DLSI en Tweet-norm 2013: Normalización de Tweets en Español*

DLSI at Tweet-norm 2013: Normalisation of Spanish Tweets

Alejandro Mosquera
DLSI-Universidad de Alicante
Alicante
amosquera@dlsi.ua.es

Paloma Moreda
DLSI-Universidad de Alicante
Alicante
moreda@dlsi.ua.es

Resumen: La gran variedad léxica y su facilidad de acceso a un gran volumen de información convierten a la Web 2.0 en un recurso importante para el Procesamiento del Lenguaje Natural. Sin embargo, la frecuente aparición de fenómenos lingüísticos no normativos pueden dificultar el procesado automático de estos textos. En este trabajo se describe la participación en el taller sobre Normalización de Tweets en el congreso de la SEPLN (Tweet-norm 2013). El taller propone una única tarea con el objetivo de estandarizar textos no normativos en español extraídos de Twitter. Para dicha tarea, se ha hecho uso de TENOR, una herramienta de normalización multilingüe para textos de la Web 2.0.
Palabras clave: Normalización, Tweets

Abstract: The lexical richness and its ease of access to large volumes of information converts the Web 2.0 into an important resource for Natural Language Processing. Nevertheless, the frequent presence of non-normative linguistic phenomena that can make any automatic processing challenging. In this paper is described the participation in the Text Normalisation Workshop at the SEPLN conference (Tweet-norm 2013). The Workshop includes one unique task focused on the normalisation of Spanish tweets. For this task we have used TENOR, a multilingual lexical normalisation tool for Web 2.0 texts.
Keywords: Normalisation, Tweets

1. Introducción

Desde la aparición de los primeros medios de comunicación sociales, las aplicaciones de la Web 2.0 han ganado popularidad en Internet. Enciclopedias colaborativas como Wikipedia, sitios de micro-blogging como Twitter o redes sociales como Facebook se encuentran entre los primeros puestos por número de visitas de la Red[1].

Estas aplicaciones han transformado el flujo de la información que se comparte en Internet. Este cambio de paradigma se centra en los usuarios, quienes generan y consumen dicha información. La naturaleza informal de dicho intercambio y la diversidad geográfica y social de los usuarios se refleja en su lenguaje escrito, siendo frecuente la aparición de fenómenos lingüísticos no normativos tales como emoticonos, supresión de grafemas y sustituciones léxicas entre otros.

Por ejemplo, en el caso particular de Twitter[2], el número máximo de caracteres por mensaje está limitado a 140, por lo que es común encontrar abreviaciones y contracciones no-estándar. Así, igual que en los mensajes SMS, algunas palabras o sílabas pueden ser representadas por letras o números que tienen la misma pronunciación pero cuyo tamaño es menor. Por ejemplo, *cansados* tiene una pronunciación equivalente a *cansa2*. De la misma forma, la sílaba o conjunción *que* puede ser sustituida por *k* o *q*. Otra forma de acortar palabras es omitir ciertas letras, normalmente vocales. Por ejemplo, la palabra *trabajo* puede ser acortada como *trbj*. Por otra parte, la expresión de emociones o estados de ánimo suele hacerse mediante emoticonos, empleándose para ello acentos, comas y otros símbolos de puntuación, pudiendo es-

* Este artículo ha sido cofinanciado por el Ministerio de Ciencia e Innovación (proyecto TIN2009-13391-C04-01), y la Conselleria d'Educación de la Generalitat Valenciana (proyectos PROMETEO/2009/119, ACOMP/2010/286 y ACOMP/2011/001)

[1] http://www.alexa.com/topsites

[2] http://www.twitter.com

tos ser omitidos deliberadamente en el resto del texto.

La gran variedad léxica unida al gran volumen de textos disponible en la Web 2.0 la convierten en un recurso importante para el procesamiento del lenguaje natural (PLN). Sin embargo, sus características informales complican el procesado de este tipo de textos de forma automática. Entre los estudios que abordan este problema destacan los que hacen uso de técnicas de normalización. Entendiendo el concepto de normalización como un proceso que permite «limpiar» una palabra o texto transformando las variantes léxicas no-estándar del lenguaje en sus formas canónicas. Sin embargo, la gran mayoría de los trabajos realizados en esta línea han sido para el idioma inglés. Por esta razón, Tweet-norm propone la normalización léxica de Tweets en español. Con el objetivo de abordar esta tarea se ha hecho uso de la herramienta de normalización multilingüe TENOR(Mosquera y Moreda, 2012) y a continuación expondremos la metodología empleada y los resultados obtenidos..

El artículo se organiza de la siguiente forma: en la sección 2 se describe el estado de la cuestión. En la sección 3 explicamos nuestra metodología. Los resultados obtenidos en el taller son evaluados en la sección 4. Finalmente en la sección 5 se comentan las conclusiones y trabajo futuro.

2. Estado de la Cuestión

Se pueden distinguir tres tendencias principales a la hora de normalizar este tipo de textos. La primera emplea técnicas de traducción automática, la segunda se basa en corrección ortográfica y la tercera usa técnicas de reconocimiento del habla.

La aplicación de técnicas de traducción automática se ha demostrado útil para normalizar textos SMS (Aw et al., 2006) tomando como idioma origen los textos no-normativos y como idioma destino su equivalencia normalizada. Este sistema también se ha usado para traducir textos en lenguaje SMS al español (López et al., 2010), siendo la única que hemos encontrado para este idioma, y empleando el motor de traducción estadística MOSES (Hoang et al., 2007). Sin embargo, estas propuestas de traducción necesitan corpus relativamente grandes previamente normalizados y alineados para obtener buenos resultados (Kaufmann, 2010).

El uso del modelo de Shannon para canales con ruido (Shannon, 1948) se suele emplear en los sistemas de corrección automática para realizar una corrección ortográfica a nivel de palabra (Choudhury et al., 2007). Dichos errores ortográficos pueden ser intencionales para darle énfasis y sentimiento a una palabra *(goooooooooool!!)* o contracciones no-estándar homófonas para ahorrar espacio *(knsado)*. Tanto este caso como con textos de escasa longitud se dificulta considerablemente la tarea de normalización empleando este modelo, ya que el contexto no juega un papel tan relevante.

Por último, las técnicas de reconocimiento automático del habla (RAH) se basan en la hipótesis de que la mayoría de las variantes léxicas no-estándar tienen una equivalencia homófona estándar *(ksa - casa)*. Empleando algoritmos fonéticos para codificar la pronunciación de la palabra a normalizar se genera una lista de candidatos homófonos de la cual se extrae la palabra normalizada mediante modelos del lenguaje (Gouws et al., 2011). Los sistemas de normalización no-supervisada basados en esta técnica han obtenido los mejores resultados (Han y Baldwin, 2011).

3. Normalización de Tweets con TENOR

Hemos participado en Tweet-norm empleando la herramienta de normalización multilingüe TENOR, siguiendo una estrategia similar a la usada satisfactoriamente en textos de la Web 2.0 y SMS en inglés empleando técnicas de RAH pero adaptada a las singularidades del idioma español. Dado que TENOR está orientado principalmente a la sustitución de variantes léxicas se ha adaptado su funcionamiento acorde a los objetivos del taller de normalización.

En primer lugar definiremos el ámbito de la tarea propuesta en el apartado 3.1. En el apartado 3.2 explicaremos la metodología empleada.

3.1. Ámbito de la tarea

El objetivo del taller consiste en estandarizar una cantidad determinada de tweets con serios problemas de normalización. El sistema propuesto debe ser capaz de etiquetar las palabras dentro de tres grupos dependiendo si se tratan de variantes léxicas, palabras correctas o si pertenecen a otro idioma y ob-

tener su versión canónica. En el caso de las palabras pertenecientes a otro idioma si existiesen errores ortográficos también se debe proporcionar la versión correcta.

3.2. Metodología

TENOR sigue un proceso de normalización compuesto de dos pasos: En primer lugar se emplea un método de clasificación con el fin de detectar variantes léxicas no-estándar o fuera del vocabulario; En segundo lugar, se sustituyen las palabras seleccionadas en el paso anterior por su forma original normalizada.

3.2.1. Detección de palabras fuera del vocabulario

En este estudio nos referimos a las palabras fuera del vocabulario como aquellas que no forman parte del vocabulario español estándar y requieren ser normalizadas. Sin embargo, la detección de este tipo de palabras no es una tarea trivial: La presencia de palabras en otros idiomas, neologismos o siglas, así como la riqueza lingüística del español dificulta la tarea de conocer si una palabra pertenece al idioma o por el contrario es una variante léxica no-normativa. Ya que se ha usado como sistema de referencia durante el proceso de anotado de los textos del taller[3], se ha hecho uso de Freeling(Atserias et al., 2006) para dicha tarea.

3.2.2. Sustitución de variantes léxicas

En este apartado hablaremos de los diferentes pasos que se llevan a cabo para reemplazar las palabras clasificadas como fuera del vocabulario en la sección anterior por su forma normalizada. En primer lugar, se introducirán diversas técnicas de filtrado empleadas para «limpiar» los textos. En el siguiente paso, se detalla el proceso de sustitución de abreviaturas y transliteraciones. A continuación, se comentará el algoritmo de indexado fonético implementado en TENOR con el objetivo de obtener listas de palabras con pronunciaciones equivalentes. Posteriormente, este método se aplicará con el objetivo de identificar posibles candidatas para reemplazar las palabras no-normativas. Finalmente, se explica como el uso de algoritmos de similitud y modelos del lenguaje puede ayudar a seleccionar la forma canónica más apropiada para cada sustitución a partir de la lista de palabras candidatas.

Filtrado: En primer lugar, se han eliminado todos los caracteres no imprimibles y símbolos de puntuación no estándar excepto los emoticonos.

Abreviaturas y Transliteraciones: El segundo paso del análisis es comprobar que la palabra no perteneciente al vocabulario sea una abreviatura, la cual se sustituye por su equivalencia normalizada. En caso contrario, mediante reglas heurísticas se reducen las repeticiones de vocales o consonantes dentro de la palabra *(nooo!, goooooll)*. Posteriormente se analiza la presencia de números cuya pronunciación es frecuentemente utilizada para acortar la longitud del mensaje *(separa2, ning1)* o combinación alfanumérica *(c4s4)*, sustituyéndose por su transliteración más apropiada mediante una tabla de equivalencias.

De forma adicional, se ha compilado manualmente una tabla de equivalencias con 146 de las abreviaturas más comunes *(qtal, xfa)* que necesitan un tratamiento especial al ser expresiones compuestas o variantes que guardan muy poca o ninguna similitud léxica con su equivalencia normalizada.

Indexado Fonético: Se ha empleado el diccionario expandido de GNU Aspell[4] aumentado con nombres de países, ciudades, siglas y nombres propios más comunes. El léxico resultante de 931.435 palabras incluye conjugaciones en diferentes tiempos verbales y entidades nombradas. Posteriormente, se ha construido un índice fonético con las palabras de dicho léxico agrupándolas en base a su pronunciación. Esto se ha realizado de forma no-supervisada empleando el algoritmo del metáfono (Philips, 2000) adaptado al español. Este sistema permite representar la pronunciación de una palabra empleando un conjunto de reglas. Por ejemplo, el metáfono caracterizado por *(JNTS)* permite indexar las siguientes palabras *gentes, gentíos, jinetas, jinetes, juanetes, juntas y juntos* entre otras.

En la siguiente parte del proceso se obtiene el metáfono de la palabra resultante y se comprueba su presencia en el índice fonético para obtener una lista de posibles palabras candidatas en caso de encontrar una coincidencia.

[3] http://komunitatea.elhuyar.org/tweet-norm/files/2013/05/Manual_para_participantes_-_Tweet-norm.pdf

[4] http://aspell.net

Similitud Léxica: El algoritmo Gestalt (Ratcliff y Metzener, 1988) que está basado en el principio de la máxima sub-secuencia común, permite obtener un índice de similitud entre dos cadenas con valores entre 0 y 100, donde 100 es máxima similitud y 0 es ausencia de similitud. Se ha calculado la similitud de la palabra a normalizar con cada una de las candidatas fonéticas obtenidas en el paso anterior. Posteriormente, las candidatas con un índice de similitud menor de 60 han sido descartadas ya que por debajo de este umbral no se han observado resultados fiables.

Modelos del Lenguaje: Finalmente, cuando hay más de una palabra candidata con la misma similitud léxica se ha utilizado un modelo de lenguaje basado en trigramas y entrenado sobre el corpus CESS-ESP (Martí y Taulé, 2007).

4. Evaluación

4.1. Corpus utilizado:

Se ha hecho uso del corpus de test[5] proporcionado por la organización para evaluar los resultados. Dicho corpus consta de 564 tweets correspondiente a los días 1 y 2 de abril de 2013 localizados en el área geográfica de la península ibérica, eliminando aquellas regiones que tienen lenguas cooficiales. Este corpus contiene textos mayoritariamente en español.

4.2. Resultados

Se han enviado 2 ejecuciones, la primera (DLSI-Alicante-1) empleando Freeling para extraer las palabras fuera del vocabulario y la segunda (DLSI-Alicante-2) haciendo uso de las palabras fuera del vocabulario existentes en el corpus de test. Los resultados obtenidos en la tarea solamente se han evaluado en base a la precisión y se describen en el Cuadro 1.

Corpus	Ejecución	Precisión
Dev100	DLSI-Alicante-1	68.03
Dev500	DLSI-Alicante-1	57.27
Test	DLSI-Alicante-1	54.53
Test	DLSI-Alicante-2	52.11

Cuadro 1: Resultados obtenidos en la tarea de normalización.

[5]http://komunitatea.elhuyar.org/tweet-norm/files/2013/07/tweets-test-reference.txt

Los resultados son competitivos, teniendo en cuenta la dificultad de la tarea, pero no directamente comparables a los obtenidos en trabajos anteriores(Mosquera y Moreda, 2012),(Mosquera, Lloret, y Moreda, 2012) ya que si bien los objetivos generales del taller se podrían englobar dentro de la normalización de variantes léxicas hay ciertos aspectos tales como la restauración de mayúsculas/minúsculas, distinción de palabras en español de otros idiomas o la corrección de nombres propios que se podrían solapar con otras tareas como la corrección automática. Así mismo, se han tenido en cuenta no sólo variantes léxicas del español sino también palabras en otros idiomas pertenecientes a nombres propios o marcas (Ej. redbull por Red Bull).

Por otra parte, evaluando únicamente la precisión se puede beneficiar a los sistemas más conservadores que hayan detectado un número bajo de palabras fuera del vocabulario pero cuya normalización haya sido mayormente correcta. Una evaluación basada en precisión y cobertura permitiría realizar una evaluación global de los sistema propuestos en ambos niveles: detección y normalización de las palabras fuera del vocabulario.

5. Conclusiones y Trabajo Futuro

Este artículo presenta la contribución del grupo DLSI a la tarea de normalización de tweets del taller Tweet-norm. La herramienta de normalización TENOR ha obtenido resultados aceptables teniendo en cuenta el ámbito más amplio de la tarea que va más allá de la corrección de variantes léxicas informales. Así mismo, se ha tenido que integrar Freeling dentro del proceso de detección de palabras fuera del vocabulario, cuyo diccionario es de menor tamaño que el índice fonético empleado por TENOR, lo cual ha podido afectar a los resultados. Por otra parte, los problemas abordados en el taller son de gran relevancia a la hora de procesar textos de la Web 2.0 y servirán de referencia para mejorar el rendimiento del sistema propuesto en un trabajo futuro.

Bibliografía

Atserias, Jordi, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluis Padró, y Muntsa Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. En *Procee-*

dings of the 5th International Conference on Language Resources and Evaluation (LREC'06), páginas 48–55.

Aw, Aiti, Min Zhang, Juan Xiao, y Jian Su. 2006. A phrase-based statistical model for sms text normalization. *Proceedings of the COLING/ACL*, páginas 33–40.

Choudhury, Monojit, Rahul Saraf, Vijit Jain, Sudeshna Sarkar, y Anupam Basu. 2007. Investigation and modeling of the structure of texting language. En *Proceedings of the IJCAI-Workshop on Analytics for Noisy Unstructured Text Data*, páginas 63–70.

Gouws, Stephan, Donald Metzler, Congxing Cai, y Eduard Hovy. 2011. Contextual Bearing on Linguistic Variation in Social Media. *ACL Workshop on Language in Social Media (LSM)*.

Han, Bo y Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, páginas 368–378, Portland, Oregon, USA, June. Association for Computational Linguistics.

Hoang, Hieu, Alexandra Birch, Chris Callison-burch, Richard Zens, Rwth Aachen, Alexandra Constantin, Marcello Federico, Nicola Bertoldi, Chris Dyer, Brooke Cowan, Wade Shen, Christine Moran, y Ondrej Bojar. 2007. Moses: Open source toolkit for statistical machine translation. páginas 177–180.

Kaufmann, Joseph. 2010. Syntactic Normalization of Twitter Messages. *REU Site for Artificial Intelligence Natural Language Processing and Information Retrieval Research Project*, 2.

López, Veronica, Rubén San-Segundo, Roberto Martín, Julian David Echeverry, y Syaheera Lutfi. 2010. Sistema de traducción de lenguaje SMS a castellano. En *XX Jornadas Telecom I+D*, Valladolid, Spain, Septiembre.

Martí, Maria Antonia y Mariona Taulé. 2007. Cess-ece: corpus anotados del español y catalán. *Arena Romanistica. A new Nordic journal of Romance studies*, 1.

Mosquera, Alejandro, Elena Lloret, y Paloma Moreda. 2012. Towards facilitating the accessibility of web 2.0 texts through text normalisation. En *Proceedings of the LREC workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA) ; Istanbul, Turkey.*, páginas 9–14.

Mosquera, Alejandro y Paloma Moreda. 2012. Tenor: A lexical normalisation tool for spanish web 2.0 texts. En *Text, Speech and Dialogue - 15th International Conference (TSD 2012)*. Springer.

Philips, Lawrence. 2000. The double metaphone search algorithm. *C/C++ Users Journal*, 18:38–43, June.

Ratcliff, John W. y David E. Metzener. 1988. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13(7):46–72, Julio.

Shannon, Claude. E. 1948. A mathematical theory of communication. *The Bell Systems Technical Journal*, 27:379–423.