ISSN 1946-5351 (Online) Octubre 15 y 16, 2009 ISSN 1948-2353 (CD ROM)

Desarrollo de un traductor textual Lenguaje SMS – Español Tradicional – Lenguaje SMS

Ismael Esquivel Gámez, Daniel Rodríguez Angeles

Resumen— Con el auge actual de la comunicación vía electrónica, ha surgido el Lenguaje SMS, cuyo principal inconveniente es lo difícil que resulta su comprensión a personas que no suelen usarlo, representando así, un obstáculo en la comunicación entre jóvenes (mayoría de usuarios del lenguaje SMS) y no usuarios, generalmente de generaciones anteriores. El presente documento describe el desarrollo y los resultados de un sistema de traducción de Lenguaje SMS a español tradicional basado en el más grande diccionario de términos SMS, denominado "Diccionario SMS". El sistema ha sido probado con un elevado porcentaje de efectividad en un paquete de 1000 mensajes SMS recopilados de páginas de internet, salas de chat y blogs, los autores de dichos mensajes son personas de diferentes edades, intereses, estados sociales y actividades.

Palabras claves— Lenguaje SMS, traductor, SMS, lingo. Language SMS

I. Introducción

Los primeros vestigios de comunicación escrita se hallaron hace más de tres milenios, esto tratando de satisfacer la necesidad innata del hombre por comunicar sus ideas. Desde entonces la escritura ha tenido una evolución muy grande, ha cambiado de acuerdo a la época, al lugar y circunstancias históricas, una constante dentro de esta evolución es el intento idear modos de abreviar, esto es, comunicar la mayor cantidad de ideas en el menor número de caracteres.

De éste modo surgen diversas técnicas muy aceptadas, una de ellas es la taquigrafía, que según La Real Academia Española [1] es el arte de escribir tan de prisa como se habla, por medio de ciertos signos y abreviaturas, se emplean trazos breves y caracteres especiales para representar letras, palabras, incluso frases. Generalmente la escritura taquigráfica omite partes de los textos, por ende, un texto recogido por un taquígrafo no puede ser fácilmente entendido por otro que no haya escuchado el texto original.

Dentro de la taquigrafía hay dos grandes corrientes: Pitman y Gregg. La taquigrafía Gregg inventada por John Robert Gregg [2] en 1888 y establece que es un sistema de escritura fonética, lo que quiere decir que graba los sonidos del hablante, no su ortografía. En éste principio radica la facilidad para adaptarla a diversos idiomas.

Algunos años después, se encuentran más evidencias del intento por acortar lo que se escribe. Josep M. Casasús [3] comenta que Walter Lippmann, en su obra clásica *Public* opinión, en 1922, habla del código de *Phillip*, muy parecido al lenguaje SMS usado hoy en los dispositivos móviles y empleado entonces por los periodistas para enviar noticias, con ahorro de tiempo, de dinero y de energía, con el mismo principio, recortar palabras aprovechándose de sus fonemas. También indica que según Alemán Ocampo, las abreviaciones se debían inicialmente a la escasez de tinta y papel. "Y en este siglo, la condición que limita a uno es la pantalla del celular, que obliga a realizar una comunicación rápida, y por eso recurrimos a la abreviación".

Como éstas, encontramos varias otras formas de acortar lo que se escribe; las siglas que son el conjunto de letras iníciales de una expresión compleja [1]. Las abreviaturas también surgen con la idea de ahorrar espacio, de escribir lo necesario, en este caso se logra suprimiendo las letras centrales o finales y aun así la palabra es perfectamente entendible. Una de las más recientes variaciones del lenguaje escrito es el llamado Lenguaje SMS, que es la forma de escritura utilizada en los dispositivos

móviles a la hora de redactar un mensaje SMS (*Short Message Service*) SMS es una tecnología europea que apareció en 1991, n un principio se desarrolló para dar informes de la bolsa y horóscopos. El lenguaje SMS, mayormente utilizado por jóvenes, surge como herramienta para ahorrar espacio y dinero, puesto que en un mensaje de texto se cuenta con solo 160 caracteres para expresar la idea que se requiera.

A tal grado llega la aceptación de ésta nueva lengua que ya se han publicado libros escritos de ese modo. El primero, por Marso Phil [6] "La cruz secreta del emperador" fue publicado en Francia y está consagrado a los daños del tabaquismo. El autor de este libro, dice que "corre el riesgo de irritar a los defensores de la lengua francesa", pero insiste sobre la legitimidad de esta traducción. Curiosamente el autor no es un aficionado del móvil: es promotor, desde 2001, de la "Jornada mundial sin teléfono móvil", que tiene lugar todos los años el 6 de febrero.

De la sintaxis usada en éste libro y en varios otros se desprenden normas para la escritura SMS, José Luis Hernández Pacheco [7] establece ciertas reglas para el uso de éste lenguaje, como:

Uso de la "H" y de la "E". La letra "h" es muda en la pronunciación, así que en mensajes cortos se obvia y de esta forma se ahorra un carácter. La "e" al principio de palabra también se suprime y la palabra es perfectamente entendible, por ejemplo: "str" por estar; "n" por en. Por otro lado, es importante tocar temas como la lingüística computacional que según Xavier López Morrás [4] es un campo interdisciplinario que se ubica entre la lingüística y la informática: su fin es la elaboración de modelos computacionales que reproduzcan distintos aspectos del lenguaje humano. Puede considerarse una disciplina de la lingüística aplicada a la Inteligencia Artificial y tiene como objetivo la realización de aplicaciones informáticas que imiten la capacidad humana de hablar y entender. A la Lingüística Computacional se le llama a veces Procesamiento del Lenguaje Natural (PLN). Ejemplos de aplicaciones de PLN son los programas que reconocen el habla o los traductores automáticos. Ésta surgió en los EE. UU. en los años 50 como un esfuerzo para obtener computadoras capaces de traducir textos automáticamente de lenguas extranjeras al inglés, particularmente de revistas científicas rusas. Es así como surge la traducción automática (TA), también llamada MT (del inglés *Machine Translation*), que en opinión de Raquel Martínez [5], es un área de la lingüística computacional que investiga el uso de software para traducir texto o habla de un lenguaje natural a otro. En un nivel básico, la traducción por computadora realiza una sustitución simple de las palabras de un lenguaje natural por las de otro. Por medio del uso de corpora lingüísticos se pueden intentar traducciones más complejas, lo que permite un manejo más apropiado de las diferencias en la tipología lingüística, el reconocimiento de frases, la traducción de expresiones idiomáticas y el aislamiento de anomalías.

ESTADO DEL ARTE

Para el lenguaje SMS en español existe una página de internet que contiene términos con su significado, www.diccionarioSMS.com [8], cuyo objetivo es recabar los términos y abreviaturas que utilizan por los jóvenes para escribir en sus teléfonos móviles o cuando lo hacen en Internet.

Es una herramienta de consulta creada por los jóvenes, estudiosos de la lengua española, medios de comunicación, padres y educadores. Para el lenguaje SMS en ingles está www.netlingo.com [9], con la misma función del anteriormente mencionado. NetLingo y DiccionarioSMS son para sus respectivos idiomas las bases de datos con mas términos, sin embargo, NetLingo, con poco más de 5000, no tiene ni la mitad de los términos de diccionarioSMS.

Existe un trabajo de investigación semejante al presente, aplicado al idioma inglés, "*A Phrase-based Statistical Model for SMS Text Normalization*" [10] es decir, un modelo estadístico basado en frases para

la normalización de textos SMS. El método de normalización consiste básicamente en dos sub-modelos: uno basado en palabras y un modelo de mapeo léxico basado en frases. En dicha investigación se compara la técnica de normalización con varios métodos de traducción, obteniendo resultados muy adelantados. Se hace un experimento traduciendo de inglés a chino mediante los mismos algoritmos, aumentando el porcentaje de efectividad en la traducción.

PROBLEMÁTICA

Es innegable que el ser humano tiene una innata tendencia de rechazo a lo nuevo, a lo que no entiende, a lo que transgrede sus costumbres, a lo que contradice lo que cree correcto, sin embargo, más allá de los rechazos infundados, el Lenguaje SMS, al ser un tanto exclusivo de las generaciones nativas en éstas tecnologías, representa un obstáculo para la comunicación entre diferentes generaciones, si un joven de 18 años le escribe a su padre de 45 "*pa, n m sprs, vy a ygar trd xq m qdar n ksa d Bre a acr 1a tarea*", sin duda, lo dejará preocupado toda la noche. Se tiene entonces, un problema de comunicación.

PROPUESTA DE SOLUCIÓN

El desarrollo de una aplicación que sea capaz de traducir un mensaje o texto de lenguaje SMS a español tradicional, de modo que los no usuarios de ésta forma de expresión, tengan toda la facilidad para comprender lo que se les quiere comunicar. Se elige www.diccionarioSMS.com para tomar como base de la traducción por ser la recopilación de términos SMS más grande, con más de 11,000 términos supera por mucho a otros sitios que contienen diccionarios de palabras SMS, según se muestra en la tabla 1, además, cuenta con una información que ninguno otro, la popularidad del término, ésta se genera en base a las veces que los usuarios dan de alta un término.

Diccionarios electrónicos	Número de términos
www.diccionarioSMS.com	11364
www.viajoven.com	649
www.cabinas.net	142
www.lonuncavisto.com	123

Tabla 1. Comparación de diccionarios

El lenguaje de programación utilizado para el desarrollo del traductor es Rexx (*REstructured eXtended eXecutor*) un lenguaje de programación desarrollado en IBM por Michael Cowlishaw del que existen numerosas implementaciones disponibles con código abierto [11]. Es un lenguaje de programación estructurado de alto nivel diseñado para ser al mismo tiempo fácil de entender y fácil de leer.

Se decide usar éste lenguaje puesto que sus características se adaptan a las necesidades, principalmente porque cuenta con un gran conjunto de funciones, especialmente para procesado de cadenas y palabras. Otra ventaja determinante para usar Rexx es que es un lenguaje multiplataforma, es decir, sin moverle un punto ni una coma se podrá ejecutar en cualquier computadora, bajo cualquier sistema operativo y funcionará.

METODOLOGÍA

Para el desarrollo de la aplicación se siguen éstos pasos:

A. Se genera una base de datos que contiene los términos SMS conseguidos de www.diccionarioSMS.com
B. Se crea un vector con los términos ambiguos.
C. Se leen los desde archivo y se compara palabra por palabra contra el diccionario, buscando primero en el vector de términos ambiguos.
D. Si se encuentra se traduce por el significado de mayor popularidad, si no,
E. Se busca en el diccionario general y se traduce, sí no,
F. Se pasa el término tal cual,
G. Se imprimen los resultados en un archivo de salida.

Enseguida se detalla cada proceso.

Generación de base de datos de términos

Se genera en base al contenido del diccionarioSMS (www.diccionarioSMS.com), se guarda en un archivo separado por comas. Se ordenan los términos por su valor en el código ASCII, los términos con más de un significado se ordenan por el valor de su popularidad para después aprovecharlo en el la resolución de la ambigüedad que generan.

Ya ordenado el archivo se carga en 3 vectores, el primero contendrá los términos SMS, el segundo su equivalente en español tradicional y el último contiene la popularidad de cada término.

Una vez cargado el diccionario se procede a la lectura de los mensajes a traducir, mediante un proceso de búsqueda binaria, se busca cada palabra en el vector de términos SMS y si se encuentra es cambiado por el primero que se encuentra.

Después del anterior proceso los resultados aparecen como se muestra en los siguientes ejemplos:

SMS:
N KIERO IR AD+ N TNGO DINERO
ESPAÑOL:
EN QUIERO IR ADEMAS EN TENGO DINERO
- -
SMS:
OLA Q ACS DSPIERTO
ESPAÑOL:
HOLA QUE HACES DESPIERTO
- -
SMS:
KTAL?
ESPAÑOL:
¿QUé TAL?
- -
SMS:
T EXO D-
ESPAÑOL:
TE ECHO DE MENOS
- -

Detección de ambigüedad

Con los anteriores resultados se determina que los términos con más de un significado son los que generan error en la traducción, las palabras que no se encuentran son mínimas.

Para solucionar dicho problema se genera un vector más en el que se almacenarán solo los términos SMS que tienen más de un equivalente en español tradicional, algunos ejemplos se muestran en la tabla 2.

Término SMS	Equivalente en español	Popularidad
Akb	acabo	11
Akb	acabe	5
Akb	acaba	1
Akb	acabar	1
st	este	47
st	esto	6
st	esta	3
sts	estas	17
sts	estés	4
sts	estos	4

Tabla 2. Términos ambiguos

La traducción se realiza con el término encontrado con mayor popularidad, se hace de éste modo puesto que en la mayoría de los términos ambiguos existe uno que sobresale por su popularidad, según se puede ver en la tabla 3, la diferencia tan grande en la frecuencia de uso de los términos hace que se tenga una alta probabilidad de éxito en la traducción.

Término SMS	Equivalente en español	Popularidad
n	en	65
n	ni	5
n	número	2
N	norte	1
cm	como	62
cm	centímetro	2
cm	cama	1
cm	coma	1

Tabla 3. Diferencia en popularidad de términos ambiguos

Los resultados de la traducción buscando y traduciendo primero los términos que generan ambigüedad son los siguientes:

SMS:
N KIERO IR AD+ N TNGO DINERO
ESPAÑOL:
NO QUIERO IR ADEMAS NO TENGO DINERO
- -
SMS:
OLA Q ACS DSPIERTO
ESPAÑOL:
HOLA QUE HACES DESPIERTO
- -
SMS:
KTAL?
ESPAÑOL:
¿QUé TAL?
- -
SMS:
T EXO D-
ESPAÑOL:

TE HECHO DE MENOS

- -

El grado de efectividad logrado aumenta considerablemente usando el término más popular, esto se debe a la cantidad de términos ambiguos contenidos en los mensajes, que si bien no son mayoría, si llegan a ser suficientes para cambiar el sentido de la frase o dificultar su comprensión. En una traducción de 140 mensajes, cuyos detalles se muestran en la tabla 4, se encuentran alrededor de 400 términos SMS que han de ser traducidos, de los cuales el 45% tienen más de un significado.

Muestra	140 mensajes
Términos SMS	400 aprox.
Términos ambiguos	180 aprox.

Tabla 4. Estadística de términos ambiguos

El avance en la traducción se hace evidente en frases como la siguiente:

T KIERO + Q A TDO L MNDO

El primer resultado, comparando contra el diccionario general y traduciendo por el primer término encontrado era:

TE QUIERO MAS QUE A TODO LITRO MANDO

Ahora, tomando primero los términos ambiguos y usando el significado de mayor popularidad, se traduce de ésta manera:

TE QUIERO MAS QUE A TODO EL MUNDO

Definitivamente el mensaje queda en modo comprensible, regresando al ejemplo que se dio inicialmente (pa, n m sprs, vy a ygar trd xq m qdar n ksa d Bre a acr 1a tarea ”), si el padre pudiera introducir ese mensaje al traductor obtendrá el siguiente resultado:
PARA, NO ME SPRS, VOY A LLEGAR TARDE POR QUÉ ME QUEDAR NO CASA DE BRE A HACER UNA TAREA

Seguramente el mensaje ahora si sería más entendible.

Traducción Español tradicional – Lenguaje SMS

Después de desarrollado el módulo para la traducción en forma inversa, es decir, del Español tradicional a Lenguaje SMS, los resultados son los siguientes:

ESPANOL:
NO QUIERO IR ADEMAS NO TENGO DINERO
SMS:
N KERO IR AD+ N TNG MONEY
- - - - - - - - - - - - - - - - - -
ESPANOL:
HOLA QUE HACES DESPIERTO
SMS:
OLA Q ACS DSPIERTO
- - - - - - - - - - - - - - - - - -

ESPANOL:
¿QUé TAL?
SMS:
KTAL?
- - - - - - - - - - - - - - - - - -
ESPANOL:
TE HECHO DE MENOS
SMS:
T EXO D -
- - - - - - - - - - - - - - - - - -

CONCLUSIONES

En este proyecto se ha realizado un prototipo de traductor de la Lenguaje SMS a español tradicional basado en la comparación con un diccionario de términos SMS. Admite como entrada frases en lenguaje SMS que pueden provenir de cualquier archivo o ser directamente capturadas. La aplicación desarrollada permite una mejor comunicación entre usuarios del lenguaje SMS y personas a las que se les dificulta la comprensión del mismo.

El grado de efectividad es afectado por el uso de términos poco populares en los mensajes, ya sea que son traducidos con un significado erróneo o simplemente no se encuentran en el diccionario. Como todo diccionario, el diccionarioSMS se irá enriqueciendo con el tiempo, aumentando de este modo la efectividad del traductor. La adaptación del diccionario al Lenguaje SMS usado en México también elevará la calidad de la traducción.

TRABAJO FUTURO

Se continuará con la investigación y desarrollo de mejores métodos de traducción que permitan un mayor porcentaje de efectividad.

Se aumentará el diccionario SMS conforme crezca en la base de datos de www.diccionarioSMS.com y se creará un módulo para poder dar de alta palabras directamente.

Se tiene planeado proponer al sitio anterior, agregar el traductor a su página web, como aportación de los autores y de la Universidad Veracruzana.

REFERENCIAS

1. [1] REAL ACADEMIA ESPAÑOLA © TODOS LOS DERECHOS RESERVADOS DICCIONARIO DE LA LENGUA ESPAÑOLA VIGÉSIMA SEGUNDA EDICIÓN
2. [2] TAQUIGRAFÍA
3. HTTP://ES.WIKIPEDIA.ORG/WIKI/TAQUIGRAFÍA VISITADO EN MAYO DE 2009
4. [3] CASASÚS JOSEP M.
5. HTTP://WWW.OBSERVATORIODAIMPRENSA.COM.BR/ARTIGOS.ASP?COD=305VOZ008
6. "QUE EL LENGUAJE RÁPIDO NO NOS EMPOBREZCA" COPYRIGHT LA VANGUARDIA VISITADO EN MAYO DE 2009
7. [4] LÓPEZ MORRÁS XAVIER
8. HTTP://WWW.AUCEL.COM/PLN/K-ES.HTML
9. ¿QUÉ ES LA LINGÜÍSTICA COMPUTACIONAL O PLN? VISITADO EN JUNIO DE 2009
10. [5] MARTÍNEZ RAQUEL
11. PRINCIPALES PROBLEMAS DE LA TRADUCCIÓN AUTOMÁTICA. UNIVERSIDAD JUAN CARLOS I.
12. [6] MARSO PHIL
13. LA CRUZ SECRETA DEL EMPERADOR
14. [7] HERNÁNDEZ PACHECO JOSÉ LUIS
15. MIRAFLORES GÓMEZ EMILIO LENGUAJE SMS: LA ALFABETIZACIÓN DE LOS JÓVENES EN EL SIGLO XXI

16. [8] DICCIONARIO SMS
17. WWW.DICCIONARIOSMS.COM VISITADO EN MAYO DE 2009

18. [9] NETLINGO.COM
19. NETLINGO® THE INTERNET DICTIONARY AT HTTP://WWW.NETLINGO.COM VISITADO EN JUNIO DE 2009

20. [10] MIN ZHANG, XIAO JUAN, SU JIAN
21. INSTITUTE OF INFOCOMM RESEARCH
22. Heng Mui Keng Terrace Singapore 119613

23. [11] MERTZ DAVID
24. REXX FOR EVERYONE: UNA INTRODUCCIÓN
25. IBM DEVELOPERWORKS.

AUTORES

Dr. Ismael Esquivel Gámez se desempeña como maestro de tiempo completo de Facultad de Administración, Región Veracruz de la U.V., iesquivel@uv.mx
Lic. Daniel Rodríguez Angeles daniel.rguez@live.com.mx (**autor corresponsal)**