Introducción a la tarea compartida Tweet-Norm 2013: Normalización léxica de tuits en español*

Tweet Normalization Workshop at SEPLN 2013: An overview

Iñaki Alegria[1], Nora Aranberri[1], Víctor Fresno[2], Pablo Gamallo[3] Lluis Padró[4], Iñaki San Vicente[5], Jordi Turmo[4], Arkaitz Zubiaga[6]

(1) IXA. UPV/EHU (2) UNED (3) USC (4) UPC (5) Elhuyar (6) City University of New York
tweet-norm@elhuyar.com

Resumen: En este artículo se presenta una introducción a la tarea *Tweet-Norm 2013*: descripción, corpora, anotación, preproceso, sistemas presentados y resultados obtenidos.
Palabras clave: Normalización léxica, Twitter

Abstract: An overview of the shared task is presented: description, corpora, annotation, preprocess, participant systems and results.
Keywords: Tweet-normalization

1. Introducción

En la actualidad, la normalización lingüística de tuits (Han y Baldwin, 2011) supone una tarea de gran interés en diversos campos como, por ejemplo, la traducción automática y el análisis de sentimiento, dentro del procesamiento del lenguaje natural. La normalización de SMS y tuits en inglés ha generado gran interés recientemente; sin embargo, la normalización de este tipo de textos en español se ha estudiado poco.

Partiendo de esta carencia como base, varios grupos de investigación participantes en diversos proyectos hemos visto la necesidad de fomentar la investigación en este área, con el fin de facilitar y mejorar los resultados obtenidos con tareas subsiguientes. Así, organizamos un taller para llevar a cabo una tarea compartida de Normalización léxica de tuits en español, el cual es parte del programa de la conferencia SEPLN 2013 en Madrid. A su vez, este taller puede ser visto como complemento del *Taller de Análisis de Sentimientos en la SEPLN* (TASS)[1] organizado en 2012 y 2013.

Esta tarea ha conllevado un reto científico importante, y creemos que la competición entre los grupos de investigación ha creado un marco ideal para proporcionar un banco de pruebas de referencia, con el que se ha impulsado la aplicación de técnicas y algoritmos propuestos recientemente, y estudiar su mejora o adaptación. Así, los grupos participantes han podido evaluar y comparar métodos, algoritmos y recursos lingüísticos de los que disponen. En este artículo vamos a hacer un repaso de las características de la tarea, los corpus usados, el etiquetado del mismo y la forma de evaluación, así como un resumen de los sistemas que se han presentado a la evaluación. Detalles adicionales pueden ser consultados en la web oficial[2] y en el resto de artículos del workshop.

2. Trabajos relacionados

Una buena introducción al tema de normalización de tuits es el artículo (Eisenstein, 2013), donde se revisa el estado del arte en NLP sobre variantes SMS y tuit, y cómo la comunidad cientifica ha respondido por dos caminos: normalización y adaptación de herramientas.

Como se ha dicho, el artículo (Han y Baldwin, 2011) es una referencia importante en el campo de la normalización. Para el inglés ellos estudian un corpus de 449 tuits en el que identifican 254 tokens a ser normalizados. Para detectar las palabras fuera de diccionario (OOV) usan GNU aspell y, como

* Gracias a todos los miembros del Comité de Organización y a los proyectos *Tacardi*, *Xlike*, *Celtic*, *TextMESS2* y *Skater* por su colaboración.

[1] http://www.daedalus.es/TASS2013/about.php

[2] http://komunitatea.elhuyar.org/tweet-norm/

en nuestro caso, las identificaciones de tuits (@usuarios), los hashtags o etiquetas (#etiqueta) y los URLs son excluidos. Estudian la distribución de las formas a normalizar y proponen un sistema basado en 3 pasos: (1) generación del conjunto de confusión, donde para una palabra OOV generan los posibles candidatos; (2) identificación de las palabras a normalizar usando un clasificador, distinguiéndolas de las que deben permanecer inalteradas; (3) selección de candidatos. Evalúan los resultados comparándolos con los modelos *noisy-channel* y SMT obteniendo una precisión de alrededor del 75 %.

Sobre adaptación de herramientas es interesante el trabajo (Liu et al., 2011) que replantea el tema de reconocimiento de entidades nombradas en corpus de tuits. Para el español se ha prestado atención al análisis de sentimiento (Villena Román et al., 2013) en el marco del citado taller TASS pero apenas se ha publicado nada sobre normalización. Existen otros trabajos relacionados con normalización (Gomez-Hidalgo, Caurcel-Díaz, y del Rio, 2013) (Mosquera, Lloret, y Moreda, 2012) (Oliva et al., 2011) principalmente sobre mensajes SMS, pero que no abordan la normalización de tuits en su conjunto.

3. Descripción y características de la tarea

Hemos elegido el término normalización léxica porque la tarea se centra en normalizar palabras detectadas como no conocidas (abreviaturas, formas no normalizadas, palabras con letras repetidas...); intentando, en la medida de lo posible, aislar este problema de otros fenómenos como variantes sintácticas, de estilo etc. Por lo tanto, y en la misma línea que (Han y Baldwin, 2011), sólo serán tratadas las palabras que en el preproceso se consideran OOV. Además, estas palabras se evaluarán individualmente. Los sistemas deben decidir si proponen normalizarlas o mantenerlas como están, ya que pueden ser palabras que no se deben normalizar por ser palabras correctas (nuevas entidades nombradas, por ejemplo), escritas en otro idioma, etc. Desde la organización del taller se decidió anotar un conjunto de 600 tuits para distribuirlo anotado entre los participantes (para la adaptación y ajuste de sus sistemas), y otros 600, que se han mantenido en secreto, para la evaluación de los sistemas (ver sección 5).

3.1. Colección de documentos

Entre las múltiples opciones que ofrece la API de Twitter[3], se optó por descargar tuits geolocalizados, los cuales vienen marcados con las coordenadas desde donde cada tuit ha sido enviado. La API de Twitter, a su vez, permite descargar tuits para la zona geográfica especificada. Aprovechando esta característica, optamos por una zona amplia dentro de la península ibérica, evitando incluir zonas con lenguas cooficiales, para así aumentar la posibilidad de que un gran número de los tuits estuvieran escritos en español. Así, el área escogida abarca, aproximadamente, el rectángulo comprendido entre Guadalajara como extremo al noreste, y Cádiz como extremo al sudoeste. Tras almacenar los tuits geolocalizados enviados desde esa zona durante los días 1 y 2 de abril de 2013, obtuvimos una colección de un total de 227.855 tuits. A partir de esta gran colección, generamos dos subconjuntos aleatorios de 600 tuits, los cuales fueron distribuidos a los participantes, el primero como conjunto de entrenamiento, y el segundo como conjunto de test para la evaluación final. Los tuits restantes fueron distribuidos a los participantes, sin anotaciones manuales, por si consideraban conveniente hacer uso de él.

3.2. Preproceso

Se decidió distinguir dentro de los tuits las palabras fuera del diccionario (OOV) usando el analizador morfológico de la librería FreeLing (Padró y Stanilovsky, 2012). Se analizan los tweets con los módulos básicos (diccionario, sufijos, detector de números, fechas, etc.) y si la palabra no es reconocida por ninguno de ellos, se considera OOV.

Para ello, se usó una version adaptada del tokenizador, de forma que mantuviera como un solo token las palabras del tipo `@usuario` y `#etiqueta`, así como las direcciones de e-mail, URLs, y los *smileys* más frecuentes. Igualmente, se activó una instancia del módulo *usermap*, que aplica una batería de expresiones regulares a cada token, y asigna un análisis a los que cumplen alguna de ellas. De este modo, se descartan como OOVs dichos patrones, ya que obtienen un análisis.

A continuación, se aplicó un analizador morfológico básico, con los módulos por defecto, excepto el reconocedor de multipala-

[3] https://dev.twitter.com/docs/api

bras (para evitar aglutinación de varias palabras en un solo token), el reconocedor de entidades con nombre (dado que queremos mantenerlas como OOV), y el módulo de probabilidades léxicas (dado que aplica un guesser que asignaría al menos una etiqueta a todas las palabras).

Al final de este preproceso, las palabras que no han recibido ningún análisis de ningún módulo del morfológico se consideran OOVs.

4. Proceso de anotación

Durante la fase de anotación, se procedió a la anotación manual de las palabras identificadas por FreeLing como palabras OOV. Cada OOV podía ser etiquetada como *correcta*, *variante* o *NoES* (otro idioma) y en el segundo caso había que asignarle su forma normalizada. En el corpus de desarrollo tres expertos etiquetaron independientemente cada OOV y posteriormente se consensuaron las anotaciones definitivas. Durante este proceso se fue completando un manual. El corpus de test fue etiquetado independientemente por dos expertos que consensuaros posteriormente la anotación final.

Los criterios de anotación por los cuales se rigió el grupo de anotadores se recogieron en el Manual de anotación y se resumen de la siguiente manera:

Palabra incluida en diccionario RAE

En todo caso se anotará como correcta sin modificación alguna, aunque por su contexto se dedujera que dicha palabra no es la adecuada.

Palabra con categoría de nombre propio no incluida en diccionario RAE

Si es un acrónimo originalmente compuesto, todo con mayúscula o con alguna letra en minúscula, tanto la forma originaria como su forma totalmente en mayúsculas serán etiquetadas como correctas sin ninguna modificación (p.e., *CoNLL*, *CONLL*, *IBM* e *I.B.M.*).

Si no es un acrónimo, está formado por las letras requeridas y su inicial está en mayúsculas e incorpora los acentos requeridos, será etiquetada como correcta, ya sea un nombre propio en diminutivo, un apodo u otra forma alternativa de la entidad (p.e., *Tony*, *Anita*, *Yoyas*)

Si se expresa con alguna falta de ortografía o con alguna alteración no aceptada en los puntos anteriores, se anotará como variante y se especificará su forma correcta, según se define con dichas reglas. (p.e., *sanchez* → *Sánchez*, *tamagochi* → *Tamagotchi*, *abc* → *ABC*, *a.B.c.* → *A.B.C.*, *CONL* → *CONLL*)

Palabra no incluida en el diccionario RAE sin ser nombre propio

Si es un neologismo o extranjerismo compuesto correctamente según reglas de buena formación se etiquetará como correcta sin ninguna modificación. (p.e., *mouriñistas*, *retuitear*, *retweetear*)

Si es un diminutivo o superlativo compuesto correctamente según reglas de buena formación se etiquetará como correcta sin ninguna modificación. (p.e., *supergrande*)

Si se expresa con alguna falta ortográfica o alteración (repetición, eliminación, permutación de letras, etc), se etiquetará como variante y se especificará su forma correcta. (p.e., *horrooorr* → *horror*, *hacia* → *hacía*)

Si es una abreviatura o un acortamiento se etiquetará como variante, especificando su forma correcta. (p.e., *admin* → *administración*, *sr* → *señor*)

Si es una onomatopeya con alguna alteración (normalmente repetición de letras), de una o varias formas existente según la RAE, entonces se etiquetará como variante de todas esas formas. Si no existe en el diccionario RAE se anotará como correcta. (p.e., *aaaahhh* → *ah*, *jajajajas* → *ja*)

Si es una concatenación de palabras, entonces se etiquetará como variante y se especificará la secuencia correcta de palabras.

Si es una palabra (o cadena de palabras) de otro idioma o un emoticón se etiquetará como *NoEs*.

El manual describe las líneas generales de casos. Sin embargo, la casuística encontrada fue amplia e hicieron falta varias puestas en común para detallar las reglas y mantener la continuidad y rigurosidad de la anotación. El límite no siempre claro entre palabras extranjeras y préstamos ya aceptados en la lengua española, títulos de películas y series, y errores ortográficos intencionados fueron, entre otros, motivo de discrepancia antes de unificar anotaciones.

Por ejemplo:

- `El hashtag #7añosSLQH ocupó el sábado 30, la 3ª posición en el Top10 de los Trending Topics de Málaga`
- `que estafa de tablet`

- `Me dispongo a ver Game of Thrones.`
- `Habril luisma con h...`

Una dificultad adicional de la anotación, la cual añadió cierto grado de subjetividad a la tarea, radicó en la necesidad de interpretar los acortamientos y/o abreviaturas utilizados por los usuarios. Cuando el contexto no era suficiente para descifrar la intención del usuario, algo nada sorprendente dada la limitación de caracteres impuesta en los tuits, los anotadores se vieron en la tesitura de interpretar dicha intención y ofrecer la corrección acorde a ésta. Como podemos ver en este ejemplo

- `cariiii k no te seguia en twitter!!!mu fuerte!!!..yasoy tu fan....muak....se te exa d menos en el bk....sobreto en los cierres jajajajas`

la OOV *bk* es de libre interpretación, ya que podría tratarse del acortamiento de cualquier lugar de ocio. En este caso se optó por *Burger King*, considerada la opción más general y reusable. En ciertos casos se optó por incluir más de una posible corrección. La corrección de onomatopeyas, cuya intención no siempre es clara, también ha sido discutida:

- `me da igual JUUUM!!`

5. *Corpus anotados y medidas de evaluación*

5.1. Corpus de desarrollo y test

A partir del corpus inicial descrito en la sección 3.1 se han generado dos subconjuntos: uno compuesto por 500 tuits que constituye el corpus de desarrollo y otro compuesto por 600 tuits que constituye el corpus de evaluación. En el corpus de desarrollo fueron anotadas manualmente 653 palabras OOV, mientras que en el de evaluación se anotaron 724. Cabe mencionar que, debido a las restricciones de uso del API de Twitter[4], está prohibido redistribuir corpus que contiene información sobre usuarios. Por esta razón, ambos corpus fueron distribuidos a los participantes utilizando únicamente los identificadores de tuits. Cada participante podía bajar el contenido de dichos identificadores a través de búsquedas a la API de Twitter mediante el script *Twitid*[5].

[4]https://dev.twitter.com/terms/api-terms

[5]http://komunitatea.elhuyar.org/tweet-norm/iles/2013/06/download_tweets.py

Una vez finalizado el plazo de participación, comprobamos que los tuits que seguían públicamente disponibles en ese momento para generar el corpus de evaluación era menor al conjunto original. Así el corpus de evaluación que finalmente ha sido considerado consta de 562 tuits, un número que varía ligeramente con respecto al volumen inicial de 600 tuits.

La distribución de las tres categorías (0-correcta, 1-variante y 2-NoES) en los corpus de desarrollo y de evaluación se muestran en la tabla 1. Como se puede comprobar, la distribución de las tres categorías sobre el total de palabras OOV no varía significativamente en los dos corpus, lo que ha permitido a los participantes desarrollar sus sistemas comprobando su eficacia sobre un conjunto de datos comparable al que se ha ofrecido posteriormente para evaluar la tarea.

Corpus	**#OOV**	**0**	**1**	**2**
Desarrollo	653	497	93	63
Test	662	531	98	33

Tabla 1: Datos de los corpus. *Se reducen los 724 OOVs de test anotados a 662 debido al problema de disponibilidad de los tuits.*

5.2. Medidas de evaluación

La tarea consistió únicamente en la corrección de errores, y no en la clasificación de las distintas categorías de palabras OOV (0, 1 y 2). De esta manera se pretende evaluar exclusivamente la capacidad de corrección de los sistemas participantes, ya que una fase de clasificación previa introduciría un factor de acumulación de errores, haciendo más difícil evaluar el rendimiento de los sistemas. Por tanto, la evaluación sólo tiene en cuenta si la forma propuesta es correcta, en base a los siguientes criterios:

Correcta: si la forma original era correcta (categoría 0) o NoES (categoría 2) y no se ha realizado ninguna normalización, o si la forma original era una variante (categoría 1) y la normalización propuesta es correcta.

Errónea: En cualquier otro caso.

Como medida de evaluación para calcular los resultados oficiales se ha utilizado la precisión sobre el total de palabras OOV en el corpus de evaluación. La fórmula de la precisión mide el número de decisiones realiza-

das correctamente sobre el total de palabras OOV a tratar en el corpus de evaluación.

6. *Resultados y resumen de los sistemas*

Sobre 20 grupos inscritos inicialmente 13 participaron finalmente con sus respectivos sistemas; y sólo seis de ellos hicieron uso de la posibilidad de evaluar dos sistemas.

6.1. Resultados

La tabla 2 muestra los resultados de precisión de los trece grupos participantes. Además de estos resultados se muestran otros dos resultados a tener en cuenta como referencia de la tarea. Por un lado se ha calculado cuál sería el rendimiento mínimo de un sistema (*baseline*), dando como correctas todas las palabras OOV. Este sistema obtendría una precisión por debajo del 20 %. Por otro lado se ofrece el rendimiento máximo (*upper-line*) al que se podría aspirar con los sistemas presentados. El *upper-line* incluye todas aquellas palabras OOV que han sido correctamente corregidas por al menos uno de los sistemas participantes.

El anexo 1 muestra la lista de las palabras OOV (7,25 %, 39) que ningún sistema ha corregido. La lista incluye una casuística muy amplia: por ejemplo, *filosofia*/*Filosofía*, que requiere corrección ortográfica y mayúsculas; **yaa**/*allá*, que está muy lejos de su forma correcta en cuanto a similitud de cadena, y *ya* es a priori un candidato mucho más probable para esa forma.

6.2. Resumen de las técnicas y recursos utilizados

Destacan las buenas prestaciones del sistema de la RAE, que supera claramente al resto de los sistemas y supera el 78 % de precisión. La mayoría de los sistemas, sin embargo, están en un intervalo entre el 54 % y el 67 %. Se podría explicar la diferencia del mejor sistema por el tratamiento meticuloso de cada uno de los fenómenos posibles, la combinación estadística de los componentes y la calidad y cobertura de los recursos utilizados.

Los fenómenos a los que varios sistemas hacen frente explícitamente son los siguientes:

- Errores ortográficos habituales (h → 0).
- Cambios fonológicos habituales (k → c).
- Omisión de tildes (á → a).

Rank	Sistema	Prec1	Prec2
-	Upperline	0,927	-
1	RAE	0,781	-
2	Citius-Imaxin	0,663	0,662
3	UPC	0,653	-
4	Elhuyar	0,636	0,634
5	IXA-EHU	0,619	0,609
6	Vicomtech	0,606	-
7	UniArizona	0,604	-
8	UPF-Havas	0,548	0,491
9	DLSIAlicante	0,545	0,521
10	UniMelbourne	0,539	0,517
11	UniSevilla	0,396	-
12	UJaen-Sinai	0,376	-
13	UniCoruña	0,335	-
-	Baseline	0,198	-

Tabla 2: Precisión obtenida por los sistemas presentados.

- Omisiones de letras (principalmente vocales y letras finales, especialmente en participios). P.e. *encantao* → *encantado*.
- Uso de abreviaturas o reducción de las palabras a los primeros caracteres. P.e. *exam* → *examen*.
- Énfasis repitiendo letras (*Felicidadeeees* → *Felicidades*).
- Restauración de mayúscula (*felicidades* → *Felicidades*).
- Unión de palabras contiguas (yuxtaposición de palabras). P. e. *esque* → *es que*.
- Logogramas y pictogramas. (x → *por* 2 → *dos*).
- Onomatopeyas (*ahahahah* → *ah*).

Respecto a los léxicos utilizados se usan principalmente diferentes diccionarios de español (o correctores ortográficos o el propio Freeling[6] usado en el preproceso) para buscar propuestas normalizadas. Algunos sistemas utilizan diccionarios de inglés para detectar OOVs que no deben modificarse, Wikipedia [7] para añadir o detectar entidades nombradas, pequeños diccionarios de variantes y *slang* (en inglés existen más extensos) o listas de frecuencias a partir de corpus para detectar y normalizar cambios habituales propios de Internet/Twitter.

[6] http://nlp.lsi.upc.edu/freeling/
[7] es.wikipedia.org

También diversos corpus de español son usados para construir modelos de lenguaje. Son usados tantos corpus de propósito general como corpus de tuits. También un sistema ha utilizado la API de un buscador para filtrar términos multipalabra.

Respecto a herramientas podemos destacar los ya nombrados correctores ortográficos (aspell[8], hunspell[9], Jazzy[10]), que se usan también para obtener propuestas de normalización. Junto a ellos varios sistemas usan *foma*[11] para escribir, compilar en transductores. y aplicar reglas de transformación de grafemas/fonemas. En algún caso se han aprendido reglas de transformación basadas en modelos de lenguaje (compuestos grafemas/fonemas) (p.e. usando Phonetisaurus[12]). Para seleccionar entre las propuestas (además de frecuencias basadas en corpus) varios sistemas usan modelos de lenguaje de bigramas o trigramas de palabras (usando p. ej. OpenGrm[13] o SRILM[14])

6.3. Breve descripción de los sistemas

RAE (Porta y Sancho, 2013): Se basa en transductores de estados finitos con pesos que son combinados estadísticamente usando la composición en tres pasos (variantes, posibles variantes, modelo de lenguaje) . A partir de reglas generan transductores para prácticamente todos los fenómenos comentados además de un modelo de lenguaje (LM) basado en trigramas de palabras. Los recursos léxicos más reseñables son el diccionario DRAE, las 100.000 palabras inglesas más frecuentes del BNC, y un corpus de páginas web (Wacky).

Citius-Imaxin (Gamallo, Garcia, y Pichel, 2013): A partir de diversos recursos léxicos, generan dos tipos de candidatos, primarios y secundarios; los cuales son ordenados de diferentes maneras en el proceso de selección del mejor candidato. Escriben reglas para tres tipos de errores: mayúscula/minuscula, caracteres repetidos y errores ortográficos comunes. Utilizan una lista de normalización (principalmente obtenida del corpus de desarrollo), el DRAE y un diccionario de nombres propios obtenido de la Wikipedia. También utilizan un LM basado en un corpus de RSS periodísticos.

UPC (Ageno et al., 2013): Usan una batería de módulos (divididos en tres grupos; palabras sueltas, términos multipalabra y expresiones regulares) para generar diferentes propuestas de corrección para cada palabra desconocida. Usan *foma* para realizar búsquedas aproximadas de términos simples o multipalabra similares. La corrección definitiva se elige por votación ponderada según la precisión de cada módulo. Los recursos mencionados son: lista de acrónimos, lista de emoticones multicaracter y lista de onomatopeyas, diccionarios de español (con variantes) y de inglés y listas de nombres propios.

Elhuyar (Saralegi y San-Vicente, 2013): Usan una estrategia compuesta por dos pasos: generación de posibles candidatos de corrección y selección del candidato utilizando un modelo de lenguaje. Para la generación de candidatos además de la habitual distancia de edición tratan abreviaturas comunes, coloquialismos, caracteres repetidos e interjecciones. También restauración de mayúsculas y nombres propios. Usan SRILM para el LM de bigramas de palabras, entrenándolo con la Wikipedia (también para la lista de nombres propios) y un corpus de EFE.

IXA-EHU (Alegria, Etxeberria, y Labaka, 2013): Usa también *foma* para reglas que se aplican incrementalmente, para la mayoría de los fenómenos nombrados, pero a diferencia del sistema RAE no usa pesos, salvo para los cambios ortográficos que aprende automáticamente del corpus de desarrollo. Para este aprendizaje usa un modelo de lenguaje basado en grafemas aprendido del corpus de desarrollo (utilizando *Phonetisaurus*). El LM de palabras es de unigramas (frecuencia de las palabras) basado en corpus de tuits base vueltos a recuperar y filtrados con *Freeling* (también se usa para obtener los nombres propios más frecuentes). Un buscador de Internet es usado para filtrar los términos multipalabra propuestos.

Vicomtech (Ruiz, Cuadros, y Etchegoyhen, 2013): Usan reglas de preproceso, un modelo de distancias de edición adecuado al dominio y tres LM de 5-gramas de palabras, usando KenLM, para seleccionar candidatos de corrección según el contexto. Además de la distancia de edición adaptada con pesos

[8] http://aspell.net/
[9] http://hunspell.sourceforge.net/
[10] http://jazzy.sourceforge.net/
[11] https://code.google.com/p/foma/
[12] http://code.google.com/p/phonetisaurus/
[13] http://www.opengrm.org/
[14] http://www.speech.sri.com/projects/srilm/

usan aspell y hunspell como diccionario, listas de nombres propios (JRC Names y SAVAS), un corpus de tuits recolectado por ellos y un corpus extraído de *Europarl*. Hacen un interesante estudio de los casos de variantes.

UniArizona (Hulden y Francom, 2013): Estudian dos sistemas alternativos de reglas escritas por un experto o inducción de las mismas. Los resultados son algo mejores para el primer sistema. Para el primer método escriben reglas para ser compiladas en transductores sin pesos usando foma. Las reglas afrontan varios de los fenómenos mencionados (restauración de tildes, repeticiones de caracteres, errores ortográficos habituales y abreviaturas). Para el segundo método inducen pesos para los cambios. Las propuestas se ordenan usando un LM de unigramas (frecuencia de palabras). Para manipulación de pesos en los transductores usan Kleen.

UPF-Havas (Muñoz-García, Suárez, y Bel, 2013): Hacen uso de datos abiertos extraídos de recursos publicados en la Web desarrollados de manera colectiva, entre los que se encuentran la Wikipedia y un diccionario de SMS. No afronta específicamente la mayoría de los problemas enumerados, salvo las tildes y las mayúsculas. Realiza búsquedas en el diccionario de SMS y si no tiene éxito usa la primera propuesta del corrector Jazzy.

DLSIAlicante (Mosquera-López y Moreda, 2013): empleando la herramienta de normalizacion multilingüe TENOR, siguiendo una estrategia similar a la usada en SMS en inglés empleando técnicas de reconocimiento del habla, pero adaptada al español. Usan aspell ampliado con nombre de países como diccionario, y representan el léxico fonéticamente usando el algoritmo del metáfono adaptado al español. Para distancia entre palabras usan el algoritmo Gestalt y para ordenar las propuests un LM (basado en el corpus CESS-ESP).

UniMelbourne (Han, Cook, y Baldwin, 2013): Basándose en su experiencia para el inglés, construyen un léxico de normalizacion a partir de un corpus (compuesto de millones de tuits en español) utilizando similitud distribucional basada en distancia de edición/fonológica, y este léxico se combina con un diccionario slang de jerga de Internet en español (obtenido de dos sitios web).

UniSevilla (Cotelo-Moya, Cruz, y Troyano, 2013): Aparte de caracterizar la fuente de error/variación usan reglas de transformación (implementación propia) y distancia de edición para proponer normalización y detección de palabras en otros idiomas (basado en trigramas de caracteres). Usan el diccionario de español Libreoffice y dos pequeños diccionarios de emoticones y variantes en tuits (generados por ellos).

UJaen-Sinai (Montejo-Ráez et al., 2013): Para proponer formas normalizadas hacen una serie de conversiones a partir de lexicones de reemplazamiento (abreviaturas y onomatopeyas) y un corrector ortográfico (aspell enriquecido con nombres de ciudades, interjecciones, neologismos de Internet y otras entidades nombradas).

UniCoruña (Vilares, Alonso, y Vilares, 2013): Es un sistema conceptualmente sencillo y flexible que emplea pocos recursos (diccionario SMS, tratamiento de onomatopeyas, repeticiones, diacríticos y errores ortográficos) y que aborda el problema desde un punto de vista léxico.

7. Conclusiones

El taller Tweet-Norm-2013 ha sido un primer paso académico conjunto para estudiar y mejorar el problema de normalización de tuits en español. La participación de 13 sistemas demuestra el interés en el tema. Es de resaltar la diversidad de procedencia de los participantes y la variedad de recursos utilizados.

A la espera de un análisis todavia más detallado de los resultados creemos que los corpus desarrollados y las publicaciones realizadas ayudarán a la mejora de los resultados en el futuro.

Desde los participantes se han recibido propuestas de mejora sobre ciertos aspectos del preproceso que pueden ser mejorados (entidades comunes que se han marcado como OOV) y algunos casos de anotación que pueden ser discutibles.

Los corpus anotados se pondrán en breve plazo a libre disposición de toda la comunidad científica (consultar el sitio oficial: komunitatea.elhuyar.org/tweet-norm/).

Creemos que en el futuro una tarea similar puede ser planteada, aunque creemos necesario algún tipo de evaluación combinada con otras tareas (traducción, análisis de sentimiento...). Además sería interesante dar un paso más allá de la normalización léxica, y afrontar también la normalización sintáctica.

Bibliografía

Ageno, Alicia, Pere R. Comas, Lluís Padró, y Jordi Turmo. 2013. The talp-upc approach to tweet-norm 2013. En *Proc. of the Tweet Normalization Workshop at SEPLN 2013. IV Congreso Español de Informática*.

Alegria, Iñaki, Izaskun Etxeberria, y Gorka Labaka. 2013. Una cascada de transductores simples para normalizar tweets. En *Proc. of the Tweet Normalization Workshop at SEPLN 2013. IV Congreso Español de Informática*.

Cotelo-Moya, Juan M., Fermín L. Cruz, y Jose A. Troyano. 2013. Resource-based lexical approach to tweet-norm task. En *Proc. of the Tweet Normalization Workshop at SEPLN 2013. IV Congreso Español de Informática*.

Eisenstein, Jacob. 2013. What to do about bad language on the internet. En *Proceedings of NAACL-HLT*, páginas 359–369.

Gamallo, Pablo, Marcos Garcia, y José Ramom Pichel. 2013. A method to lexical normalisation of tweets. En *Proc. of the Tweet Normalization Workshop at SEPLN 2013. IV Congreso Español de Informática*.

Gomez-Hidalgo, Jose M., Andrés A. Caurcel-Díaz, y Yovan Iñiguez del Rio. 2013. Un método de análisis de lenguaje tipo sms para el castellano. *Linguamática*, 5(1):31–39.

Han, Bo y Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. En *ACL*, páginas 368–378.

Han, Bo, Paul Cook, y Timothy Baldwin. 2013. unimelb: Spanish text normalisation. En *Proc. of the Tweet Normalization Workshop at SEPLN 2013. IV Congreso Español de Informática*.

Hulden, Mans y Jerid Francom. 2013. Weighted and unweighted transducers for tweet normalization. En *Proc. of the Tweet Normalization Workshop at SEPLN 2013.IV Congreso Español de Informática*.

Liu, Xiaohua, Shaodian Zhang, Furu Wei, y Ming Zhou. 2011. Recognizing named entities in tweets. En *ACL*, páginas 359–367.

Montejo-Ráez, Arturo, Manuel Díaz-Galiano, Eugenio Martínez-Cámara, Teresa Martín-Valdivia, Miguel A. García-Cumbreras, y Alfonso Ureña-López. 2013. Sinai at twitter-normalization 2013. En *Proc. of the Tweet Normalization Workshop at SEPLN 2013.IV Congreso Español de Informática*.

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-López, Alejandro y Paloma Moreda. 2013. Dlsi en tweet-norm 2013: Normalización de tweets en español. En *Proc. of the Tweet Normalization Workshop at SEPLN 2013. IV Congreso Español de Informática*.

Muñoz-García, Oscar, Silvia Vázquez Suárez, y Nuria Bel. 2013. Exploiting web-based collective knowledge for micropost normalisation. En *Proc. of the Tweet Normalization Workshop at SEPLN 2013. IV Congreso Español de Informática*.

Oliva, Jesús, José I. Serrano, María D. Del Castillo, y Ángel Iglesias. 2011. Sms normalization: combining phonetics, morphology and semantics. En *Advances in Artificial Intelligence*. Springer, páginas 273–282.

Padró, Lluís y Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2012). Istanbul.*

Porta, Jordi y José Luis Sancho. 2013. Word normalization in twitter using finite-state transducers. En *Proc. of the Tweet Normalization Workshop at SEPLN 2013. IV Congreso Español de Informática*.

Ruiz, Pablo, Montse Cuadros, y Thierry Etchegoyhen. 2013. Lexical normalization of spanish tweets with preprocessing rules, domain-specific edit distances, and language models. En *Proc. of the Tweet Normalization Workshop at SEPLN 2013. IV Congreso Español de Informática*.

Saralegi, Xabier y Iñaki San-Vicente. 2013. Elhuyar at tweet-norm 2013. En *Proc. of the Tweet Normalization Workshop at SEPLN 2013.IV Congreso Español de Informática*.

Vilares, Jesus, Miguel A. Alonso, y David Vilares. 2013. Prototipado rápido de un sistema de normalización de tuits: Una aproximación léxica. En *Proc. of the Tweet Normalization Workshop at SEPLN 2013. IV. Congreso Español de Informática*.

Villena Román, Julio, Sara Lana Serrano, Eugenio Martínez Cámara, y José Carlos González Cristóbal. 2013. Tass-workshop on sentiment analysis at sepln.

Anexo I: Listado de palabras no corregidas

A continuación se detallan las variantes del corpus de test que ningún sistema ha propuesto corregido correctamente, junto la normalización anotada.

```
FYQ Física_y_química
sisiii sí_sí
yaa allá
picolos picoletos
nainonainonahh nainonainoná
gordys gorditas
JUUUM hum
Tuitutil TuitÚtil
crst Cristo
mencantaba me_encantaba
diitaas diítas
soo eso
queeee qué
Teinfiniteamo Te_amo_infinitamente
aber a_ver
Hum Humedad
L. l.
Muchomuchacho Mucho_Muchacho
Hojo Jo
jonaticas jonáticas
gafis gafitas
her hermano|hermana
MIAMOR mi_amor
guapii guapita
WAPAHHH guapa
EAEA ea_ea
Acho Macho
tirantitas tirantitos
HMYV MHYV
filosofia Filosofía
nah nada
FAV favorito
JIIIIIIIIOLE Olé
Fotazo fotaza
gor gorda|gordo
coner con_el
shh sí|sé
primera+ primera_más
salobreja Salobreja
```