatizando gran parte de los procesos que realiza, la pregunta que surge de forma natural es: ¿podemos crear sistemas que puedan automatizar el proceso de traducción, o al menos crear herramientas que ayuden al traductor en su trabajo? Quizás el punto en común sea que todas estas herramientas pretenden ayudar al traductor profesional para que su trabajo sea más rentable, fiable, consistente y menos repetitivo.

Por otro lado, la disponibilidad de textos bilingües en formato digital ha hecho posible el diseño de métodos automáticos de extracción de información lingüística.

Los programas de traducción automática, están concebidos para que ayuden a realizar traducciones técnicas, en dominios del lenguaje bien definidos y normalmente carecen de figuras retóricas y literarias.

La traducción automática se está imponiendo cada vez más por varias razones:

- Demanda de traducciones de ámbito global.

- Los textos a traducir son cada vez más técnicos.

- Es más barata que la traducción humana, a pesar de que en muchos casos requiere trabajo de post-edición.

Uno de los sistemas de Traducción Automática más utilizado en el actualidad es el sistema estadístico.

Objetivos del proyecto.

Este proyecto se desarrolla en el marco de la Traducción Automática Estadística (TAE) , que utiliza un corpus paralelo a nivel de oración como entrenamiento.

Nosotros trabajaremos sobre estos corpus con el fin de alcanzar los siguientes objetivos:

- **Mejora del corpus de entrenamiento.**

  Por diferentes motivos un par de frases puede contribuir de forma negativa al sistema de traducción. Nuestro objetivo será establecer los criterios que determinen cuando una frase bilingüe es conveniente o no para nuestro sistema. Esto implica realizar un análisis de los textos bilingües, utilizando diferentes herramientas con el fin de eliminar las frases cuya aportación a nuestro sistema de traducción sea negativa o nula.

- **Categorización de números y páginas web**.

  Hasta este punto no se ha realizado categorización en el sistema de TAE de la UPC, de modo que tanto los números en general como las páginas web que aparecen en los textos no reciben un tratamiento especial. Nuestro objetivo será la detección y procesado de dichos elementos del texto.

Estructura del proyecto.

Siguiendo esta introducción, el proyecto se estructura de la siguiente manera:

- **Capítulo 1 : Estado del Arte.**
  Descripción de cual es el estado de desarrollo del sistema de traducción hasta el momento de comenzar nuestro proyecto.

- **Capítulo 2 : Filtrado estadístico.**
  Descripción de las técnicas empleadas para realizar el filtrado estadístico de los corpus de entrenamiento y experimentos realizados aplicando dichas técnicas a diferentes corpus de entrenamiento.

- **Capítulo 3 : Categorización.**
  Localización y posterior tratamiento de palabras o construcciones que poseen unas características particulares, como son números y páginas web.

- **Capítulo 4 : Conclusiones.**
  Resumen de los análisis de los resultados, obtenidos en la aplicación de las diferentes técnicas implementadas en este proyecto.

Capítulo 1

Estado del Arte

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 $P(e|f)$,

$$\tilde{e} = \underset{e}{\operatorname{argmax}}\, P(e|f) \tag{1.1}$$

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

$$\tilde{e} = \underset{e}{\operatorname{argmax}}\, P(f|e)P(e) \tag{1.2}$$

A esta aproximación se la conoce como la aproximación del modelo de canal ruidoso [Brown, P., Della Pietra,S., Della Pietra, V., and Mercer, R., 1993], 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.

Los primeros sistemas TAE seguían la aproximación del canal ruidoso, como ya hemos comentado, y trabajaban a nivel de palabras (las unidades bilingües se componían de palabras aisladas).

Recientemente, los sistemas de TAE tienden a utilizar secuencias de palabras, denominadas tuplas, como unidades básicas del modelo de traducción [Koehn, Och, and Marcu, 2003], con el objetivo de introducir el contexto en

dicho modelo.

Particularmente el sistema propio de la UPC usa los n-gramas bilingües. Las unidades bilingües (llamadas tuplas) se extraen a partir de cada par de oraciones. Una secuencia de tuplas sólo dependerá de los alineamientos internos entre las palabras de la oración.

Estos sistemas llevan a cabo la traducción mediante la maximización de una combinación loglineal (como alternativa al modelo de canal ruidoso) de los logaritmos de la probabilidad asignada a la traducción por el modelo de traducción y otras características [Och, F.J. and Ney,H., 2002].

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

$$U = \sum_i \lambda_i h_i(d, f) \tag{1.3}$$

Formada por la combinación lineal de distintas características $h_i(d, f)$ relativas a pares bilingües de oraciones traducidas 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.

1.1. El sistema de traducción

1.1.1. Alineamiento

El sistema tiene que aprender de alguna manera un diccionario de tuplas, que son segmentos bilingües en los que dividimos nuestro corpus de oraciones bilingües.

Necesitamos 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 entre sí de una

y otra lengua. Para ello utilizamos una herramienta de código abierto.

Empezamos por alinear el corpus de entrenamiento utilizando una herramienta de alineación. Mediante dicha herramienta se realiza el alineamiento de los textos bilingües paralelos del material de entrenamiento.

El alineamiento nos permite establecer las probabilidades, denominadas probabilidades IBM1, de traducción entre las palabras fuente y destino.

Figura 1.1: *Alineamiento y obtención de probabilidades.*

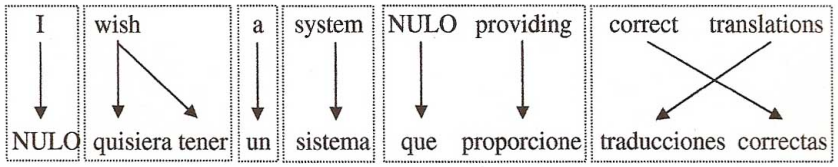

Figura 1.2: *Par de oraciones bilingües en el que, mediante flechas, se indican las palabras vinculadas en la traducción. Mediante recuadros se muestran los pares bilingües de segmentos (tuplas) en los que se segmenta monótonamente el par de oraciones.*

En la figura 1.2 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, J. M., Costa-jussà, M. R., Mariño, J., and Fonollosa, J. A., 2005] puede encontrarse una comparación entre algunos de ellos.

1.1.2. Extracción de tuplas.

Nuestro sistema utiliza un subconjunto de estos segmentos bilingües, que llamamos tuplas.

Básicamente, el criterio de extracción de tuplas [Och, F.J., 2003] 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.
- Cada tupla no puede ser descompuesta en segmentos bilingües más pequeños sin violar las condiciones anteriores.

En la figura 1.2 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 1.2, 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. 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 anterior o a la siguiente en función de la probabilidad IBM1.

En la figura 1.2 también se puede observar el concepto de tupla incrustada (embedded):

(correct translations, traducciones correctas)

En este caso no tenemos traducción para una de las palabras aislada, ya sea "correct.º "translations", porque forman una única tupla.
Una solución para la mayoría de los casos en los que hay un cruce de vínculos (links) directo, es desglosar la tupla.

A la hora de realizar la extracción de tuplas podemos ejecutarlo de manera normal (*regular*) o bien desglosado en unidades menores (*unfold*), para ello disponemos de un parámetro mediante el que indicaremos cual será el tipo de extracción. En la figura 1.3 mostramos un ejemplo de cual es la diferencia entre un tipo y otro.

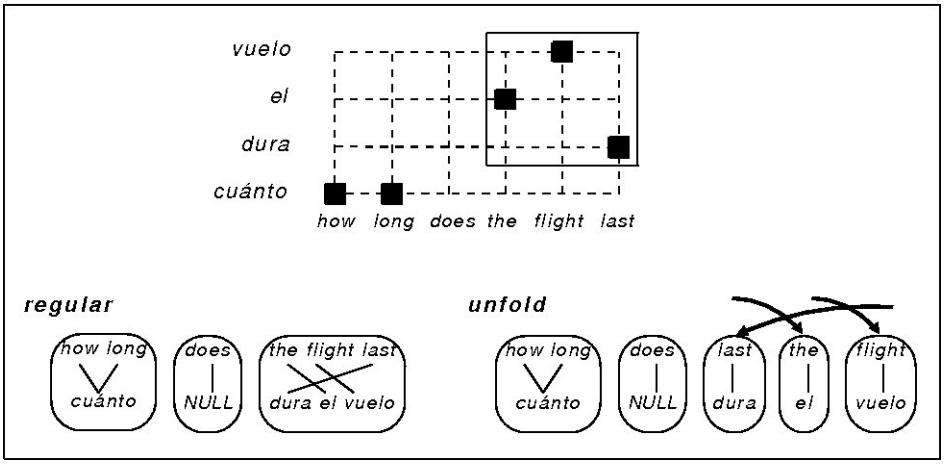

Figura 1.3: *Ejemplo de extracción de tuplas.*

1.1.3. El modelo de traducción

Se basa en sequencias de palabras bilingües o tuplas [Mariño, J.B., Banchs, R.E., Crego, J.M., de Gispert, A., Lambert, P., Fonollosa, J.A.R., and Costa-jussà, M.R., 2006]. Cada frase de entrenamiento se constituye de dos frases monolingües una de las cuales se supone la traducción de la otra. Cada frase la

podemos separar en diferentes segmentos bilingües, o sea en tuplas. Una frase monolingüe es una secuencia de palabras. Por lo tanto, la principal idea de una traducción basada en tuplas radica en segmentar la oración fuente en tuplas, entonces traducir cada tupla y finalmente hacer la composición de la oración destino a partir de estas tuplas traducidas.

Si se realiza el alineamiento 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 el modelamos según el modelo de N-grama. La extracción de tuplas y estimación de las probabilidades de las frase de traducción son el núcleo central del modelo de traducción.

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:

$$p(d, f) = P_r\{t^K\} = \prod_{k=1}^{K} p(t_k|t_{k-1}, ..., t_{k-N+1}) \tag{1.4}$$

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

1.1.4. Las funciones 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:

$$h_1(d, f) = \log \prod_{k=1}^{K} p(t_k|t_{k-1}, ..., t_{k-N+1}) \tag{1.5}$$

Actualmente, el sistema N-grama 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 la oración que constituye cada tupla $t_k = (d_k, f_k)$. Ambas probabilidades se consideran informaciones independientes.

$$h_2(d, f) = \log \prod_{k=1}^{K} p(d_k|f_k) \tag{1.6}$$

$$h_3(d, f) = \log \prod_{k=1}^{K} p(f_k|d_k) \tag{1.7}$$

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

$$h_4(d, f) = \log \prod_{i=1}^{I} p(d_i|d_{i-1}, ..., d_{i-N+1}) \tag{1.8}$$

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

$$h_5(d, f) = I \tag{1.9}$$

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

- Penalización por número de palabras.

  Se trata de utilizar dos informaciones simples muy utilizadas en la bibliografía [Zens, R., 2004] y [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.

- Modelo de lenguaje

  Este modelo se combina con la probabilidad de traducción tal y como se muestra en la ecuación (1.2). Pretende dar coherencia al texto destino que se obtiene con la concatenación de unidades.

  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 [Kneser, R., 1995] e interpolación.

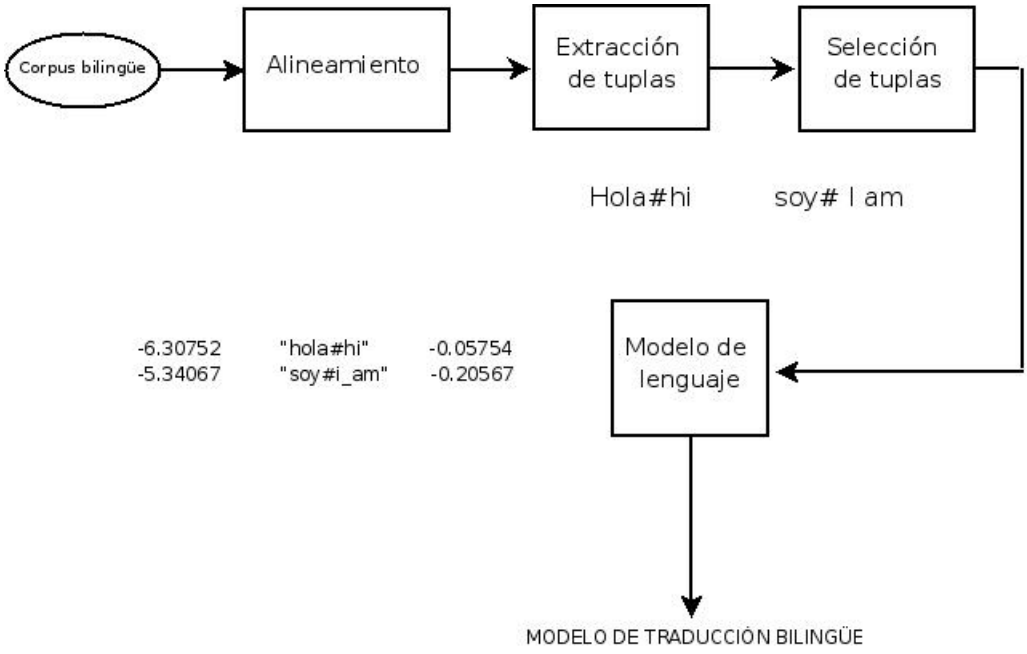

Figura 1.4: *Obtención del modelo de traducción bilingüe.*

Una vez obtenido el modelo de traducción bilingüe utilizaremos el decodificador para traducir el texto deseado.

1.2. 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 (1.2).

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. Tal probabilidad máxima, se calcula como

combinación lineal de los modelos utilizados en el sistema de traducción.

La traducción del material de test se lleva a cabo mediante la herramienta MARIE (Crego, Mariño y de Gispert, 2005), que maximiza la función (1.2) mediante un algoritmo de programación lineal de búsqueda de haz. La búsqueda construye traducciones parciales, 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 mantiene las mejores hipótesis y aquellas que tienen asignada una puntuación próxima a la mejor hipótesis de la lista.

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.

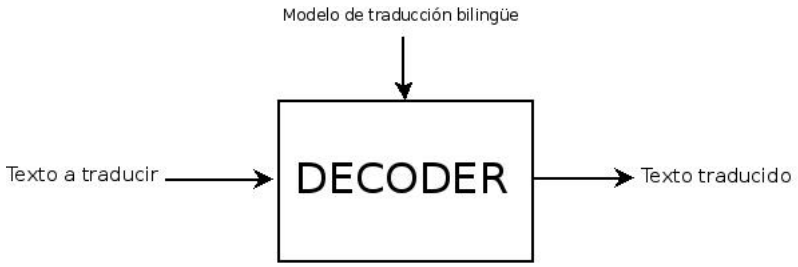

Figura 1.5: *Traducción mediante decodificador.*

1.3. Evaluación de la traducción.

La calidad del sistema elegido para realizar la traducción está univocamente relacionada con la calidad de los textos traducidos mediante los distintos sistemas. Queremos realizar una traducción de forma automática. Por ello utilizaremos una serie de métricas popularmente utilizadas dentro del área de la Traducción Automática. A continuación exponemos las métricas utilizadas:

- WER (*word error rate*): Esta medida ha sido heredada de la evaluación de sistemas de reconocimiento del habla dado que en este campo ha sido aceptada como estándar de evaluación. El WER consiste en el cálculo del mínimo número de operaciones de edición (sustituciones, borrados e inserciones) necesarias para convertir una frase generada por un traductor en otra considerada como traducción de referencia de la frase entrada. Este

mínimo se calcula mediante un algoritmo de programación dinámica, normalmente conocido como la distancia de Levenstein.

- PER (*position-independent word error rate*): Un problema del WER es que requiere que el orden de las palabras sea perfecto con respecto a las frases de referencia. Esto en particular es problemático para idiomas en los que el orden de las palabras puede ser bastante diferente, es decir, que a pesar de que el orden de las frases en la frase generada y la frase de referencia sea muy distinto, ambas podrían ser traducciones válidas de una frase de entrada. Este efecto no puede ser captado por el WER con lo que también utilizaremos la medida PER la cual hace exactamente lo mismo que el WER pero sin tener en cuenta el orden de las palabras, de modo que en el caso de palabras de la frase generada que no tengan contrapartida en la frase referencia se contarán como errores de sustitución. En el caso en que las longitudes de ambas frases sean distintas, el resto de palabras darán lugar adicionalmente a errores de inserción o borrado. Evidentemente, el PER siempre será menor o igual que el WER.

- BLEU *score*: Esta medida mide la precisión de unigramas, trigramas y cuatrogramas con respecto a todo un conjunto de traducciones de referencia, penalizando aquellas frases muy cortas. Por contra de las medidas anteriores el BLEU mide aciertos, con lo que a mayor *score* mejor traducción. Se trata de una media geométrica ponderada de correspondencias entre tuplas, modificada para penalizar tanto la sobregeneración de expresiones correctas, como la producción de oraciones menores a las de referencia. El resultado de la comparación da una métrica numérica que indica la cercanía de las oraciones de prueba respecto a las de referencia. BLEU aporta valores entre 0 y 1, siendo 1 la traducción perfecta y 0 incorrecta.

- NIST *score*: El método NIST es similar a la método BLEU en la medida en que también utiliza n-gram concurrencias de precisión. Sin embargo, se toma la media aritmética de los n-gramas.

Capítulo 2

Filtrado estadístico

La aproximación estadística al traducción automática es una aproximación basada en corpus. Un corpus es una recopilación de textos. Los corpus utilizados en la traducción automática están alineados a nivel de frase, a estos corpus los denominamos corpus paralelos. Nosotros trabajaremos con los corpus bilingües.

El problema que se nos plantea es que no todos los corpus de que disponemos se adecúan a nuestras necesidades, en ellos aparecen frases no paralelas lo cual provoca que obtengamos tuplas erróneas, recordemos que las tuplas definen una segmentación única y monotónica del par de oraciones. Algunos ejemplos de frases no paralelas en diferentes corpus:

Corpus Español - Inglés

EN línea 70 *We see that the French Government has sent a mediator .*
ES línea 70 *Constatamos que la situación cambia día a día .*
EN línea 69 *I think that in view of present events in the Middle East , we ought to ask whether the Council can make a statement on Wednesday afternoon about the way things are going . We note that the situation is changing every day .*
ES línea 69 *Debido a los acontecimientos que están teniendo lugar en estos momentos en Oriente Medio , creo que deberíamos preguntar a el Consejo si el miércoles por la tarde no puede emitir una declaración sobre la situación*

Figura 2.1: ***Frases no paralelas***

Como vemos en el primer ejemplo la traducción de las frases no es correcta y en el segundo caso lo que se considera una frase en el primer idioma se ve fragmentado en dos en el segundo. Dichos errores afectan negativamente a nuestro sistema de traducción, en estos casos es mejor eliminar la información antes que incorporar unidades de traducción erróneas a nuestro sistema.

Disponemos de un corpus bilingüe de entrenamiento, nuestro objetivo en este punto va a ser analizar dicho corpus y modificarlo para intentar obtener un nuevo corpus que nos permita mejorar nuestro sistema de traducción, concretamente nuestro objetivo será reducir las tuplas erróneas y mejorar el vocabulario bilingüe. Para ello debemos implementar algún método que nos permita la detección de frases dentro del corpus que no sean paralelas, es decir necesitamos una herramienta para detectar y eliminar posibles errores en el alineamiento de las frases.

2.1. Planteamiento de la solución 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.

Dicho modelo surge de la necesidad de establecer un alineamiento entre las palabras de un par de oraciones, dado 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:

Modelo IBM1

$$h_{LEX}(s,t) = log\frac{1}{(I+J)^J}\prod_{j=1}^{J}\sum_{i=0}^{I} p_{IBM_1(t_j^n|s_i^n)} \tag{2.1}$$

Donde t_j^n y s_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 $p_{IBM_1(t_j^n|s_i^n)}$ serán las probabilidades de traducción en la dirección fuente-destino $p(t_k/s_k)$ 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:

Modelo IBM1 inverso

$$h_{LEX}inv(s,t) = log\frac{1}{(J+I)^I}\prod_{i=1}^{I}\sum_{j=0}^{J} p_{IBM_1(s_i^n|t_j^n)} \tag{2.2}$$

Donde tenemos $p_{IBM_1(t_j^n|s_i^n)}$ serán las probabilidades de traducción en la dirección $p(s_k/t_k)$ asignada por el modelo IBM1.

En nuestro caso disponemos de dos ficheros que contienen las probabilidades IBM1 palabra fuente - palabra destino y la probabilidad palabra destino - palabra fuente, que tienen el siguiente formato:

PALABRA_FUENTE	PALABRA_DESTINO	PROBABILIDAD
PALABRA_DESTINO	PALABRA_FUENTE	PROBABILIDAD

Haremos un análisis frase a frase del corpus de entrenamiento, para ello calcularemos la probabilidad de cada frase eliminando los pares de frases del corpus con menor probabilidad.

2.2. Selección de frases utilizando PER

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

El sistema basado en PER consiste simplemente en utilizar el sistema de traducción original, con él traduciremos uno de los dos textos que componen

el corpus bilingüe, para luego evaluar dicha traducción. La evaluación de la traducción se realiza por comparación del texto traducido con el texto original del corpus.

Disponemos de una herramienta que nos permite calcular el PER de cada par de líneas paraleas de los textos. Nuestro criterio de selección de frases será, eliminar los pares de líneas cuyo PER sea peor que el del resto, en este caso el umbral que marcará el límite entre frases aceptadas y eliminadas vendrá determinado por el sistema basado en el modelo IBM1, esto se debe a que deseamos comparar la eficiencia de un sistema respecto del otro. Así pues el número de frases eliminadas deberá ser el mismo en ambos casos.

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.

2.3. Modificaciones del Modelo IBM 1

Generalmente el modelo descrito anteriormente se utiliza para computar el peso léxico de dos tuplas, pero en nuestro caso no estamos analizando tuplas, sino líneas de texto. Si calculáramos las probabilidades de frases paralelas directamente, podríamos llegar a conclusiones que no se ajustaran completamente a la realidad, puesto que debemos tener en cuenta varios factores que influyen en el cálculo de la probabilidad, como son la longitud de la frase y la repetición de palabras con una gran probabilidad de coincidencia.

2.3.1. Normalización

Vemos que las ecuaciones 2.2 y 2.1 aparece un factor que condiciona la probabilidad a la longitud de las frases origen y destino.

En el caso de no tener en cuenta la longitud de la frase, podemos ver que al comparar dos frases del corpus la longitud es inversamente proporcional a la probabilidad de que sean paralelas. A mayor número de palabras más probable será que dos frases o incluso partes de la frase no sean paralelas. Teniendo en cuenta que debemos comparar frase de muy distinto tamaño parece coherente pensar que debemos normalizar las frases por su longitud, es decir por su tamaño en número de palabras, para que las probabilidades sean del mismo orden de magnitud

Por lo tanto normalizaremos el valor de la probabilidad obtenida, por el número de palabras que forman la oración, nuestro corpus está formado por frases de muy diversa longitud, y los resultados obtenidos, para cada par de frases, sin normalizar no serían comparables entre sí.

2.3.2. Stopwords

Por otra parte 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 par de idiomas de un corpus.

2.4. Eliminación de frases iguales.

En algunos casos el corpus de entrenamiento contiene una gran cantidad de frases iguales, es decir, frases alineadas en los dos textos que son exactamente iguales. Esto suele ocurrir en idiomas como el castellano y el catalán que comparten una zona geográfica y por lo tanto, el conocimento de ambos idiomas hace que algunos fragmentos tales como diálogos o nombres propios no se traduzcan.

La existencia de estas frases no aporta nada a nuestro sistema de traducción, pero incrementa el tamaño del corpus y con ello el coste computacional que se requiere. Así pues en alguno de los experimentos realizados utilizaremos la eliminación de frases iguales.

2.5. Proceso experimental

Disponemos de diferentes corpus de entrenamiento donde vamos aplicar nuestro criterio de eliminación de frases paralelas, con el fin de optimizar dichos corpus y obtener posteriormente un modelo de traducción más eficaz que el actual.

Para ello haremos un análisis de las probabilidades de frases paralelas y una posterior eliminación de las frases que consideremos que no aportan nada positivo al sistema de traducción, para cada uno de los corpus de que disponemos. En este momento dispondremos de dos textos alineados de frases paralelas.

Para evaluar le necesidad de modificar el corpus de entrenamiento utilizaremos el método siguiente, con los dos textos alineados y diferentes herramientas, crearemos un modelo de traducción. Una vez obtenido dicho modelo traduciremos un texto de referencia y evaluaremos automáticamente mediante las medidas WER, PER, BLEU y NIST.

Este proceso lo realizaremos con el corpus original y posteriormente con varios corpus modificados, con el fin de evaluar las mejoras obtenidas en el proceso de selección de frases paralelas.

2.6. Experimentos

Realizamos el proceso de análisis para los siguientes corpus:

-CHINO - INGLÉS

-CASTELLANO - CATALÁN

-CASTELLANO - INGLÉS

2.6.1. Descripción del sistema de referencia

Alineamiento

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 (esta categorización es uno de los objetivos de este proyecto y será definida posteriormente). Se han eliminado los pares bilingües en el que una de las oraciones contenía más de 50 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++ [Och, F.J. and Ney, H., 2003]. se realizó el alineamiento de los textos bilingües paralelos del materia de entrenamiento, ejecutándose 4 iteraciones del modelo IBM1, 5 iteraciones del modelo HMM 3 iteraciones del modelo IBM4 y ninguna del modelo IBM3. Se obtuvo 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 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 mejor cobertura de los enlaces entre las palabras de ambas lenguas, que es importante para generar elementos bilingües correctos. El segundo genera enlaces con alta precisión, que serán usados para la traducción de palabras.

Selección de tuplas

Una vez obtenido el alineamiento unión se procedió a la segmentación en tuplas del material de entrenamiento. Tras analizar las tuplas en función del número de apariciones en el entrenamiento y el número de traducciones diferentes que las tuplas ofrecen para una misma palabra fuente, puede observarse que la mayor parte de las tuplas aparecen muy pocas veces y abundan las tuplas que ofrecen un número reducido de traducciones alternativas.

A efectos de simplificar el sistema de traducción, el vocabulario de tuplas se limitó a aquellas que tengan una longitud máxima de 15 palabras tanto en el lenguaje fuente como en el lenguaje destino.

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 del 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, R., 1995]. El modelo generado fue un trigrama (N=3) de tuplas.

Decodificación

La traducción del material de test fue llevada a cabo mediante la herramienta MARIE [José B.˜Marino, 2006], que maximiza la función U en (1.3) mediante un algoritmo de programación lineal de 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 aquellas 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 para simplificar el sistema.

2.6.2. Corpus CHINO - INGLÉS

En primer lugar vamos a aplicar nuestro criterio de análisis al corpus proporcionado por el NIST. Se corresponde a la parte del corpus de noticias del corpus utilizado como entrenamiento en la evaluación del NIST 2006.——

Disponemos de dos textos alineados con las siguientes características:

Cuadro 2.1: *news.zh - news.en Parallel corpus.*

CORPUS	sent	words	vocab.	Lmax	Lmean
news.zh	1020248	26322847	157206	100	25.8
news.en		27246624	215828	100	26.7

Nuestra intención es hacer una representación gráfica del comportamiento de todos los pares de frases del corpus en función de probabilidad de que sean paralelas. Expresaremos la probabilidad en logaritmo negativo, con el fin de que el análisis de las gráficas sea más intuitivo, y nos referiremos a ella con el nombre de coste.

Distribución del coste para el corpus CHINO - INGLÉS

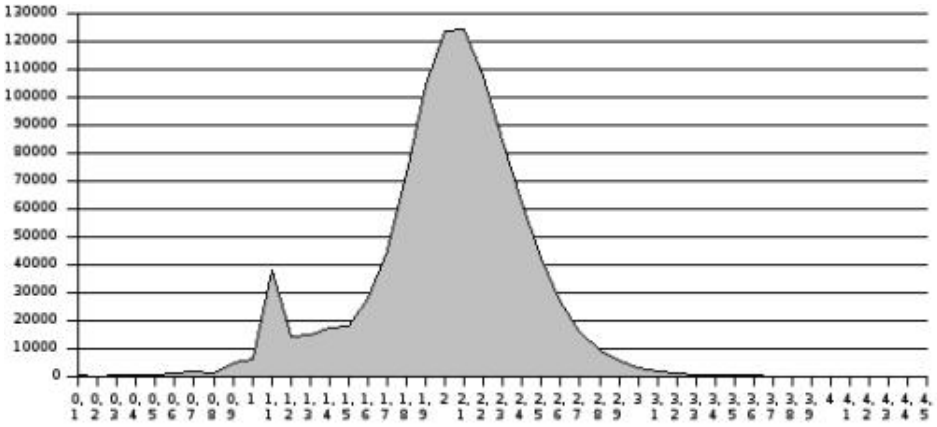

Figura 2.2: *Modelo IBM1 sin Stopwords.*

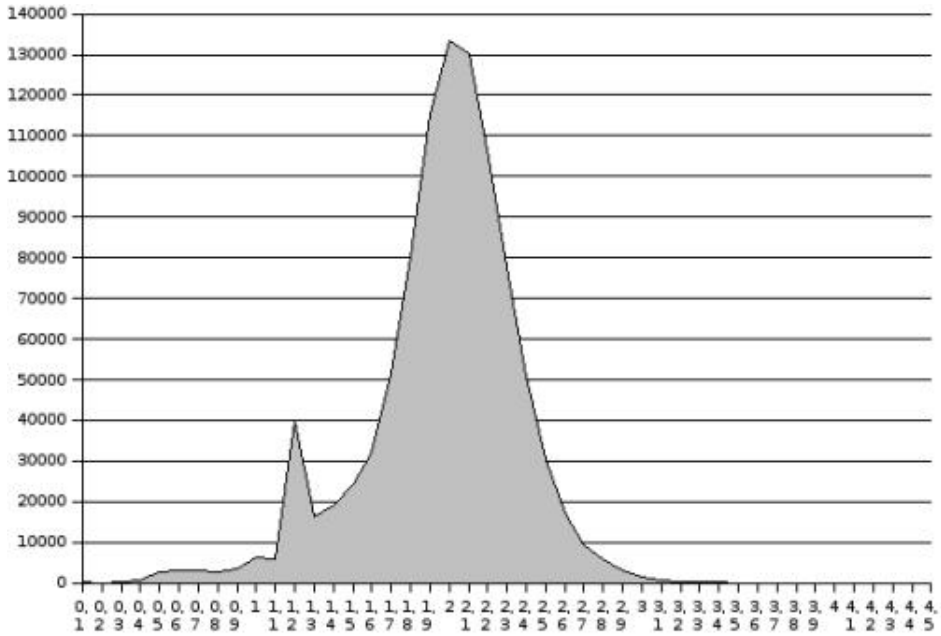

Figura 2.3: *Modelo IBM1 inverso sin Stopwords.*

En los gráficos 2.2 y 2.3 mostramos los resultados obtenidos de la expresiones 2.1 y 2.1, donde indicamos el número de oraciones (eje Y) con una determinada probabilidad de ser paralelas (la probabilidad está expresada en logaritmo negativo o sea coste).

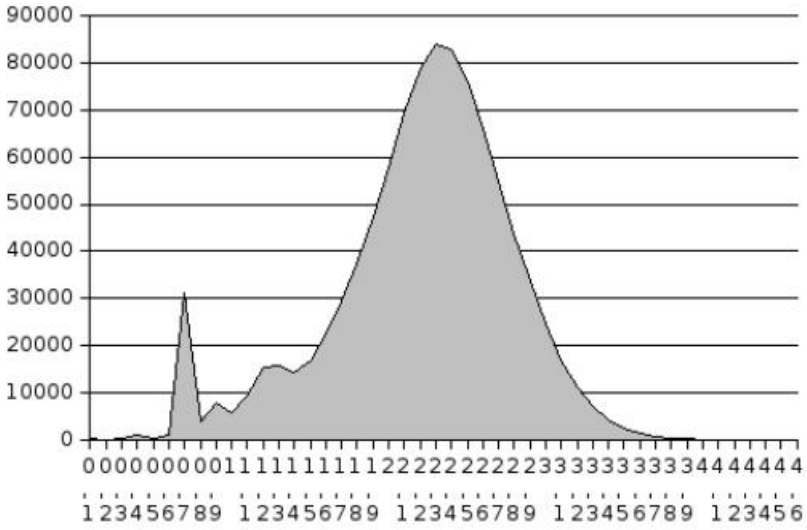

Figura 2.4: *Modelo IBM1 con Stopwords.*

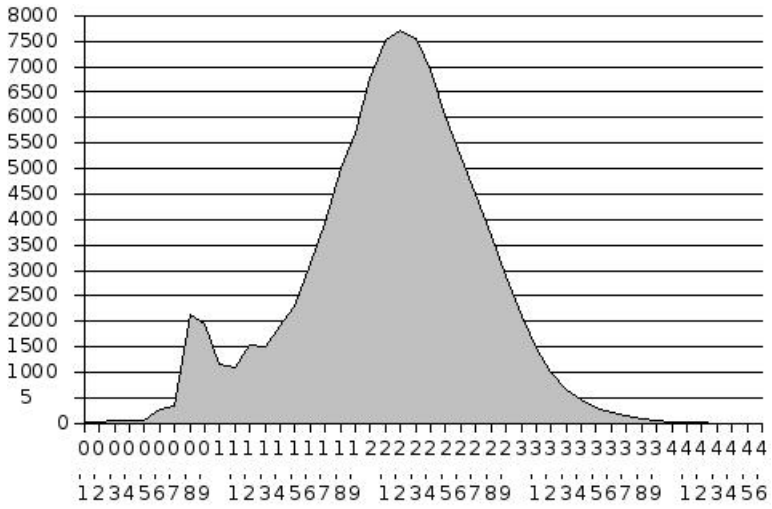

Figura 2.5: *Modelo IBM1 inverso con Stopwords.*

Los gráficos 2.4 y 2.5 muestran los resultados obtenidos al no tener en cuenta el coste aportado por las Stop Words.

A continuación mostramos las palabras mas comunes de los textos en ambos idiomas, las stop words:

STOP WORDS INGLÉS - CHINO

INGLÉS: /the/ /,/ /./ /of/ /and/ /to/ /in/ /a/ /“/ /for/ /)/ /(/ /on/ /is/ /that/ /by/ /whith/ /be/ /at/ /will/

CHINO: /的/ /。/ /在/ /、/ /及/ /「/ /」/ /是/ /）/ /（/ /一/ /有/ /於/ /和/ /为/ /,/ /：/ /会/ /了/

Como vemos el hecho de utilizar stopwords modifica la distribución de la probabilidades, aunque la forma de la gráfica no varía, si se ven modificados los valores máximos, es decir hay menos frases que comparten la probabilidad más común, . Es un resultado lógico tras haber eliminado las palabras y signos de puntuación más comunes, que incrementan considerablemente la probabilidad IBM1, pero sin aportar demasiada información a la hora de traducir un texto.

Una vez tenemos la distribución de las probabilidades por número de frases, podemos establecer un umbral a partir del cual consideremos que la probabilidad de que un par de frases sean paralelas es tan baja que podemos eliminar dicho par de frases.

Debemos encontrar el umbral adecuado para conseguir mejorar nuestro corpus, si escogemos un umbral demasiado bajo y eliminamos más pares de frases de las necesarias, el corpus de entrenamiento resultante se verá tan reducido que los resultados obtenidos serán peores a los iniciales. Puesto que el modelo no es simétrico debemos escoger un umbral para el modo directo y otro para el modo inverso. Utilizaremos el procedimiento de prueba y error con diferentes umbrales para ver con cual obtenemos mejores resultados

Umbrales propuestos para la selección de frases.

	Umbral directo	Umbral inverso	Frases descartadas
Corpus modificado C1	3.1	3.1	5.3 %
Corpus modificado C2	3	2.95	8.1 %
Corpus modificado C3	2.9	2.85	13 %
Corpus modificado C4	2.7	2.65	24 %

Análisis de los resultados CHINO - INGLÉS

En primer lugar haremos un análisis de las frase eliminadas, para ver si podemos justificar que la probabilidad de ser paralelas sea tan baja. Como es de esperar dicho análisis requiere de un cierto conocimiento del idioma, en nuestro caso podemos llegar a extraer alguna conclusión observando las frase en

inglés pero difícilmente apreciaremos alguna característica del chino.

Realizamos un breve estudio estadístico y observamos que el corpus original tiene una media de 147 palabras por linea. Por otro lado nuestro corpus modificado tiene unas 130 palabras por linea, mientras que las frases descartadas tiene 221 palabras por linea. Eso nos hace pensar en la influencia de la longitud de la frase a la hora de establecer una probabilidad de paralelismo.

Este resultado es consecuencia directa de la normalización del la fórmula IBM1 2.1 por la longitud de las frases origen y destino, al normalizar las frase largas disminuye la probabilidad de que sean paralelas.

Por otro lado vemos que las frases descartadas, que no son excesivamente largas, tienen gran cantidad de nombres propios o números entre sus palabras, es lógico que así suceda puesto que en las tablas de probabilidad IBM1 no aparecen todos los números ni tampoco los nombres propios.

A continuación mostramos varias frases que muestran comentado con anterioridad:

- The 12 recipients honoured today with the GBM are : Mr Wong Ker Lee , Mr Ann Tse Kai , Mr Cha Chi Ming , Dr Chung Sze Yuen , Mr Simon Li Fook Sean , Mr Tsang Hin Chi , Mr Chuang Shih Ping , Mrs Elsie Tu , Mr Xu Simin , Mr Lee Quo Wei , Mr Fok Ying Tung , and Mr Lo Tak Shing .

- 今日 将 获 颁 大 紫荆 勋章 的 十二 位 人士 为 : 黄 克立 先生 安子介 先生 查济民 先生 钟 士元 博士 李 福善 先生 曾 宪梓 先生 庄 世平 先生 杜 叶 锡恩 女士 徐 四民 先生 利国伟 先生 霍 英东 先生 罗 德丞 先生

- Sha Tau Kok August 25 and 26 , 1997 Cheung Chau August 26 and 27 , 1997

- 流动 注射 队 前往 渔港 的 时间 如下

Resultados con nuevos modelos.

A continuación mostramos los resultados de la utilización de diferentes modelos, creados a partir del corpus original o de los diferentes corpus filtrados por la selección de frases.

Tipo	Corpus Inicial	C1	C2	C3	C4
BLEU score	12.10	12.20	12.21	12.22	12.20
NIST score	5.37	5.37	5.38	5.38	5.31
PER score	53.82	53.75	53.90	53.79	54.17
WER score	70.35	70.32	70.32	70.35	70.64

Cuadro 2.2: *Resultados obtenidos con los diferentes corpus de entrenamiento.*

Como podemos observar analizando los resultados del BLEU, todos los superan al corpus original, los mejores resultados se obtienen en los corpus C2 y C3, según el criterio de evaluación del NIST estos dos corpus también mejoran los resultados originales, cosa que no ocurre con los corpus C1 y C4. A la hora de elegir entre el corpus C2 y C3 vemos que aunque los resultados son muy similares, la evaluación del BLEU es ligeramente mejor en el corpus C3 que en el C2.

Así pues debemos elegir el modelo que proviene del corpus C3, es decir de aquel cuyo umbral directo es de 2.9 y el inverso era de 2.85 y que descartaba el 13 % de las frases por no ser paralelas.

Criterio selección OR.

Implementamos una nueva condición del modelo, en la que tenemos en cuenta un criterio diferente en la selección de frases.Hasta este momento para que una frase fuera seleccionada, tanto la probabilidad directa como la inverso debían superar el umbral que hemos establecido. Este nuevo criterio consiste en admitir una frase como válida si alguna de las dos probabilidades supera el umbral establecido, por ello le denominamos *criterio de selección OR*, ya que *P. directa* $\geq$ *P. umbral* o *P. inversa* $\geq$ *P. umbral*.

Establecemos un umbral que nos permita eliminar el mismo número de líneas que en el corpus C3. Ello lo conseguimos estableciendo un umbral tanto directo como inverso de 3.

Construimos el modelo con el nuevo corpus al que denominamos C5 y obtenemos los siguientes resultados:

Tipo	C5
BLEU score	12.53
NIST score	5.60
PER score	53.39
WER score	70.68

Cuadro 2.3: *Resultados Corpus entrenamiento C5*

Resultados obtenidos con el sistema PER.

Aplicamos el sistema de selección de frases basado en PER. Para ello utilizaremos utilizaremos un umbral de PER que nos permita eliminar el mismo número de frases que las eliminadas por el modelo que nos ha dado mejores resultados. En este caso tomaremos como referencia el corpus C5 que elimina el 13 % de las frases siguiendo el modelo IBM1 y con el criterio de selección OR, puesto que ha sido el que mejores resultados nos ha dado..

Bien una vez eliminado el 13 % de las frases con este método reconstruimos el sistema con el nuevo corpus que denominaremos C6 y realizamos el test de traducción, obteniendo los siguientes resultados.

Tipo	C6
BLEU score	12.68
NIST score	5.70
PER score	52.99
WER score	70.49

Cuadro 2.4: *Resultados Corpus entrenamiento C6*

2.6.3. Corpus CASTELLANO - CATALÁN

Disponemos de un corpus de entrenamiento para los idiomas Español y Catalán, facilitado por el diario El Periódico de Cataluña. Este corpus tiene las siguientes características:

Cuadro 2.5: *Corpus paralelo train.es - train.ca.*

CORPUS	sent	words	vocab.	Lmax	Lmean
train.es	2178796	42375299	390372	107	19.4
train.ca		44430128	381879	108	20.4

Distribución de probabilidad para el corpus CASTELLANO - CATALÁN

Al igual que en el corpus anterior vamos a proceder a hacer una representación gráfica del coste de que dos oraciones sean paralelas y a seleccionar un umbral que determine que oraciones no van a ser consideradas paralelas.

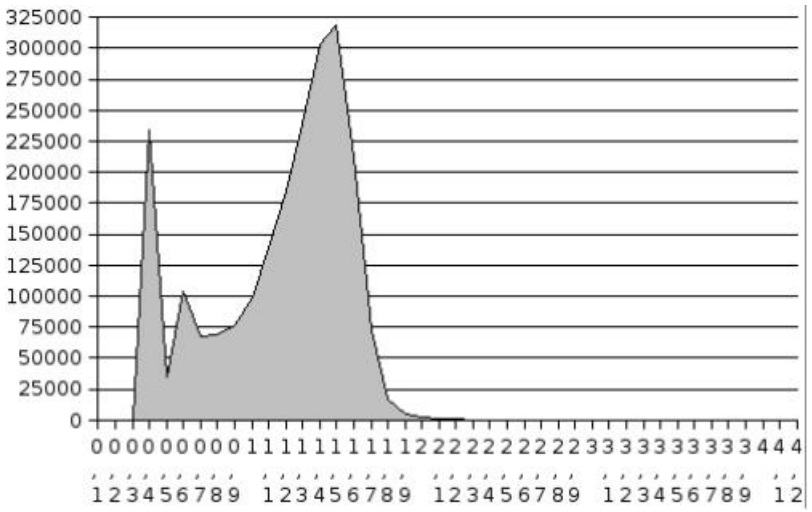

Figura 2.6: *Modelo IBM1 sin Stop Words.*

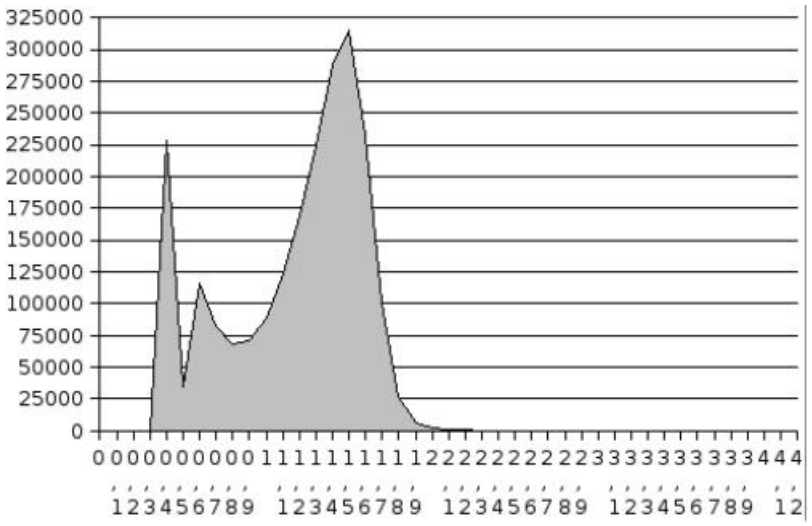

Figura 2.7: *Modelo IBM1 inverso sin Stop Words.*

Los gráficos 2.6 y 2.7 muestran los resultados obtenidos teniendo en cuenta la probabilidad aportada por las Stop Words. Por otro lado los gráficos 2.8 y 2.9 muestran los resultados obtenidos sin tener en cuenta esa probabilidad.

STOP WORDS CASTELLANO - CATALÁN

CASTELLANO: /de/ /,/ /./ /el/ /la/ /que/ /en/ /a/ /y/ /"/ /los/ /se/

/un/ /las/ /por/ /-/ /una/ /con/ /para/ /)/ /(/ /es/ /ha/ /más/ /como/ /lo/ /ayer/ /sus/ /:/ /le/

CATALÁN: /de/ /,/ /./ /el/ /a/ /la/ /que/ /i/ /l'/ /va/ /els/ /d'/ /“/ /en/ /per/ /un/ /les/ /-/ /una/ /amb/ /no/ /es/ /ha/ /més/ /és/ /van/ /)/ /(/ /s'/ /com/

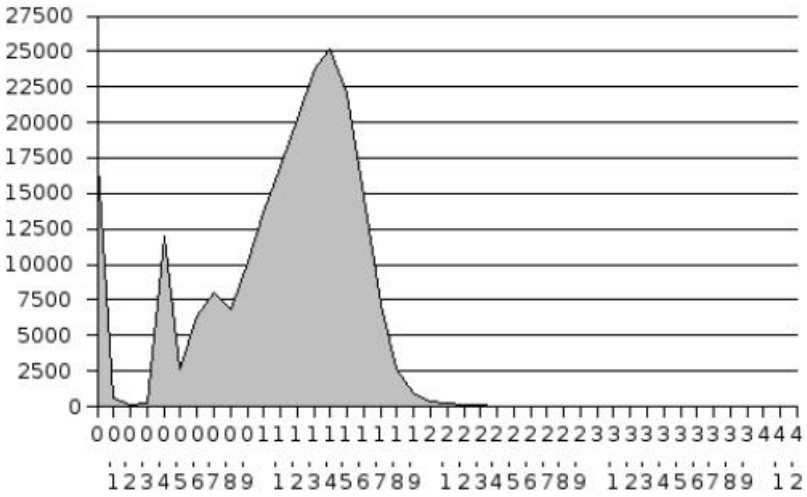

Figura 2.8: *Modelo IBM1 con Stop Words.*

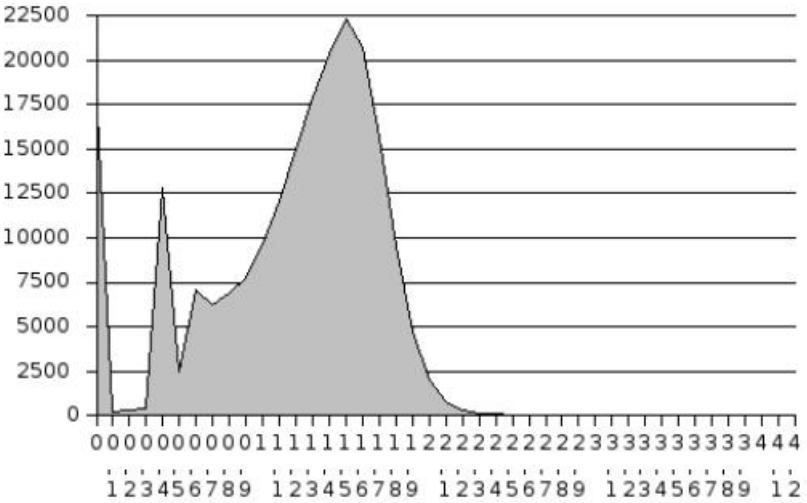

Figura 2.9: *Modelo IBM1 inverso con Stop Words.*

Umbrales propuestos para la selección de frases.

En este caso hacemos una primera eliminación de frases, que corresponde aproximadamente al 10 % de las frases existentes y otra que descarte el 20 %. Para ello determinamos, para cada caso, un umbral directo y otro inverso, de tal manera que la eliminación de frases sea equitativa en los dos sentidos del cálculo de la probabilidad.

	Umbral directo	Umbral inverso	Frases descartadas
Corpus modificado C1	1.7	1.8	11 %
Corpus modificado C2	1.58	1.72	20 %

Análisis de los resultados CASTELLANO - CATALÁN

En el análisis de frases eliminadas de este corpus, vemos ciertas similitudes con respecto al corpus anterior (CHINO - INGLÉS). Por un lado analizaremos la longitud de las frases. Tenemos que en el corpus original aparecen unas 20 palabras por frase, mientras que en el grupo de frases eliminadas aparecen casi 40 palabras por frase. Confirmamos la influencia de la longitud de las frases en nuestro criterio de análisis.

Por otro lado el hecho de que conozcamos los dos idiomas que forman el corpus nos permite discernir si las frases descartadas han sido descartadas correctamente o no.

A continuación podemos ver algunos ejemplos de frases descartadas:

- Para reforzar la idea de permanencia , el arquitecto Josep Maria Tera estudia la reforma de esta escuela nacida en enero de 1929 gracias a medio_millón de pesetas de un confitero culto .

- Per donar forç a la idea de permanència , lárquitecte Josep Maria Tera estudia la reestructuració de la Massana , aquella escola creada el gener de el 1929 amb mig milió de pessetes dún pastisser culte.

- Hi ha uns nivells molt variats de sordeses , que van des de els que no hi senten gairebé gens , i per tant no els serveixen de res els audiòfons , fins a altres (entre els quals es troba el senyor Vallhonrat) que tenen la sort de sentir -hi més i de poder aprofitar els avantatges de els audiòfons .

- Hay unos niveles muy amplios de sordera que van desde los que no oyen casi nada y por lo tanto no hay audífonos que les valga , hasta otros (entre los que se encuentra el señor Vallhonrat) que tienen la suerte de oír más y

aprovechar las ventajas de los audífonos .

- La gestió de els tres membres de el Govern de els quals el president possiblement vol prescindir (Isabel Tocino , Rafael Arias-Salgado i Margarita Mariscal de Gante) presenta moltes ombres , sobretot si es compara amb la de Rodrigo Rato , Jaime Mayor Oreja o Mariano Rajoy , per esmentar -ne tres més .

- La gestión de los tres miembros de el Gobierno de los que el presidente posiblemente quiere prescindir (Isabel Tocino , Rafael Arias-Salgado y Margarita Mariscal de Gante) presenta muchas sombras , sobre todo si se compara con la de Rodrigo Rato , Jaime Mayor Oreja o Mariano Rajoy , por nombrar otros tantos .

Vemos que en ningún caso dichas frases deberían ser descartadas puesto que su traducción es totalmente correcta. Si bien el hecho que aparezcan nombres propios y de que su longitud sea más larga que la del resto de frases, han influido en que formen parte de nuestro grupo de frases descartadas.

Estos resultados no deben sorprendernos, ya que el criterio que hemos seguido a la hora de establecer un umbral a partir del cual íbamos a considerar las frases como no paralelas, ha sido el de eliminar un porcentaje dado de frases a eliminar. Pero si observamos la gráfica de distribución de la probabilidad, veremos que el corpus que estamos tratando tiene una probabilidad de que las oraciones del corpus sean paralelas es muy elevada, por eso, al eliminar oraciones según un criterio porcentual, estamos eliminando en realidad oraciones paralelas, con menos probabilidad que el resto, pero paralelas al fin y al cabo.

Podemos llegar a la conclusión de que el corpus CASTELLANO - CATALÁN es un buen corpus de entrenamiento en cuanto a la alineación de oraciones paralelas se refiere, y que pocas mejoras obtendremos aplicando esta herramienta de selección de oraciones paralelas.

Resultados con nuevos modelos.

Realizamos la evaluación de todas maneras para ver que resultados obtenemos utilizando los nuevos modelos provenientes de los corpus tratados

Podemos ver que los resultados no mejoran, sino que incluso empeoran. Esto concuerda con lo anteriormente comentado, puesto que podemos llegar a eliminar frases correctas, prescindiendo así de su aportación a la hora de crear el modelo de traducción. Debemos buscar otros métodos para mejorar este corpus, puesto que como acabamos de demostrar la selección de frases paralelas

Tipo	Corpus Inicial	C1	C2
BLEU score	84.69	83.21	83.02
NIST score	14.03	13.72	13.70
PER score	8.3	9.02	9.12
WER score	9.51	10.37	10.46

Cuadro 2.6: *Resultados Corpus entrenamiento inicial,corpus C1 y corpus C2*

utilizando la probabilidad IBM1 no es útil en este caso.

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. Para efectuar dicha comparativa no emplearemos el mismo corpus que hemos utilizado en el experimento anterior sino una versión más reducida del mismo. El motivo de dicha modificación es que un corpus tan amplio como el anterior, junto con el elevado coste computacional que requiere el sistema de filtrado estadístico utilizando el PER dilataría mucho la obtención de los resultados, y lo que realmente nos interesa es el valor relativo de los resultados, es decir, el resultado de cada filtrado estadístico comparado con el resultado original.

Al igual que en el experimento inicial realizaremos un filtrado del 11 % de las frases para los distintos sistemas de filtrado.

Una vez eliminado el 11 % de las frases con este método reconstruimos el sistema con el nuevo corpus que denominaremos C3 y realizamos el test de traducción, obteniendo los siguientes resultados.

Tipo	Corpus Inicial	C IBM1	C IBM1 OR	C PER
BLEU score	83.70	83.62	83.65	83.66
NIST score	13.76	13.75	13.76	13.75
PER score	8.85	8.92	8.89	8.89
WER score	10.10	10.15	10.12	10.13

Cuadro 2.7: *Comparativa de los resultados obtenidos utilizando las diferentes técnicas de filtrado estadístico.*

Como vemos los resultados obtenidos tras el filtrado empleando cualquiera de los métodos propuestos son peores que los resultados originales, eso no hace mas que corroborar el análisis realizado inicialmente. El corpus que estamos tratando es un buen corpus de entrenamiento en cuanto a la alineación de frases paralelas se refiere, así que estas herramientas no nos serán útiles para mejorar

nuestro sistema de traducción.

Eliminación de frases no bilingües.

Puesto que como hemos visto, el criterio de eliminación de frases utilizando el método basado en IBM1 no es útil para este corpus en concreto, intentaremos utilizar la eliminación de frases iguales con el fin de mejorar el corpus de entrenamiento.

En este caso al analizar el corpus de entrenamiento vemos que hay un 20 % de las frases que son iguales, y que procederemos a eliminar de nuestro corpus, para así obtener un nuevo corpus de tamaño considerablemete reducido, al que llamaremos C4.

A continuación mostramos algunos ejemplos de frases que aparecen en ambos textos:

- *El Periódico on line .*
- *Sallent (Bages) .*
- *MAYKA NAVARRO .*
- *J .*
- *CASABELLA / L .*
- *DIEZ .*
- *Barcelona / Madrid .*

Construimos el sistema con el nuevo corpus y obtenemos los siguientes resultados en la evaluación de la traducción:

Tipo	C4
BLEU score	84.71
NIST score	14.03
PER score	8.36
WER score	9.51

Cuadro 2.8: *Resultados Corpus entrenamiento C4*

Como podemos observar los resultados son prácticamente iguales que los obtenidos utilizando el corpus original, peor los hemos conseguido utilizando un corpus mucho más reducido, con lo cual se reduce el coste computacional.

Ejemplo de frase traducida utilizando la eliminación de frases iguales y sin utilizarla.

Utilizando eliminación de frases
TRG 1: *L '11 - S va ser catastròfic per a Mickey i els seus parcs americans.*
Sin utilizar eliminación de frases
TRG 2: *L '11 - S va ser catastròfic per a Mickey y sus parcs americans*

Se ha aplicado la eliminación de frase iguales para mejorar el traductor CASTELLANO - CATALÁN de la UPC[1].

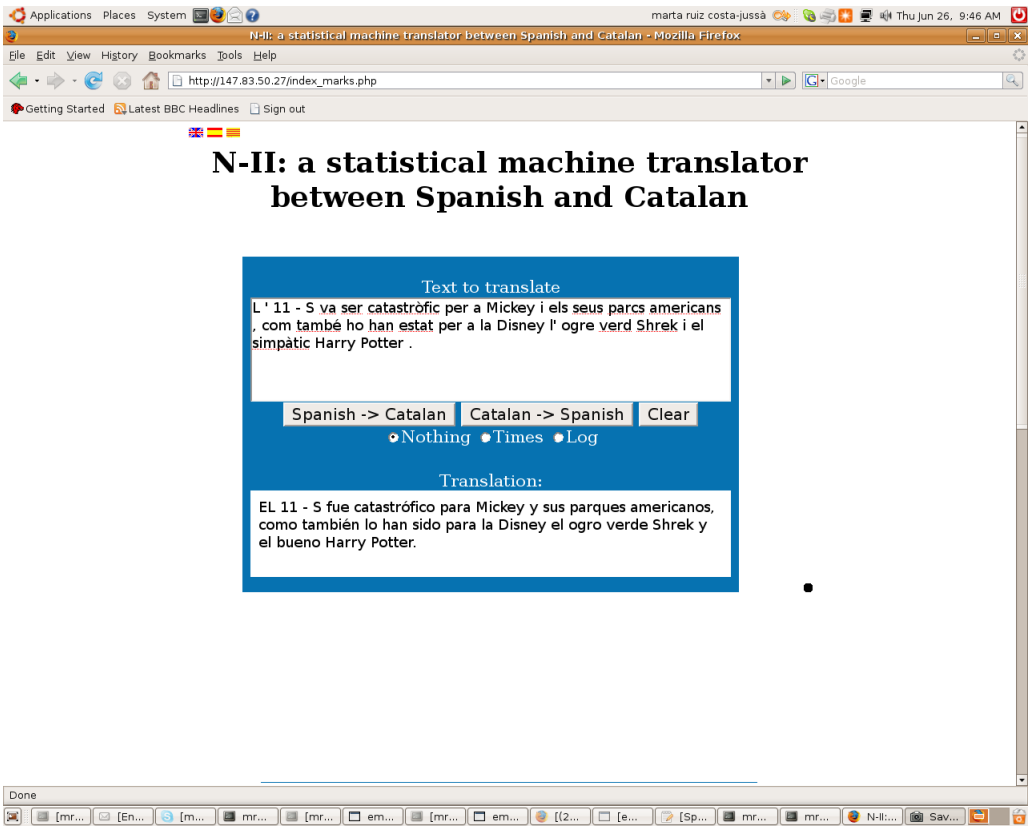

Figura 2.10: *Traductor Catalán Castellano N-II.*

[1] www.n-ii.org

2.6.4. Corpus CASTELLANO - INGLÉS

Disponemos de un corpus de entrenamiento para los idiomas Español e Inglés. Este corpus tiene las siguientes características:

Cuadro 2.9: *Corpus paralelo train.eng - train.spa.*

CORPUS	sent	words	vocab.	Lmax	Lmean
train.eng	1356754	37002435	109827	100	27.3
train.spa		39509651	147551	110	29.1

Distribución de probabilidad para el corpus CASTELLANO - INGLÉS

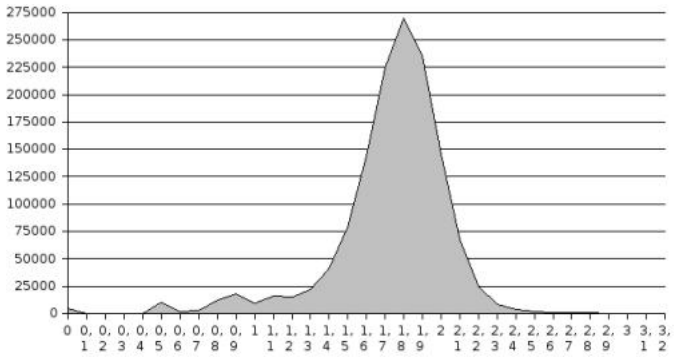

Figura 2.11: *Modelo IBM1 sin Stop Words.*

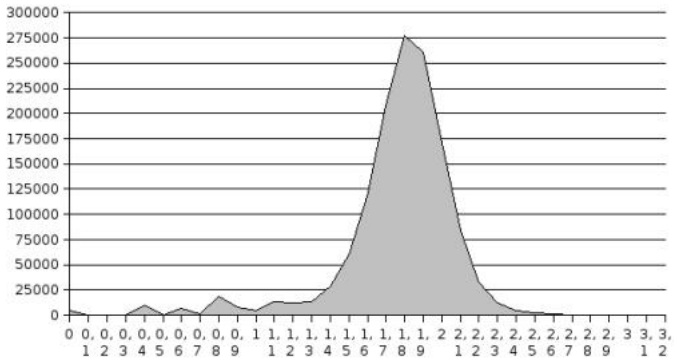

Figura 2.12: *Modelo IBM1 inverso sin Stop Words.*

Los gráficos 2.11 y 2.12 muestran los resultados obtenidos teniendo en cuenta el coste aportado por las Stop Words, mientras que los gráficos 2.13 y 2.14 no tiene en cuenta dicho coste.

A continuación mostramos las palabras que se consideran Stop Words para cada idioma:

STOP WORDS CASTELLANO - INGLÉS

CASTELLANO: /de/ /,/ /./ /el/ /la/ /que/ /en/ /a/ /y/ /"/ /los/ /se/ /un/ /las/ /por/ /-/ /una/ /con/ /para/ /)/ /(/ /es/ /ha/ /más/ /como/ /lo/ /ayer/ /sus/ /:/ /le/
INGLÉS: /the/ /,/ /./ /of/ /and/ /to/ /in/ /a/ /“/ /for/ /)/ /(/ /on/ /is/ /that/ /by/ /whith/ /be/ /at/ /will/ /this/ /I/ /we/ /not/ /are/ /have/ /as/ /has/ /-/ /it/

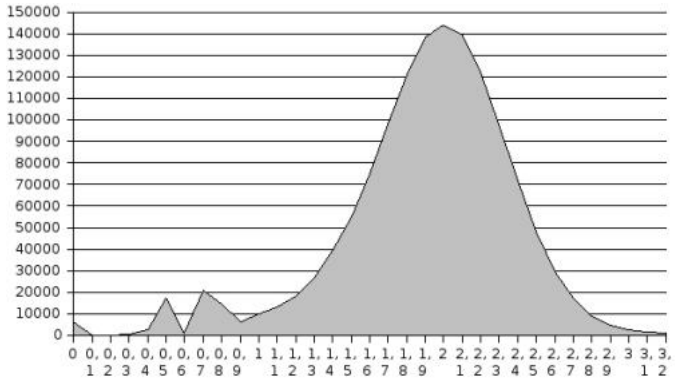

Figura 2.13: *Modelo IBM1 con Stop Words.*

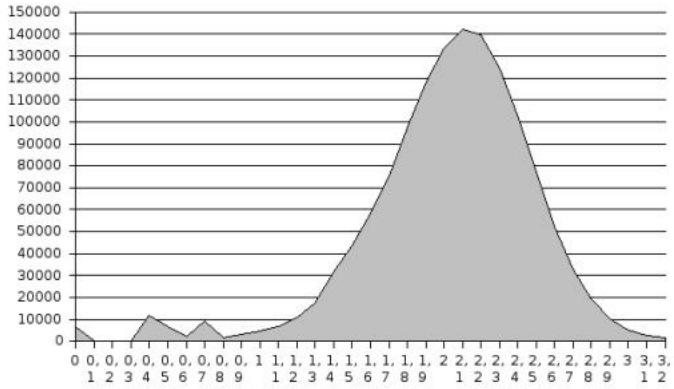

Figura 2.14: *Modelo IBM1 inverso con Stop Words.*

Umbrales propuestos para la selección de frases.

	Umbral directo	Umbral inverso	Frases descartadas
Corpus modificado C1	2.51	2.62	12 %
Corpus modificado C2	2.7	2.8	5 %

Análisis de los resultados CASTELLANO - INGLÉS

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 .
En el ejemplo anterior vemos que hay un desplazamiento de línea de un texto a otro, dicho par de lineas no debe ser tomado en cuenta, y por lo tanto lo eliminaremos de nuestro corpus, ya que la información que nos aporta es errónea.
Al igual que en corpus anteriores 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.
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 .*

- *Las consideraciones van especialmente en esta dirección .*
también es el caso de:

- *To conclude , let me ask what lessons we should be learning .*

- Permítanme que finalice mi intervención con una serie de propuestas .

Evaluación nuevos modelos.

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

Tipo	Corpus Inicial	C1	C2
BLEU score	43.30	45.20	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

Cuadro 2.10: *Resultados Corpus entrenamiento inicial,corpus C1 y corpus C2*

Criterio selección OR.

Establecemos un umbral que nos permita eliminar el mismo número de líneas 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 al que denominamos C3 y obtenemos los siguientes resultados:

Tipo	C3
BLEU score	44.50
NIST score	9.75
PER score	31.01
WER score	40.68

Cuadro 2.11: *Resultados Corpus entrenamiento C3*

En este caso no mejoramos los resultados obtenidos mediante el corpus C1.

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 sistem y obtenemos un corpus al que denominaremos C4 que procedemos a evaluar como hemos hecho en cada caso, obteniendo los siguientes resultados.

Podemos observar que los resultados aunque mejoran los obtenidos con el corpus original, son inferiores a los obtenidos utilizando en el modelo creado a partir del corpus C1, que es el que provee de unos resultados más óptimos. En este caso, el sistema inicialmente propuesto basado en el módulo IBM1 supera a el resto de sistemas analizados.

Tipo	C4
BLEU score	44.30
NIST score	9.73
PER score	31.86
WER score	40.71

Cuadro 2.12: *Resultados Corpus entrenamiento C4*

2.7. Conclusiones

De manera general, podemos concluir que las técnicas desarrolladas para el filtrado estadístico, nos van a ser útiles en dos ámbitos diferentes:

- Establecer la calidad de un corpus en cuanto a la existencia de oraciones paralelas.
- Mejorar los corpus, eliminado aquellas oraciones que no aportan nada positivo a nuestro sistema de traducción. Esta técnicas nos son útiles para detectar frases paralelas de baja calidad, al descartar estas frases ganamos en eficiencia y mantenemos o incluso mejoramos ligeramente la calidad de la traducción.

Se han presentado dos técnicas de filtrado estadístico de corpus bilingües. Dichas técnicas se basan en el Modelo IBM1 y en el PER.

La utilización de una u otra técnica, o incluso la no utilización de ellas, dependerá de varios factores.

En primer lugar debemos tener en cuenta que la eficacia de las técnicas de filtrado propuestas, dependen directamente de la calidad del corpus de entrenamiento tratado. Refiriéndonos a esa calidad como a la buena alineación de frases paralelas. Es decir un corpus de entrenamiento con una muy buena alineación de frases paralelas no requiere de la utilización de estas técnicas, pudiendo ser incluso perniciosas, puesto que llegaríamos a eliminar frases cuya aportación es positiva para el sistema.

Por otro lado la técnica basada en el PER requiere un mayor coste computacional que la basada en el Modelo IBM1, es por eso que incluso en corpus donde los resultados sean mejores utilizando esta técnica deberemos tener en cuenta la disponibilidad de recursos y de tiempo.

En general, 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 realizados en el corpus Chino-Inglés ambas técnicas mejoran ampliamente los resultados originales, en este caso los resultados obtenidos mediante la técnica PER mejoran levemente los obtenidos por el Modelo IBM1, pero los recursos empleados son muy superiores, así pues queda a nuestra elección la utilizacón de una u otra en función de nuestras necesidades.

Por otro lado en los experimentos que hemos presentado para el corpus de Español-Inglés se mejora en más de 1 punto BLEU. Mientras que en los experimentos realizados para el corpus Castellano-Catalán, los resultados obtenidos son inferiores al original, con lo cual podemos determinar que es un corpus con una buena alineación de frases paralelas.

Capítulo 3

Categorización

En cualquier idioma existen cierto tipo de palabras o construcciones que poseen unas características particulares, tienen una estructura específica que nos permite tener una información adicional. Para utilizar esta información en traducción es necesario detectar estas palabras y traducirlas independientemente.

No utilizamos un lenguaje de reglas para este tipo de palabras, sólo tablas de traducción y otras herramientas que detallaremos que nos permiten tratar estas palabras de forma diferente al resto.

Concretamente, nos estamos refiriendo a números o páginas web. Este tipo de palabras o construcciones tienen una estructura concreta, el análisis de estas estructuras nos permite establecer unos criterios que nos permita la detección de las mismas dentro de un texto dado.

La detección de este tipo de datos es fundamental, pero sólo es una parte del trabajo que se debe realizar, puesto que una vez detectados en el lenguaje fuente, debemos encontrar algún método que nos permita encontrar su equivalente en el lenguaje destino.

En este capítulo vamos a tratar dos de estos casos especiales, como son los números y las páginas web.

3.1. Tratamiento de Números.

Con el fin de mejorar nuestro sistema de traducción, implementaremos una serie de herramientas que nos permitan realizar un tratamiento especial para los números que aparecen en nuestros textos de traducción. La idea general es etiquetar todos los números que aparezcan en los textos para poder tratarlos separadamente..

Posteriormente realizaremos un proceso transformación de las etiquetas en los números correspondientes.

En definitiva se trata de hacer un preprocesado, tanto del corpus original como del texto a traducir y finalmente un postprocesado del texto traducido.

3.1.1. Preprocesado del Corpus bilingüe.

En el preprocesado que vamos a realizar del corpus de entrenamiento implementaremos dos funciones básicas, el etiquetado de los números y la creación de un diccionario de los números que en el corpus aparezcan.

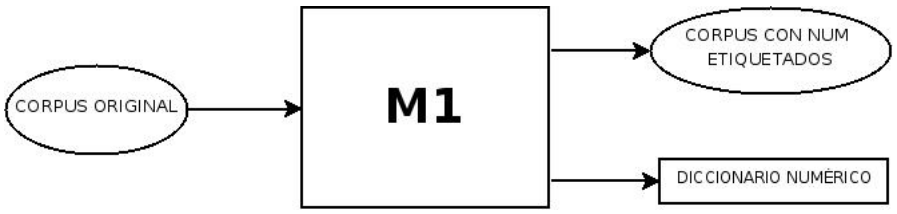

Figura 3.1: *M1: Preprocesado del corpus bilingüe.*

Esta herramienta utiliza el corpus bilingüe original, analiza frase a frase todo el corpus y sustituye cada uno de los números encontrados por una etiqueta. En nuestro caso hemos determinado sustituir los número cuya presencia en una frase sea única, es decir si sólo aparece un único número en la frase por **@NUM**. Mientras que en el caso en el que encontremos más de un número por frase la etiqueta que utilizaremos será @NUMX siendo X el ordinal que identifica a cada número (i.e @NUM1, @NUM2 ..... , @NUMX). Utilizaremos esta notación puesto que realizaremos un nuevo alineamiento de los textos y de esta manera no aseguramos que los números etiquetados quedan alineados de manera correcta. Si utilizáramos la etiqueta @NUM, perderíamos la información del alineamiento.

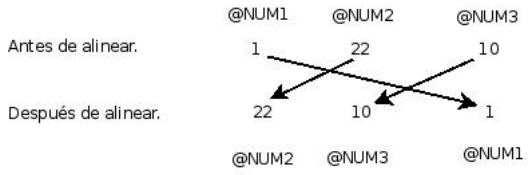

Figura 3.2: *Alineado utilizando etiquetas @NUM*

A continuación mostramos unos ejemplos de como funciona el proceso. Utilizamos para este ejemplo un corpus bilingüe castellano-catalán.

Corpus bilingüe original

SRC 1: *El índice de puntualidad cayó por debajo de el 46 % .*
SRC 2: *Las dificultades de navegación obligaron a cancelar 40 vuelos , 17 de salida y 23 de llegada , de los que 11 pertenecían a Iberia.*

Figura 3.3: ***Frase del lenguaje origen***

TRG 1: *L'índex de puntualitat va caure per sota de el 46 % .*
TRG 2: *Les dificultats de navegació van obligar a cancel·lar 40 vols , 17 de sortida i 23 d´ arribada , 11 de els quals pertanyien a Iberia.*

Figura 3.4: ***Frase del lenguaje destino***

Corpus bilingüe etiquetado:

SRC 1: *El índice de puntualidad cayó por debajo de el @NUM % .*
SRC 2: *Las dificultades de navegación obligaron a cancelar @NUM1 vuelos , @NUM2 de salida y @NUM3 de llegada ,de los que @NUM4 pertenecían a Iberia.*

Figura 3.5: ***Frase del lenguaje origen etiquetada***

TRG 1: *L'índex de puntualitat va caure per sota de el @NUM % .*
TRG 2: *Les dificultats de navegació van obligar a cancel·lar @NUM1 vols , @NUM2 de sortida i @NUM3 dárribada , @NUM4 de els quals pertanyien a Iberia.*

Figura 3.6: ***Frase del lenguaje destino etiquetada***

Por otro lado el denominado diccionario de traducción numérico consiste en una lista, que en el caso de este corpus castellano catalán tendrá el siguiente aspecto:

Lista de números

69_millones#69_milions

1998#1998

0,6#0,6

22_millones#22_milions

cuatro#quatre

061#061

dos#dues

7.000#7.000

ocho#vuit

22#22

cuatro#quatre

3.1.2. Preprocesado del texto a traducir.

El preprocesado que realizaremos del texto que deseamos traducir es considerablemente diferente, al que hemos utilizado para el corpus bilingüe. En primer lugar en este caso debemos tratar un único texto y no un par de textos paralelos, y en segundo lugar no podemos utilizar esta herramienta de preprocesado para generar ningún diccionario numérico, es precisamente el desconocimiento de los números que aparecen, el escollo que pretendemos salvar.

El formato del texto saliente es análogo al del corpus bilingüe etiquetado, es decir sustituimos en número por @NUM si es el único en la frase, y en el caso de que haya mas de uno utilizaremos @NUM1, @NUM2 ...

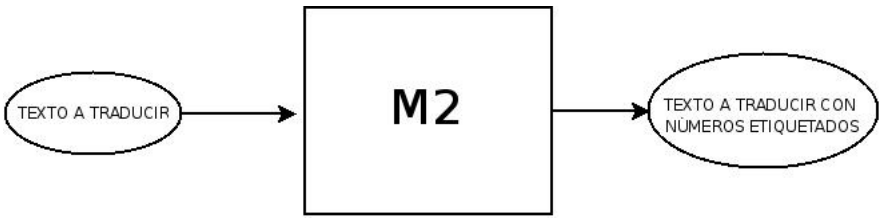

Figura 3.7: *M2: Preprocesado del texto que deseamos traducir.*

3.1.3. Postprocesado del texto traducido.

Una vez hemos preprocesado el corpus bilingüe, construimos nuestro baseline con el nuevo corpus, en el que hemos cambiado los números por etiquetas. Generamos el nuevo modelo de traducción bilingüe, utilizamos el decodificador

para obtener la traducción de nuestro texto etiquetado a traducir. La traducción obtenida arrastra las etiquetas utilizadas en este proceso, por lo tanto debemos tratar este texto para recuperar los números originales.

Para ello utilizaremos la lista o diccionario numérico que hemos obtenido en el preprocesado del corpus bilingüe.

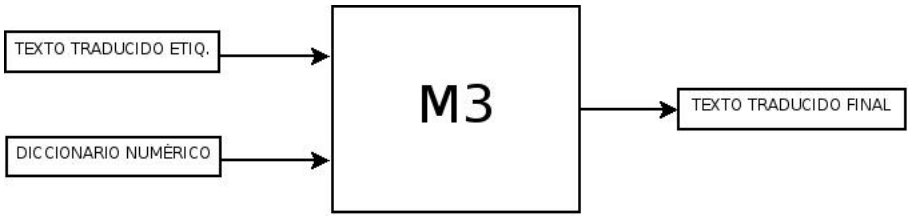

Figura 3.8: *M3: Postprocesado del texto traducido.*

Tras esta explicación teórica del proceso para la categorización de los textos en cuanto a su contenido numérico se refiere, procedemos a su implementación y evaluación, con el fin de observar las posibles mejoras que este procesado puede aportar a nuestro sistema de traducción.

3.2. Experimentos de la categorización numérica.

En el momento de implementar nuestras herramientas de preprocesado, nuestro deseo hubiera sido que estas fueran genéricas para cualquier corpus bilingüe, debido a las particularidades de cada lenguaje optamos por hacer un tratamiento específico en cada par de lenguajes.

Debemos realizar una detección de los números diferente para cada idioma, es por ello que nuestra herramienta de preprocesado está particularizada para cada corpus bilingüe.

Vamos a realizar experimentos con dos corpus de entrenamiento, uno que formado por los idiomas Chino-Inglés y otro formado por los idiomas Castellano-Catalán.

3.2.1. Corpus Chino-Inglés.

Preprocesado.

El tratamiento de los textos chinos con el fin de identificar los números que en ellos aparecen, es considerablemente más complejo que en otros idiomas que nos son más comunes. Tras un elaborado estudio hemos determinado ciertas reglas y modificasiones que nos han permitido automatizar el proceso de detección numérico.

Uno de los problemas con los que nos encontramos fue que en algunos casos en el idioma chino aparecen algunos números en diferentes palabras, dificultando notablemente su detección, eso nos obligó a hacer un primer análisis y modificación del texto, buscando los números que respondieran a ese comportamiento y uniéndolos en una sola palabra, eso nos permitía su posterior etiquetado.

Un ejemplo de lo anteriormente citado es el siguiente:

-frase original : "我们 的 营业 时间 是 从 十 九点 到 二 十 四 点"

donde: 二 = dos 十 = diez 四 = cuatro

-frase compactada : "我们 的 营业 时间 是 从 十 九点 到 二十四 点"

donde 二十四 = veinticuatro

La interpretación real nos la da la frase compactada, de no haber realizado este proceso de fusión la traducción propuesta por un sistema automático sería totalmente incorrecta, ya que interpretaría cada número por separado. No podemos unir una serie de números solo por el hecho de que están próximos unos a otros, puesto que en muchos casos realmente son números diferentes, en realidad solo aplicamos la fusión de palabras cuando alguno de los números implicados es una decena, centena, millar.

Generación de números.

Una vez hemos preprocesado el texto, y hemos conseguido que un número esté constituido por una única palabra, debemos utilizar alguna herramienta para detectar los números dentro de un texto, posteriormente categorizarlos y etiquetarlos. Como es de suponer necesitamos también, algún método que nos permita identificar un número en chino con su equivalente inglés.

Esta tarea es muy complicada para nosotros, debido al desconocimiento que tenemos del lenguaje chino. Con el fin de salvar este escollo nos pusimos en contacto con Erik Peterson, un programador y lingüista de origen chino que desarrolla su trabajo de investigación en el Instituto de Tecnologías del Lenguaje en la Universidad Carnegie Mellon. Podemos consultar su trabajo en *http://www.mandarintools.com/*.

Como parte de su investigación ha desarrollado un gran número de herramientas tanto on-line como off-line, diseñadas para ayudar a la comprensión y el aprendizaje del Chino. Nos proporcionó un autómata que detecta números en chino y otro que dado un número en chino genera el correspondiente en inglés. Integramos estos autómatas en nuestro sistema y construimos varias aplicaciones para la detección y traducción de números.

Resultados.

Aplicando las herramientas anteriormente descritas a nuestro corpus de entrenamiento, es decir preprocesamos los textos en chino del corpus de entrenamiento, realizamos un realineado del mimo, puesto que el número de palabras del texto chino ha cambiado, y por lo tanto la equivalencia con palabras del texto en inglés también habrá cambiado. A continuación mostramos un ejemplo de como se vería modificado el alineado una vez se ha preprocesado el texto en chino:

SRC: 可以 的 大约 有 四 十 分钟	*Yes . It'll take about forty minutes .*
TRG: 可以 的 大约 有 四十 分钟	*Yes . It'll take about forty minutes .*
ALIGN: *1-1 2-2 3-3 4-4 5-5 6-5 7-6*	*1-1 2-2 3-3 4-4 5-5 6-6 7-7 8-8*

Figura 3.9: ***Ejemplo de modificación de alineado debido al preprocesado***

Vemos como se ve modificado el número de palabras tras el preprocesado, en origen teníamos 7 palabras en chino que se correspondían a una frase en inglés de 8 palabras, pero tras el preprocesado tenemos una frase en chino de 6 palabras que se corresponde con la misma frase en inglés, puesto que esta no se ve alterada por el preprocesado. Así debemos realizar un realineado del texto para tener una correspondencia adecuada.

Así pues tras etiquetar todos los números que aparecen en el corpus de entrenamiento y realizar el alineado procedemos a la construcción de nuestro sistema de traducción. Esto es lo que hemos denominado anteriormente como *Preprocesado del corpus bilingüe - M1* (3.1.1)

SRC: 哦 那个 航班 是 C 三 零 六
- Sure . It is flight C three zero six .
TRG: 哦 那个 航班 是 C *@NUM1 @NUM2 @NUM3*
- Sure . It is flight C @NUM1 @NUM2 @NUM3 .

Figura 3.10: ***Resultado del preprocesado y etiquetado***

Preprocesamos y etiquetamos de igual manera el texto que deseamos traducir, no importa que varíe el número de palabras, puesto que en este caso solo tenemos un texto fuente, no hay alineado que se vea modificado. Esto será el *Preprocesado del texto que deseamos traducir - M2* (3.1.2)

Procedemos a la traducción del texto utilizando el modelo de traducción que hemos creado con el baseline etiquetado.

SRC: ..水深 大约 在 三 米
SRC': ..水深 大约 在 *@NUM* 米
UNITS: 水深*#unk[2,0]* 在*#at[3,2]* 大约*#about[2,1] @NUM#@NUM[2,3]*
米*#meters[3,4]*
TRG: *at about @NUM meters*

Figura 3.11: ***Traducción del texto etiquetado***

Seguidamente utilizamos la herramienta que hemos implementado para recuperar el significado del texto etiquetado. Es decir, en el texto traducido tenemos todos los números etiquetados como @NUM o @NUMx y debemos recuperar los números que se esconden tras estas etiquetas, para ello utilizaremos el generador de números implementado.

En el caso de que aparezca un número desconocido, es decir que no encontremos su equivalente en el listado de números generado, nuestra herramiente procede a la generación de una traducción de dicho número. Este es un gran avance respecto a un método que únicamente se base en un aprendizaje a partir de un corpus de entrenamiento.

SRC: *at about @NUM meters*
TRG: *at about three meters*

Figura 3.12: *Generación de números*

A continuación mostramos los resultados de la evaluación, del texto tra-

ducido sin utilizar nuestro sistema de categorización y el resultado obtenido al utilizarlo.

System	BLEU	NIST	mWER	mPER
Zh2En				
NB	19.04	5.97	64.77	49.57
+CATEG	19.43	6.12	64.49	49.30

Cuadro 3.1: *Resultados para el test Zh2En .*

Podemos observar que existe una mejora considerable en la evaluación del texto traducido, utilizando nuestro sistema, respecto al texto traducido sin utilizarlo.

Análisis de los resultados.

Si analizamos los dos textos, tanto el que hemos traducido sin utilizar la categorización, como el que hemos traducido siguiendo todo el proceso que hemos explicado anteriormente, vemos en que radican las diferencias que hacen que los resultados de la evaluación sea mejores.

Por un lado el preprocesado es fundamental para obtener mejores resultados, puesto que el valor del los números es completamente diferente en un caso y en otro.

Si no utilizamos la categorización encontramos errores del tipo:

1.- Errores de traducción. 请 问 你 能 再 说 一 遍 吗 ## could you say it again please
En chino aparece el número 1 como traducción de 一, pero en este contexto no es su significado real.

2.- Varias maneras de escribir cero: 她 想 要 个 啤酒 *## she wants a beer*

(个 es zero)

我 想 订 两 个 双人 标准间 *## i 'd like to reserve two twin rooms*

(个 es zero)

北美 的 营业额 增长 多 于 百分之 五 吗 *## did the turnover in north america rise by more than five percent*

(百分之 es traducido como zero cuando en realidad no es su significado)

3.- Números consecutivos pero en palabras diferentes.

好 呃 一 二 三 二 十 分 硬币 呃 一 个 五十分 的 硬币 和 两 个 五分 的 硬币 啊 我 有 了 *## well uh one two three twenty cent coins and uh one fifty cent coin and two five cent coins ah i have it*

please change this one hundred dollar into one fifty dollar bill and five ten dollar bills please ## 请 把 这 一 百 美元 换成 一 张 五 十 美元 的 钞票 和 五 张 十 美元 的 钞票

one hundred – 一 百 es 100 y no 1 100 que sería la traducción que nos proporcionaría el sistema original.

一 千 二 百 美元 的 旅行 支票 三 百 美元 和 五 万 日元 的 现金 *## twelve hundred dollars in traveler 's checks three hundred dollars and fifty thousand yen in cash* 一千二百 es 1200 pero se traducía como 一 千 二 百 es 1 1000 2 100.

3.2.2. Corpus Castellano-Catalán.

- En el caso del corpus bilingüe castellano-catalán es muy diferente al anteriormente descrito, puesto que tenemos mucha más información que relaciona ambos idiomas, de hecho disponemos de una categorización que identifica que palabras son números, dicha categorización nos la proporciona una herramienta denominada FreeLing [Atserias, Casas, Comelles, González, Padró, and Padró.], con lo cual el etiquetado es mucho más sencillo. También la extracción de un diccionario numérico resulta más sencilla que con otro corpus, puesto que además de tener identificado cada número en ambos textos, disponemos del alineado de las palabras de los dos textos.

Identificación de Números.

Para poder substituir los números por etiquetas, puesto que este es nuestro objetivo, debemos encontrar algún método para detectar los números. En ciertos casos como este mismo, el de Castellano - Catalán, disponemos de alguna herramienta que nos facilita la identificación

Disponemos de un tageador [Atserias et˜al.] , cuya funcionalidad es identificar diferentes tipos de palabras, ese decir nombres, adjetivos,etc..., de tal manera que cada palabra queda clasificada según un identificador.
Una frase en castellano quedaría tageada de la siguiente manera.

SRC: *CATALUNYA DESPIDIÓ 1998 SIN TRAUMAS .*
SRC': *RG VMIS3S0 Z SPS00 NCMP000 Fp*

Figura 3.13: ***Tageado de una frase en castellano***

Donde quedan clasificados los siguientes elementos

Nombre	CATALUNYA	RG
Verbo	DESPIDIÓ	VMIS3S0
Número	1998	Z
Preposición	SIN	SPS00
Nombre	TRAUMAS	NCMP000
Puntuación	.	Fp

Figura 3.14: ***Equivalencias***

Una frase en castellano como *CATALUNYA DESPIDIÓ 1998 SIN TRAUMAS .* Quedaría tageada como *RG VMIS3S0 Z SPS00 NCMP000 Fp*

Como podemos observar, los números quedan tageados o etiquetados con la letra **Z**, utilizando esta información podremos identificar los números en nuestro texto y substituirlos, en este caso por @NUM.

Traducción de Números.

Por otro lado el hecho de disponer conocimientos de ambos idiomas, la similitud entre ellos y disponer de datos previos para la implementación de nuestro sistema, que por supuesto ya han sido utilizados en la creación del sistema de traducción, como puede ser la categorización, nos lleva a suponer que las mejoras que podamos conseguir no serán extraordinarias.

Para la traducción de números, es decir para substituir las etiquetas que hemos colocado antes de realizar la traducción de nuestros textos de evaluación, utilizaremos tablas que relacionan los números en castellano con números en catalán.
Estas tablas las hemos extraído del análisis del corpus de entrenamiento, al ser unos textos bastante extensos nos permiten obtener una importante tabla de equivalencias. Para extraer dichas tablas utilizaremos los archivos tageados, que por cierto tienen extensión *.POS* y el conocimiento del alineado de textos paraleos, ello será suficiente para extraer una tabla con todos los números que el tageador haya sido capaz de tagear.
El tageador tiene un porcentaje de acierto en la detección de elementos superior al 95 %.

Resultados.

Hemos utilizado un texto para evaluar el sistema original, sin la introducción de la categorización. Posteriormente evaluamos el mismo texto con el sistema una vez hemos introducido la categorización y evaluamos ambos sistema.

Aplicamos la categorización al corpus de entrenamiento, a continuación mostramos como sería la evolución de una frase en los diferentes pasos.

-SRC: *Ella , cada cinco minutos , le acariciaba y le decía çariño ".*
-TAG *PP3FS000 Fc DI0CS0 Z NCMP000 Fc PP3CSD00 VMII3S0 CC PP3CSD00 VMII3S0 Fe NCMS000 Fe Fp*
-TRG *Ella , cada @NUM minutos , le acariciaba y le decía çariño ".*

Figura 3.15: ***Tageado del corpus de entrenamiento***

System	BLEU	NIST	mWER	mPER
Es2Ca				
NB	84.02.	14.03	11.87	8.37
+CATEG	83.23	13.91	10.77	8.20

Cuadro 3.2: *Resultados para el test Es2Ca .*

Análisis de los resultados.

Como podemos ver en la figura 3.16, el hecho de utilizar la categorización empeora los resultados de la traducción, en este caso podemos ver claramente como perdemos información acerca del género. La técnica de categorización no tiene en cuenta que algunos lenguajes dotan de género a algunos de sus números y es por ello que al utilizarlo perdemos dicha información.

-Referencia: *'SIN NOTICIAS DE DIOS ÉNFRONTA DUES ESTRELLES .*
-TRG Sin CAT. *'SIN NOTICIAS DE DIOS ÉNFRONTA DUES ESTRELLES .*
-TRG Con CAT *'SIN NOTICIAS DE DIOS ÉNFRONTA DOS ESTRELLES .*

Figura 3.16: ***Ejemplo traducción con y sin categorización***

Como podemos observar en este caso los resultados obtenidos utilizando nuestra técnica de sustitución de números son peores que los originales, como vemos el BLEU baja en casi un punto. Esto es debido a la ambigüedad que aparece en determinados números y que nuestro sistema tal y como a sido concebido no puede detectar.

Un ejemplo claro de dicha ambigüedad es la existencia de género en los números de determinados idiomas, por ejemplo el catellano y el catalán tiene números masculinos y femeninos, cosa que ocurre con otros idiomas como el chino o el inglés.Este es un motivo claro de porque en un corpus nuestro sistema funciona eficientemente, no siendo así en el corpus Castellano-Catalán.

3.3. Tratamiento de Páginas Web.

El proceso a seguir para la categorización de las páginas webs va a ser similar al que hemos realizado con los números, vamos a analizar los textos con el fin de identificar y etiquetar todas las páginas web que en ellos aparezcan. Como es lógico el proceso será completamente diferente, aunque el resultado será similar, la idea es tratar nuestro corpus bilingüe original y modificarlo de tal manera que obtengamos un nuevo corpus en le que hayamos sustituido las páginas web por una etiqueta definida por nosotros. En este caso la etiqueta elegida para determinar una página web será **@WEB**.

Así pues una frase tal que :

The Code is also accessible through the HKMA internet website (http://www.info.gov.hk/hkma) .

Debe tratarse para que quede de la forma:

The Code is also accessible through the HKMA internet website (@WEB) .

Al igual que en la categorización realizada para los números, el sistema que implementamos consta de una fase de preprocesado y otra de postprocesado. En este caso no es necesario realizar una proceso de traducción de los elementos etiquetados, puesto que las páginas web no requieren de traducción.

3.3.1. Preprocesado del texto a traducir.

En la primera fase, tratamos tanto el corpus bilingüe como el texto a traducir, realizamos un análisis de los textos con el fin de detectar el mayor número posible de páginas web.

Al efectuar el análisis nos percatamos del gran número de webs que aparecen fraccionadas en diferentes palabras, nos vemos ante un problema similar al acaecido en el análisis numérico del corpus Chino-Inglés, esto nos obliga a unir una serie de palabras, que serán las que consideramos que forman parte de un misma dirección de página web.

Así pues aparecen una serie de palabras como por ejemplo:

51 . 迄今为止 , 已 有 26 个 国家 的 行动 计划 可 在 万 维 网 (*http: //www.un.org/womwnwatch/list.h*) 上 查 到 , 供 广大 公众 查阅 。

Esta frase la debemos tratar para que quede de la forma:

51 . 迄今为止 , 已 有 26 个 国家 的 行动 计划 可 在 万 维 网 (*http://www.un.org/womwnwatch/list.h*) 上 查 到 , 供 广大 公众 查阅 。

Una vez realizada esta unión, realizamos el etiquetado de la misma, obteniendo como resultado la frase siguiente:

51 . 迄今为止 , 已 有 26 个 国家 的 行动 计划 可 在 万 维 网 (*@WEB*) 上 查 到 , 供 广大 公众 查阅 。

Una vez realizado el análisis y etiquetado del corpus bilingüe, debemos percatarnos de que hemos cambiado el número de palabras por frase, por lo tanto, el alineado de frases paralelas que disponíamos inicialmente ha dejado de sernos útil, puesto que no es real.

Debemos efectuar un alineado de los nuevos textos paralelos, para así poder construir un nuevo baseline. Una vez realizado construimos nuestro sistema de traducción de nuevo.

Aplicamos el mismo criterio para el texto que deseamos traducir, en este caso guardamos un archivo con las páginas webs detectadas y la posición que ocupan en el texto.

Realizamos la traducción de nuestro texto, en el que hemos sustituido las páginas web por la etiqueta @WEB.

3.3.2. Postprocesado del texto a traducir.

En la fase de postprocesado, tratamos el texto traducido con el fin de recuperar el contenido lo más fielmente posible. Así pues en esta fase buscamos las etiquetas del texto y junto con el archivo extraído del preprocesado donde guardamos las páginas web y su posición, realizamos la recuperación del valor original de las mismas.

3.4. Experimentos de la categorización de páginas web.

Una vez realizada la explicación teórica del proceso que vamos a seguir para categorizar las página webs y poder así mejorar nuestro sistema de traducción, nos disponemos a realizar algunos experimentos para confirmar que nuestra mejora del sistema da realmente sus frutos.

Para ello utilizamos los textos habituales que nos provee por ejemplo el NIST (Test Oficial MT EVAL, NIST 2002, 2003 ,20004), pero analizando estos textos observamos que la presencia de webs en sus líneas es francamente escasa o inexistente. Son textos fundamentalmente compuestos por noticias y en este entorno, actualmente, todavía la presencia de web es escasa.

Hemos analizado tests de los últimos tres años y observamos que la presencia de páginas webs está entorno a la cinco webs por cada millar de frases.

Este hecho representa un problema para nosotros, puesto que aunque el sistema mejore, la casi inexistencia de webs en nuestros tests haría inapreciable cuantitativamente esta mejora. Para solventar este problema optamos por crear nuestro propio test. Seleccionamos líneas que contenían páginas web, estas fueron extraidas de los textos de las naciones unidas (UN) LDC2004E12-UN_CE_parallel_text y LDC2004T08HK_parallel_text.

Así pues construimos un texto en el que aparece al menos una web por linea:

1.- 科委 将 备 有 附件 一 所 列 文件 。 除 通常 的 分发 对象 以外 , 还 将 在 互联网络 上 的 地址 (http : //www.unccd.ch) 向 秘书处 的 万 维 网 提供 文件 。
2.- 亚洲 - 太平洋 人口 信息网 通过 互联网络 广泛 提供 资料 服务 和 产品 , 网址 为 : http: /www.un.org/depts/escap/pop/welcome.htm 。
3.- 51 . 迄今为止 , 已 有 26 个 国家 的 行动 计划 可 在 万 维 网 (http: //www.un.org/womwnwatch/list.htm) 上 查 到 , 供 广大 公众 查阅 。
4.- [15] China adopts 14001 as State Policy , China Begins Third Phase of ISO 14001 Program , 1997 , globeNet , http : //www.iso14000.net , Global Environment & Technology Foundation , Annandale , VA , United States 。

En la tabla podemos observar los resultados obtenidos al traducir el texto de referencia, sin utilizar la categorización y utilizando nuestra técnica de cate-

gorización de páginas web.

Tipo	Sin categorización	Con categorización
BLEU score	6.11	33.3
NIST score	2.83	5.11
PER score	90.34	59.98
WER score	111.35	76.02

Cuadro 3.3: *Resultados obtenidos con la categorización de webs.*

Como podemos observar los resultados obtenidos utilizando categorización son ampliamente superiores a los obtenidos sin usarla. En este caso los valores del BLEU originales son muy bajos, ello es debo por un lado a que los resultados del sistema de traducción aplicado al Chino-Inglés no produce tan buenos resultados, y por otro lado debemos tener en cuenta que el texto utilizado ha construido para este experimento con un gran número de webs por línea, con el fin de demostrar la obtención no tanto de buenos resultados globales, sino la mejora de resultados respecto a la traducción sin categorización.

De hecho podemos ver que el BLEU es cinco veces superior por el hecho de utilizar nuestra técnica de categorización. Es de esperar que esta mejora no sea tan elevada en otros textos, ya que la mejora siempre será proporcional al número de webs que aparezcan en el texto.

En la figura 3.17 vemos diferentes ejemplos de los resultados de la categorización de páginas web.

-Referencia: *The Government of Canada (http://canada.gc.ca) .*
-TRG Sin CAT. *canadian government (www.info.gov.hk /) .*
-TRG Con CAT *the canadian government (http://canada.gc.ca).*

-Referencia: *17 The Internet address of DSBB is http://dsbb.imf.org .*
-TRG Sin CAT. *17 media standard bulletin board : www.info.gov.hk / .*
-TRG Con CAT *17 media standards as bulletin board : http://dsbb.inf.org .*

Figura 3.17: ***Ejemplo traducción con y sin categorización***

Como podemos ver en los ejemplos de traducción, el sistema sin categorización no traduce correctamente la página web. El sistema sin categorización no hace un tratamiento especial de las páginas web, eso lo hace susceptible a la aparición de este tipo de errores.

3.5. Conclusiones

Analizando globalmente los resultados obtenidos mediante las técnicas de categorización planteadas, podemos concluir que la categorización de páginas web es siempre positiva, mientras que las técnicas aplicadas para la categorización numérica serán efectivas en función de los idiomas que queramos traducir y de las herramientas que podamos implementar o utilizar para atacar la detección y generación de números, puesto que las herramientas de categorización de números deben ser específicas para cada par de idiomas.

En los experimentos realizados para el corpus de Chino-Inglés en cuanto a categorización numérica se refiere, hemos obtenido una mejora considerable en el BLEU pasando de un 19.04 a un 19.43. No ocurre lo mismo con el corpus Castellano-Catalán, en el que los resultados han sido inferiores a los originales.

En cuanto a la categorización de páginas web, la mejora de los resultados ha sido notable, obteniendo resultados cinco veces mejores del BLEU, en los experimentos realizados. No podemos extrapolar esta mejora a cualquier tipo de texto, puesto que la utilidad de nuestra técnica irá directamente vinculada al número de webs que aparezcan en los textos. A mayor número de webs mayor será la contribución de esta técnica a una correcta traducción

Capítulo 4

Conclusiones

4.1. Filtrado estadístico.

En general, 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.

Se han presentado dos técnicas de filtrado estadístico de corpus bilingües. Dichas técnicas se basan en el Modelo IBM1 y en el PER.

Ahora bien, la utilización de una u otra técnica, o incluso la no utilización de ellas, dependerá de varios factores.

En primer lugar debemos tener en cuenta que la eficacia de las técnicas de filtrado propuestas, dependen directamente de la calidad del corpus de entrenamiento tratado. Refiriéndonos a esa calidad como a la buena alineación de frases paralelas. Es decir un corpus de entrenamiento con una muy buena correspondencia de traducción de frases paralelas no requiere de la utilización de estas técnicas, pudiendo ser incluso perniciosas, puesto que llegaríamos a eliminar frases cuya aportación es positiva para el sistema.

Por otro lado la técnica basada en el PER requiere un mayor coste computacional que la basada en el Modelo IBM1, es por eso que incluso en corpus donde los resultados sean mejores utilizando esta técnica deberemos tener en cuenta la disponibilidad de recursos y de tiempo.

En general, la técnica basada en el Modelo IBM1 obtiene resultados ligeramente superiores a la técnica basada en el PER.

En los experimentos que hemos realizados en el corpus Chino-Inglés ambas técnicas mejoran ampliamente los resultados originales, en este caso los resultados obtenidos mediante la técnica PER mejoran levemente los obtenidos por el Modelo IBM1, pero los recursos empleados son muy superiores, así pues queda a nuestra elección la utilización de una u otra en función de nuestras necesidades.

Por otro lado en los experimentos realizados con el corpus Español-Inglés se mejora en más de 1 punto BLEU. Mientras que en los experimentos realizados para el corpus Castellano-Catalán, los resultados obtenidos son inferiores al original, con lo cual podemos determinar que es un corpus con una buena correspondencia de traducción de frases paralelas.

4.2. Categorización de números.

Hemos generado diferentes herramientas para llevar a cabo la detección y categorización de números en diferentes idiomas. Tras llevar a cabo varios experimentos en diferentes corpus hemos podido comprobar la importancia de crear técnicas para cada idioma, específicas para dicha categorización.

El gran handicap que se nos presenta a la hora de utilizar técnicas eficaces para la categorización numérica, es el hecho de que hemos creado herramientas particulares para cada idioma, no pudiendo generar una herramienta genérica que aplicar a cualquier idioma. Tal y como lo hemos planteado se requiere un conocimiento considerable de los idiomas que componen el corpus para poder implementar herramientas que nos faciliten la detección y traducción numérica.

En los experimentos realizados hemos obtenido resultados diversos, en función del corpus que hemos tratado. Para el corpus de Chino-Inglés hemos obtenido una mejora considerable en el BLEU pasando de un 19.04 a un 19.43. Es decir para este par de lenguajes la técnica implementada es muy útil, la mejora de los resultados ha sido notable.

No ocurre lo mismo para el corpus Castellano-Catalán donde los resultados obtenidos utilizando nuestra técnica de sustitución de números son peores que los originales, en este caso el BLEU baja en casi un punto. Esto es debido a la ambigüedad que aparece en determinados números y que nuestro sistema tal y como a sido concebido no puede detectar.

Un ejemplo claro de dicha ambigüedad es la existencia de género en los números de determinados idiomas, por ejemplo el catellano y el catalán tiene números masculinos y femeninos, cosa que ocurre con otros idiomas como el chino o el inglés.Este es un motivo claro de porque en un corpus nuestro sistema funciona eficientemente, no siendo así en el corpus Castellano-Catalán.

4.3. Categorización páginas web.

En la actualidad podemos observar la cada vez mas abundante presencia de páginas web en cualquiera de los escritos con los que estamos acostumbrados a tratar, no sólo de los relacionados con temas técnicos, en los que parece ser más habitual, sino en cualquier tipo de temas.

Así pues podemos asegurar que la implementación de esta herramienta siendo muy útil en la actualidad, lo será mucho más en un futuro muy cercano, dotando de mucha mas calidad a los textos traducidos.

Este hecho queda demostrado analizando los resultados obtenidos en nuestros experimentos. La mejora de los resultados ha sido notable, obteniendo resultados cinco veces mejores del BLEU, en los experimentos realizados, pasando de 6.11 a 33.3.

Si bien es cierto que no podemos extrapolar esta mejora a cualquier tipo de texto, puesto que la utilidad de nuestra técnica irá directamente vinculada al número de webs que aparezcan en los textos. A mayor número de webs mayor será la contribución de esta técnica a una correcta traducción. En nuestro experimento hemos utilizado un texto done aparece al menos una web en cada línea, con lo cual la mejora del BLEU es muy amplia.

Bibliografía

Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA.*

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

Crego, J. M., Costa-jussà, M. R., Mariño, J., and Fonollosa, J. A. Ngram-based versus phrase-based statistical machine translation. pages 177–184, October 2005.

A. de Gispert and J. Mariño. Using X-grams for speech-to-speech translation. September 2002.

Josep M. Crego Adrià de Gispert Patrick Lambert José A. R. Fonollosa Marta R. Costa-Jussà José B. Marino, Rafael E. Banchs. N-gram-based machine translation. *Computational Linguistics, Association for Computacional Linguistics.*, 32(4):527–549, 2006.

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

P. Koehn. A beam search decoder for phrase-based statistical machine translation models. technical manual of the pharaoh decoder. 2003.

P. Koehn, F.J. Och, and D. Marcu. Statistical phrase-based translation. pages 48–54, Edmonton, Canada, May 2003.

Mariño, J.B., Banchs, R.E., Crego, J.M., de Gispert, A., Lambert, P., Fonollosa, J.A.R., and Costa-jussà, M.R. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549, December 2006.

Och, F.J. Minimum error rate training in statistical machine translation. 2003.

Och, F.J. and Ney, H. A systematic comparison of various statistical alignment models. 29(1):19–51, March 2003.

Och, F.J. and Ney,H. Discriminative training and maximum entropy models for statistical machine translation. pages 295–302, Philadelphia, USA, July 2002.

A. Stolcke. Srilm - an extensible language modeling toolkit. September 2002.

E. Vidal. Finite-state speech-to-speech translation. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, 1:111–114, 1997.

H. Ney. Zens, R., F.J. Och. Improvements in phrase-based statistical. *Proceedings of Australasian Language Technology Workshop machine translation. Proc. of the Human Language Technology Conference,*, 2004.