Sistema de traducción de lenguaje SMS a castellano

V. López, R. San-Segundo, R. Martín, J.D. Echeverry, S. Lutfi
Grupo de Tecnología del Habla - Universidad Politécnica de Madrid
veronicalopez@die.upm.es

***Resumen* — En este artículo se describe el proceso llevado a cabo para desarrollar un sistema de traducción de lenguaje SMS (Short Message Service) a castellano. En primer lugar, se genera una base de datos necesaria para desarrollar el sistema, formada por más de 11000 términos y expresiones en lenguaje SMS y sus traducciones al castellano, así como frases de ejemplo en lenguaje SMS para realizar una primera evaluación del sistema. La arquitectura completa está formada por un traductor automático estadístico basado en subfrases o secuencias de palabras y una serie de funciones implementadas para actuar sobre las frases en tiempo real. La evaluación de la arquitectura se realiza con las siguientes métricas: WER (tasa de error de palabras), BLEU ("BiLingual Evaluation Understudy") y NIST. Como resultado final, se obtiene una tasa de error de palabra de 20,2% para el mejor experimento.**

I. INTRODUCCIÓN

El uso del lenguaje SMS (Short Message Service) se propagó con el auge de la mensajería instantánea y el servicio de mensajes cortos por teléfono. Desde el punto de vista de la teoría de la comunicación, el lenguaje SMS es una codificación adicional del mensaje en el propio idioma. Su rápida propagación se debe a la necesidad, cada vez mayor, de minimizar el coste de la comunicación economizando el lenguaje.

Este lenguaje no es universal, contando cada lengua con su conjunto de reglas en función de las abreviaciones posibles y de la fonética propia de cada idioma. Pero, en general, se caracteriza por abreviar las palabras en relación con la fonética de cada lengua y el significado de las mismas, eliminando tildes y palabras que se sobreentienden según el contexto, eliminando letras como la '*h*', eliminando los signos de puntuación, incluyendo emoticonos, etc.

Como consecuencia de su rápida expansión, surge la necesidad de desarrollar sistemas de traducción del lenguaje SMS a voz. Estos sistemas pueden ser útiles para mandar mensajes SMS a teléfonos fijos, siendo numerosas las aplicaciones de este servicio. Por ejemplo, pueden ser empleados en situaciones de emergencia, para mandar SMS a ancianos que no están familiarizados con el lenguaje SMS, para enviar mensajes a personas con problemas de visión o a personas que están conduciendo.

II. ESTADO DE LA CUESTIÓN

Existen distintos tipos de software que se encargan de traducir mensajes SMS a voz. Sin embargo, la gran mayoría de ellos consiste en un diccionario de términos SMS que va traduciendo las palabras del mensaje una por una, sin atender a un modelo de lenguaje ni al contexto en que se sitúan dichas palabras. Como ejemplo de estos sistemas está el servicio Voz SMS, desarrollado por Esendex [1], cuyo software convierte el mensaje en un mensaje de voz, y a continuación, se envía a un número de teléfono móvil o fijo. Comsys también ha desarrollado SMS to Fixed [2], que permite enviar mensajes en lenguaje SMS a teléfonos fijos. También CBOSSsms2voice de la compañía CBOSS [3] convierte mensajes en lenguaje SMS (en inglés) a voz.

Telefónica también tiene un servicio de mensajes de voz que permite que el móvil lea el mensaje SMS que se acaba de recibir. La entrega del mensaje se realiza mediante un envío procedente del número que ha dejado el mensaje. En este caso, el sistema de traducción es algo más complejo que un simple diccionario de términos, existiendo modelado del lenguaje.

Existe también un sistema desarrollado en el Institute of Infocomm Research de Singapur que normaliza el texto en lenguaje SMS antes de mandarlo a un traductor automático [4]. Cabe destacar también la página Web www.diccionariosms.com, desarrollada por la Asociación de Usuarios de Internet, que permite a los usuarios añadir entradas a un diccionario de términos (o expresiones) en lenguaje SMS y su traducción al castellano. En este artículo se describe el desarrollo de un sistema de traducción de lenguaje SMS al castellano, empleando para ello un traductor automático estadístico con un modelo de traducción y un modelo de lenguaje, así como funciones de preprocesado y postprocesado de las frases para su funcionamiento en tiempo real (Figura 1).

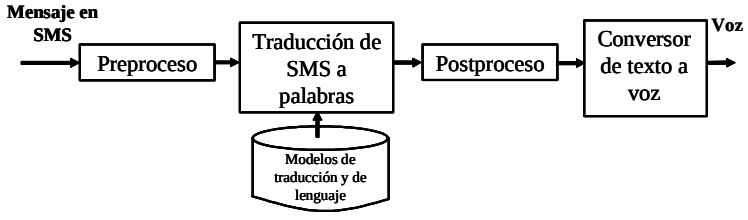

Figura 1. Arquitectura del sistema

III. BASE DE DATOS

Para obtener el corpus paralelo necesario para el entrenamiento del modelo de traducción, se utiliza, en primer lugar, el diccionario de términos extraído de la Web www.diccionariosms.com que ha sido realizado por usuarios de Internet.

Este diccionario cuenta con más de 11000 términos y expresiones en lenguaje SMS (aunque cada día aumenta su número), con sus correspondientes traducciones y un índice de popularidad en función del número de usuarios que ha registrado o están de acuerdo con ese par término-traducción. Es decir, cada término o expresión SMS, aparece con varias posibles traducciones. Por ejemplo, en la Figura 2, el término "*ma*" puede traducirse por "*madre*", "*Madrid*", "*mamá*", "*mañana*" o "*me ha*" y, según la popularidad, lo más usual es que se traduzca por "*mamá*" o "*me ha*".

	A	B	C
1	TÉRMINO SMS	SIGNIFICADO	POPULARIDAD
6046	**ma**	- madre	2
6047		- madrid	1
6048		- mamá	8
6049		- mañana	1
6050		- me ha	8

Figura 2. Fragmento de la base de datos de términos y expresiones SMS

Además de términos y expresiones, en la base de datos aparecen emoticonos, que son expresiones en ASCII que en su mayoría representan caras humanas con determinadas emociones, aunque también pueden tener significados muy diversos.

A partir de este corpus paralelo se generan los ficheros necesarios para entrenar el modelo de traducción. Así, se generan dos ficheros: uno con los términos en SMS y otro con cada una de sus traducciones en castellano. Además, para hacer uso de la popularidad como una medida de probabilidad de traducción, cada par SMS-castellano se repite en estos ficheros tantas veces como indique su popularidad.

En general, cuando se escribe un mensaje en lenguaje SMS, se pueden dejar algunas palabras en castellano sin ningún tipo de abreviatura, principalmente palabras cortas como "*no*" o "*sí*". Por ello, es necesario que en el entrenamiento de los modelos, los ficheros que contienen términos o expresiones en lenguaje SMS contengan también palabras en castellano, cuya traducción sea la misma palabra. Por tanto, se incorpora el vocabulario en castellano de la base de datos en todos los ficheros por duplicado para dotar de cierta probabilidad a la traducción de una palabra en castellano por la misma palabra.

Los términos en lenguaje SMS de todos los ficheros se tuvieron que convertir en expresiones con códigos ASCII, para evitar errores de los programas que pueden dar algunos caracteres especiales. El formato de estas expresiones es una '*e*' seguida de cada uno de los números asociados a los caracteres en el estándar ASCII. Estos números se separan por guiones. Por ejemplo:

"*la*" se transforma en "*e-108-97*"

Donde 108 es el código ASCII de la letra *l* y 97 el de la letra *a*.

Además de estos ficheros con términos y expresiones, se genera un fichero que contiene 58 frases en lenguaje SMS extraídas de páginas Web y otro fichero con las correspondientes traducciones en castellano. Algunos ejemplos de estas frases son: "*qndo akaba la peli?*", "*qtl tdo?*", "*compra 4 entrdas +*" o "*stoy n 1 atsco*". Con las 58 frases, se generan 8 listas de frases aleatorias en lenguaje SMS y las correspondientes traducciones en castellano, de manera que 7 de ellas se emplean para entrenar el modelo de lenguaje y ajustar los pesos de los modelos, y la restante para evaluar la arquitectura. Como comentaremos, las 8 listas van rotando con una estrategia de Round-Robin, para entrenar, ajustar y evaluar con todas ellas, calculando la media de los resultados (para obtener unos valores más significativos).

IV. ARQUITECTURA BASADA EN SUBFRASES

Para desarrollar el sistema de traducción de SMS a voz, se emplea, por un lado, un traductor automático estadístico basado en un modelo de subfrases o secuencias de palabras (Figura 1). Y, por otro lado, se implementan una serie de funciones para su funcionamiento en tiempo real.

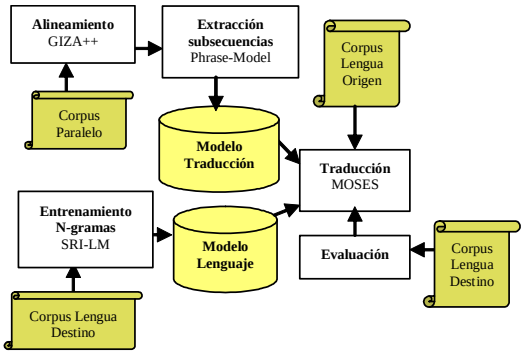

Figura 3. Arquitectura completa del sistema de traducción basado en subfrases

En la traducción automática estadística se aplica el Teorema de Bayes para calcular la probabilidad de que la cadena del idioma destino (d) haya sido generada por la cadena origen (o). De tal manera que para conseguir *p(d|o)* se calcula *p(o|d)·p(d)*, donde *p(o|d)* es la probabilidad de que la cadena origen sea la traducción de la cadena destino (modelo de traducción), y *p(d)* es la probabilidad de ver aquella cadena destino (modelo de lenguaje). El objetivo de la traducción es obtener la frase en destino que maximiza esta probabilidad.

A. Modelo de lenguaje

Para generar el modelo de lenguaje se emplea la herramienta *n-gram_count* de *SRI-LM [5]*, que realiza la estimación de los modelos de lenguaje tipo n-grama. Un modelo de n-grama determina la probabilidad de una palabra dadas las n-1 palabras previas.

B. Modelo de traducción

Para generar el modelo de traducción se emplean las siguientes herramientas: *GIZA++*, *Phrase_Extract* y *Phrase_Score*. Además, se necesita una colección de textos en lengua origen traducidos a lengua destino (corpus paralelo), para lo que se emplean los términos y expresiones de la base de datos en lenguaje SMS y su traducción al castellano (ficheros de entrenamiento). A partir del corpus paralelo, se obtiene el alineamiento con *GIZA++* ([6], [7]) en los dos sentidos: origen-destino y destino-origen. El alineamiento final, con el que se generará el modelo de traducción, puede obtenerse combinando los alineamientos anteriores de diferentes formas: teniendo en cuenta sólo el alineamiento en sentido destino-origen, sólo el alineamiento origen-destino, la intersección de los puntos de alineamiento, la unión, etc. Y, a partir de dicho alineamiento, se obtienen las probabilidades de traducción para todos los pares de palabra (w(d|o) y w(o|d)), realizándose así una estimación de la tabla de traducción léxica más probable. A continuación, de todos los pares de subsecuencias obtenidos, se escogen con el programa *phrase_extract* sólo los que sean consistentes con el alineamiento de palabras. Y, por último, con el programa *phrase_score*, se calculan las probabilidades de traducción para todos los pares de subfrases en los dos sentidos: SMS-castellano y castellano-SMS.

C. Ajuste

Los modelos de traducción y de lenguaje (en lengua destino) se combina linealmente para ser utilizados como heurístico en el proceso de búsqueda necesario para generar la traducción de un mensaje dado. Esta combinación lineal utiliza unos pesos que hay que ajustar. Para ello, con el traductor *MOSES* ([8], [9]) se traduce y evalúa una lista de frases de validación (diferente a las listas de entrenamiento o test final) cuya traducción correcta se conoce, probando con distintos pesos aleatoriamente y escogiendo los que dan los mejores resultados.

D. Traducción

Por último, como hemos comentado anteriormente, para la traducción se emplea el decodificador *MOSES*, que es un sistema de traducción automática estadística basado en subsecuencias de palabras que implementa un algoritmo de búsqueda para obtener, a partir de una frase de entrada en SMS, la secuencia de palabras que con mayor probabilidad corresponde a su traducción. Para ello, utiliza los modelos de traducción y lenguaje obtenidos anteriormente, con los pesos de sus probabilidades ajustados.

V. EVALUACIONES

Las evaluaciones del sistema de traducción (*MOSES*) se realizan traduciendo un nuevo conjunto de frases. Adicionalmente, también se desarrolla una serie de funciones para el funcionamiento en tiempo real, pero que no se incluyen en la evaluación y, por tanto, en los resultados que se muestran.

A. Medidas de evaluación

Para evaluar cada uno de los experimentos que se realizan se emplean varias métricas que comparan la traducción que realiza el sistema con una traducción de referencia. Estas métricas son: WER (“Word Error Rate”, tasa de palabras con error), PER (“Position Independent Word Error Rate”). Y también BLEU (“BiLingual Evaluation Understudy”) y NIST, que son métodos de evaluación de la calidad de las traducciones, que comparan los n-gramas generados en la frase de referencia y en la traducción del sistema.

B. Experimentos iniciales

Inicialmente se realiza un experimento entrenando el modelo de lenguaje, el modelo de traducción y ajustando los pesos de estos modelos con los ficheros obtenidos a partir del diccionario de términos de la Web. Y para la evaluación se emplean los ficheros con frases de ejemplo en lenguaje SMS. Los resultados se muestran en la Tabla I.

Tabla I. Resultados iniciales entrenando y ajustando con términos y evaluando con frases

WER	PIWER	BLEU	NIST
58,82	58,82	0,0284	1,1084

C. Principales problemas y las soluciones propuestas

Los principales problemas de traducción encontrados y las soluciones tomadas son las siguientes (donde cada letra se relaciona con la presentación de resultados):

A. *Problema:* Términos en castellano con faltas de ortografía en todos los ficheros, debido a la no revisión manual de los registros introducidos por los usuarios a la hora de generar el diccionario.
Solución: Se corrige manualmente la ortografía de toda la base de datos, ya que, como se comentó anteriormente, tanto el diccionario de términos como las frases son generadas por usuarios anónimos y contienen numerosos errores que dan lugar a traducciones erróneas

B. *Problema:* Términos SMS en las frases de evaluación que van acompañados de signos de interrogación o exclamación. Estos términos existen en el modelo de traducción, pero sin los signos de interrogación y exclamación correspondientes, por tanto, el traductor no los encuentra. También ocurre lo contrario, es decir, términos sin signos de interrogación o exclamación de las frases en lenguaje SMS que se encuentran en los modelos, pero con los signos correspondientes.
Solución: Se eliminan de la base de datos todos los signos de interrogación o exclamación (salvo los de los emoticonos), para entrenar y evaluar sin estos signos. Posteriormente, en tiempo real, se tratarán adecuadamente, quitándolos antes de enviar la frase al traductor y añadiéndolos al final de la traducción si se da el caso, tal como se explica en el apartado de postproceso.

C. *Problema:* Términos SMS (abreviaturas de otras palabras) que en realidad son también palabras válidas en castellano sin abreviatura alguna (ej: *no* negación y *no* como abreviatura de número). Al escribir en lenguaje SMS se escriben palabras también en castellano sin abreviar. El traductor las toma como términos SMS y no los encuentra en el modelo de traducción.
Solución: Se incorpora todo el vocabulario en castellano de la base de datos duplicado en los ficheros de entrenamiento. De esta manera, si algún término en una frase en lenguaje SMS está en castellano, el traductor lo encontrará en los modelos y podrá traducirlo por la misma palabra. Tendrá que ser el modelo de lenguaje en destino el que deshaga esa ambigüedad.

D. *Problema:* Términos SMS que pueden tener distintas traducciones y se escoge la que ocurre con más probabilidad, pero que no es la correcta, ya que no se entrena bien el modelo de lenguaje.
Solución: Se entrena el modelo de lenguaje con 7 de las 8 listas de frases completas en lugar de hacerlo con los ficheros de términos. Es importante entrenar el modelo de lenguaje con el mayor número de frases completas posible.

E. *Problema:* Términos SMS, en general, que no se encuentran en la base de datos y, por lo tanto, no se pueden traducir.
Solución: Para ver el efecto que tiene este problema, para realizar la evaluación se incorporan en los modelos los términos desconocidos. Sin embargo, la solución real será intentar traducir el término desconocido en tiempo real con una función que devuelve el término en castellano de mínima distancia de edición al término en lenguaje SMS.

F. *Problema:* Mucho peso del modelo de traducción en relación con el peso del modelo de lenguaje. El ajuste de los pesos no se ha realizado correctamente.
Solución: En lugar de realizar el ajuste de los pesos de los modelos con ficheros que contienen términos o expresiones cortas, se ajustan con ficheros de frases completas, pues así el modelo de lenguaje tiene más posibilidad de ofrecer información valiosa en la generación de la salida correcta con un peso mayor en la traducción.

Tomando cada una de las soluciones anteriores, en la Figura 4 y en la Tabla II, puede observarse cómo va reduciéndose la WER. Como se puede observar, la incorporación de nuevos términos en el modelo de traducción y el correcto ajuste del peso del modelo de lenguaje son muy importantes.

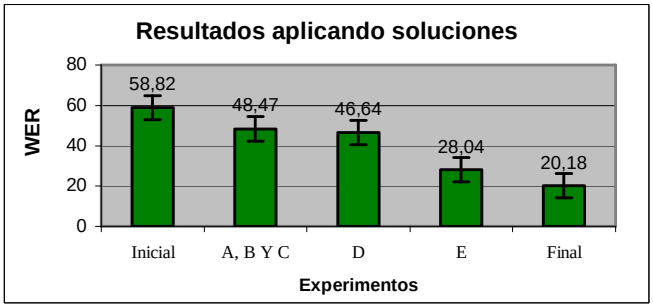

Figura 4. Variación de la tasa de error aplicando las distintas soluciones

Para el experimento final, se utilizan las 58 frases en lenguaje SMS organizados en 8 listas de frases aleatorias (junto con sus correspondientes traducciones al castellano). De manera que con 7 de ellas se entrena el modelo de lenguaje y se ajustan los pesos de los modelos, y con la restante se evalúa el sistema. Esta organización se va rotando para, al final, entrenar, ajustar y evaluar con todas. Obteniendo finalmente una tasa de error media del 20,18%.

Tabla II. Resultados obtenidos con cada uno de los problemas que se van resolviendo

	WER	PER	BLEU	NIST
Inicial	58,82	58,82	0,0284	1,1084
A, B y C	48,47	48,46	0,0370	0,9624
D	46,64	47,31	0,0382	0,9571
E	28,04	27,96	0,3578	4,5331
Final	20,18	17,55	0,4436	3,8445

VI. FUNCIONAMIENTO EN TIEMPO REAL

Para integrar el sistema de traducción en una interfaz de traducción en tiempo real es necesario implementar una serie de funciones que, por un lado, preparen las frases que se manden en tiempo real al traductor para que pueda trabajar con ellas (preproceso) y, por otro, retoquen las frases de salida del traductor, que en ocasiones puede no haber traducido bien todos los términos (postproceso).

A. Funciones de preproceso

En el preproceso se siguen los siguientes pasos (Figura 5):

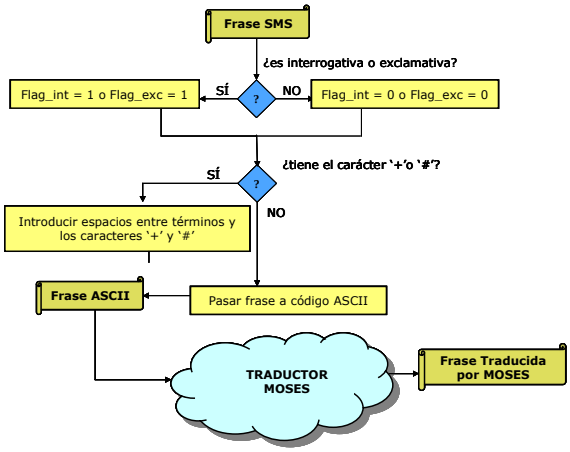

Figura 5. Diagrama de flujo del preproceso

- Comprobar si hay algún signo de interrogación o exclamación. De ser así, se elimina de la frase y se marca ese hecho (poniendo un flag a 1) para, posteriormente, tenerlo en cuenta en el postproceso.
- Comprobar si hay algún carácter '+' o '#' adyacente a algún término. De darse el caso, se introduce un espacio en blanco entre el carácter y el término. La razón es que, por lo general, estos dos caracteres aislados se traducen por las palabras "*más*" y "*número*", respectivamente. Por ejemplo: "*q+ kiers?*" que se traduce por "*¿qué más quieres?*".
- Por último, se convierten todos los términos de la frase en expresiones numéricas con los códigos ASCII, ya que los modelos de la arquitectura de traducción han sido entrenados con los términos en este formato.

B. Funciones de postproceso

La frase traducida por el traductor automático puede contener términos en castellano y términos en código ASCII (términos que no ha sido capaz de traducir). Además, la frase no tiene signos de interrogación ni exclamación. Por tanto, a continuación, se siguen los siguientes pasos (Figura 6):

- Se comprueba si el formato es el correspondiente a la secuencia de códigos ASCII. Si no lo es, se deja el término tal cual, porque significa que el traductor automático ha sabido traducirlo (con posibles errores).
- Si está en código ASCII, se pasa cada carácter a letra y se mira si el término resultante es pronunciable, para lo que se implementa una función que determina si un término es pronunciable en función de la secuencia de consonantes y vocales en castellano.
- Si es pronunciable, se deja así en la frase final, porque significa que es una palabra no abreviada o un nombre propio.
- Si no es pronunciable, se busca el término en castellano de una lista de vocabulario que se le parezca más (calculando la distancia de Levenshtein como se explica más adelante) y se vuelca dicho término a la frase final.
- Al final, cuando la frase está completa, se comprueba si la frase en lenguaje SMS era interrogativa o exclamativa (indicado con un flag). De ser así, se ponen los signos de interrogación o exclamación al principio y al final de la frase, obteniendo la frase final.

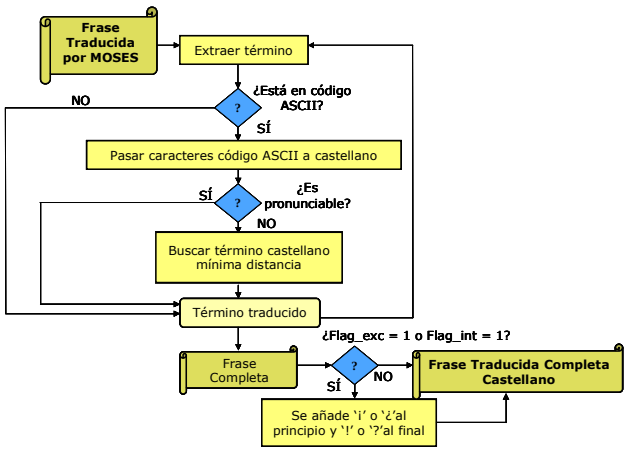

Figura 6. Diagrama de flujo del post-proceso

Como se menciona anteriormente, para traducir un término que el traductor automático no ha podido traducir, se vuelca como resultado el término en castellano de mínima distancia de edición, que es el número mínimo de operaciones que hay que realizar para transformar una cadena de caracteres en otra: borrado de un carácter, sustitución de un carácter por otro, e inserción de un carácter nuevo. A cada operación anterior se le asigna un coste 1. Sin embargo, en el sistema descrito en este artículo, se modifica la distancia de Levenshtein, de manera que se fomente la inserción de caracteres. Así, el borrado tiene un coste 4, la sustitución un coste 4 y la inserción un coste 1, ya que para pasar de un término en lenguaje SMS a un término en castellano se realizan muchas inserciones, pero pocos borrados o sustituciones. El coste final de transformar una cadena en otra será la suma de los anteriores. Por ejemplo, dado el término "*kntndo*". Su distancia a la palabra "cantando" es 6 (1 sustitución y 2 inserciones). De esta manera, la función recorre todas las palabras de una lista de vocabulario y va calculando la distancia del término a cada una de las palabras. Y, al final, se devuelve la palabra que tiene la mínima distancia.

VII. CONCLUSIONES

En este artículo se ha descrito el desarrollo de un sistema de traducción de lenguaje SMS a castellano. La arquitectura completa consta de un módulo de traducción estadística basada en subfrases con un modelo de lenguaje y un modelo de traducción, empleando como traductor el programa *MOSES*. Pero, además, se ha implementado una serie de funciones que preparan la frase en lenguaje SMS para mandarla al traductor, y funciones para retocar la frase de salida del traductor, permitiendo traducir términos en tiempo real que el traductor no sabe traducir.

Para el desarrollo del sistema se empleó como base de datos un diccionario de términos y expresiones cortas en lenguaje SMS generados por usuarios de Internet. Esta base de datos se utilizó para entrenar el modelo de traducción. Además, también se utilizaron 58 frases de ejemplo de mensajes cortos en lenguaje SMS para realizar la evaluación de la plataforma, y para entrenar el modelo de lenguaje y ajustar los pesos de los modelos.

Tras realizar varios experimentos, analizar los fallos en la traducción y aplicar distintas soluciones a los problemas encontrados, se obtuvo una WER de aproximadamente 20,2%.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el Plan Nacional de I+D con los siguientes proyectos: PAV-070000-2007-567 (Plan Avanza), ROBONAUTA (DPI2007-66846-c02-02) y SD-TEAM (TIN2008-06856-C05-03).

REFERENCIAS

[1] Esendex: http://www.esendex.es/Envio-de-SMS/Voz-SMS
[2] Comsys: http://www.comsys.net/?page=SMS_2_fixed.htm&submenu=products
[3] Cboss: http://www.cbossgroup.com/products/cbosssms2voice.html
[4] AiTi Aw, Min Zhang, Juan Xiao, Jian Su. "A Phrase-based Statistical Model for SMS Text Normalization", *ACL 2006*.
[5] A. Stolcke. "SRILM – An Extensible Language Modelling Toolkit". *ICSLP 2002*.
[6] GIZA++: http://www.fjoch.com/GIZA++.html)
[7] Franz Josef Och, Hermann Ney. "A Systematic Comparison of Various Statistical Alignment Models", *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003.
[8] Philip Koehn. "Statistical Machine Translation". *Cambridge University Press 2010*.
[9] Moses Translation System: http://www.statmt.org/moses/.