CAMPUS
DE EXCELENCIA
INTERNACIONAL

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INFORMÁTICOS

MÁSTER UNIVERSITARIO EN INTELIGENCIA ARTIFICIAL

TRABAJO DE FIN DE MÁSTER:

TWEETSC: CORRECTOR DE TEXTO PARA TWITTER

JAVIER MORENO VEGA

TUTOR DE PROYECTO:
OSCAR CORCHO GARCÍA

CO-TUTOR DE PROYECTO:
VÍCTOR RODRÍGUEZ DONCEL

`http://tweetsc.github.io`

23 de mayo de 2018

Índice

1. Introducción

1.1. Motivación

Los nuevos sistemas de comunicación como la mensajería instantánea, chats, redes sociales han generado un uso diferente de los idiomas en estos ámbitos, llamado lenguaje tipo chat (N. and Martell, 2007). Una de estas redes sociales y en la que este trabajo va a centrarse es Twitter. En esta red social predomina el Uso de emoticonos, repetición de vocales o eliminación de las mismas, uso abusivo de mayúsculas o asusencia, siglas de expresiones populares; lo que dificulta el análisis de los textos. Las ventajas que ofrece esta red social para investigar sobre ella son la cantidad de datos en tiempo real y su fácil acceso.

Uno de los principales problemas a la hora de analizar textos procedentes de las redes sociales son los errores gramaticales que suelen contener, así como la presencia de elementos propios de este tipo de foros que requieren de un procesamiento especial (i.e. hashtags, formas de mencionar a otros usuarios o emoticonos y expresiones habituales en las redes). Además, la limitación en el número de caracteres existente en Twitter la convierte en un caso singular dentro de las redes sociales, ya que los usuarios tienden a adaptar su forma de escribir a dicha limitación, omitiendo palabras y creando abreviaturas que dificultan el uso de herramientas genéricas de procesamiento del lenguaje, especialmente a la hora de realizar tareas como el Análisis de Sentimientos.

Los usuarios en twitter tienden a cometer errores tipográficos, abreviaciones, sustituciones fonéticas y estructuras no gramaticales en los mensajes cortos de texto, causando problemas en las herramientas de análisis. Esto es lo que se consideran palabras mal formadas y la detección de las palabras mal formadas es difícil debido al contexto ruidoso. El objetivo es normalizar estas palabra mal formadas.
A parte de un uso puramente de investigación, este tipo de trabajo también es beneficioso para un estudio de marcas o personas y sobre lo que las persones opinan sobre ello en las redes social, ya que sin el proceso de normalización y análisis de sentimientos estariamos ante millones de datos que costarían mucho trabajo analizar de una forma automática.

1.2. Objetivos

El objetivo principal de este trabajo es la creación de un corrector que "normalice"tweets en español.
Para cumplir con este objetivo principal se ha dividido en los siguientes subobjetivos.

- Corregir tweets (palabras y gramaticalmente)
- Procesar emoticonos deduciendo su significado en el contexto de análisis de sentimientos
- Expandir hashtags
- Procesar las conversaciones de los tweets, así como las URLs o imágenes para añadir contexto

Estos subojetivos se cumplirán con su implementación en un módulo software que además estará disponible en una aplicación web (twe).

1.3. Resumen del documento

Esta memoria explica todo el trabajo desarrollado entrando en detalle en el estado del arte y el módulo software desarrollado.

Primero se expone el estado del arte desarrollado sobre el tema de la corrección de textos, específicamente en twitter y en español.

En segundo lugar se presenta el análisis realizado, atendiendo a: metodologías de desarrollo utilizadas, análisis de requisitos, solución propuesta.

Posteriormente se plantea la implementación que se ha seguido entrando en detalle en cómo usar el software y en detalles técnicos.

Por último se muestran unas conclusiones, incluyendo un resumen del trabajo desarrollado y los objetivos conseguidos.

2. Estado del arte

2.1. Introducción

En la actualidad, la normalización lingüística de tweets (Han and Baldwin, 2011) supone un campo de gran interés y en donde la mayoría de trabajos se han realizado sobre textos en inglés y pocos en español. Además no hay ningún trabajo en donde se incluya, dentro de la normalización de tuits, el estudio de los hashtags o etiquetas y los emoticonos, y su contexto. Una 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 tweets, y cómo la comunidad científica ha respondido por dos caminos: normalización y adaptación de herramientas.

2.2. Normalización

El modelo de canal ruidoso (Shannon, 1948) ha sido tradicionalmente la primera aproximación a la normalización de textos. Supone que el texto mal formado es T y su forma normalizada es S, por lo que hay que encontrar: arg max $P(S|T)$, calculando arg max $P(T|S)P(S)$, P(S) es el modelo del lenguaje y $P(T|S)$ es el modelo de error. (Brill and Moore, 2000) caracterizan el modelo de error calculando el producto de operaciones de probabilidad en partes de cadenas de caracteres. (Toutanova and Moore, 2002) mejoraron el modelo incorporando información de la pronunciación. (Choudhury et al., 2007) modela el proceso de generación de texto a nivel de palabra para mensajes SMS considerando las abreviaturas grafémicas/fonéticas y los errores tipográficos involuntarios como transiciones de estado ocultas del modelo de Markov (HMM) y emisiones, respectivamente. (Cook and Stevenson, 2009) expandieron el modelo de error introduciendo inferencias de diferentes procesos de formación erróneos, de acuerdo con la distribución de errores muestreada.

Mientras el modelo de canal ruidoso es apropiado para normalización de textos, es difícil aproximar la normalización con exactitud, además estos métodos ignoran el contexto alrededor del OOV, el cual ayuda a resolver ambigüedades. La traducción automática estadística (SMT) se ha propuesto como un medio de normalización de texto sensible al contexto, al tratar el texto mal formado como el idioma de origen, y la forma estándar como el idioma de

destino. Por ejemplo (Aw et al., 2009). Normalización de textos como un problema de reconocimiento de voz (Kobus et al., 2008). (Beaufort et al., 2002) métodos de estado finitos combinando las ventajas de SMS y el modelo de canal ruidoso. (Kaufmann and Kalita, 2010) usan un enfoque de traducción automática con un preprocesador para la normalización sintáctica (en lugar de léxica).

El problema de estos trabajos anteriores es que requieren datos de entrenamiento anotados a gran escala, lo que limita su adaptabilidad a nuevos dominios o idiomas, mientras que el trabajo (Han and Baldwin, 2011) no. Este trabajo es una buena referencia en el campo de la normalización de tuits en inglés. En donde para detectar palabras fuera de diccionario (OOV) utilizan GNU aspell, y los usuarios (@usuario), los hashtags y las URLs son excluidas de la normalización. La normalización tiene relación con los correctores de texto (Peterson, 1980) pero difiere en que las palabras mal formadas en los mensajes de texto suelen ser intencionadas, para ahorrar caracteres, como identidad social, o debido a la convención en este subgénero de texto. La detección de las palabras mal formadas es difícil debido al contexto ruidoso. El objetivo es normalizar estas palabra mal formadas, además muchas palabras mal formadas son ambiguas y requieren el contexto para poder normalizarlas.

2.3. Adaptación de herramientas

En vez de adaptar el texto a herramientas de análisis otro de los caminos a seguir es adaptar las herramientas de análisis al texto. Destacan los trabajos de reconocimiento de voz (Gimpel et al., 2011); (Owoputi et al., 2013), reconocimiento de entidades (Finin et al., 2010); (Ritter et al., 2011); (Liu et al., 2011), análisis gramatical (Foster et al., 2011), modelización de diálogos (Ritter et al., 2010) y resumen (Sharifi et al., 2010). El trabajo (Liu et al., 2011) sobre NER (reconocimiento de entidades) replantea el tema de reconocimiento de entidades nombradas en corpus de tuits. Combina un clasificador KNN con CRF (Conditional Random Fields).

2.4. Normalización en español

Una introducción a la normalización de tuits en español es (Alegria et al., 2013)(Alegria et al., 2015). Utiliza la herramienta Freeling (Freeling) para

detectar palabras OOV. Uno de los sistemas de normalización de tuits en español, que participó en Tweet-Norm 2013 (Alegria et al., 2013), es (Ruiz et al., 2013)(Vicomtech), que usa reglas de preproceso, un modelo de distancias de edición adecuado al dominio y modelos de lengua para seleccionar candidatos de corrección según el contexto. El sistema obtuvo resultados superiores a la media en la tarea (Alegria et al., 2013)(Tweet-Norm). Una mejora a este trabajo por los mismos autores es (Ruiz et al., 2014). El trabajo que mejores resultados obtuvo en Tweet-Norm 2013 fue RAE (Gamallo et al., 2013b), consiste en transductores que se aplican a tokens OOV. Los transductores implementan modelos lingüísticos de variación que generan conjuntos de candidatos de acuerdo con un léxico. Se usa un modelo de lenguaje estadístico para obtener la secuencia de palabras más probable. En el trabajo (Cotelo et al., 2015) hace uso de una combinación de varios "módulos expertos" independientes, cada uno especializado en una tarea concreta de la normalización de tuits, en lugar de centrarse en una sola técnica. En este trabajo además realiza un estado del arte actual de la normalización de tuits y en concreto para el idioma español. Otros trabajos sobre normalización en español son (Gomez-Hidalgo et al., 2013) (Mosquera et al., 2012) (Oliva et al., 2011) principalmente sobre mensajes SMS, pero que no abordan la normalización de tuits en su conjunto.

2.5. Análisis de sentimientos

Un campo muy relacionado con la normalización de tuits es el análisis de sentimientos y un trabajo que realiza un estudio sobre técnicas de análisis de sentimientos de tuits en español es (Anta et al., 2013). El trabajo (Gamallo et al., 2013a) se centra en una técnica Naive-Bayes para el análisis de sentimientos en tuits en español.

3. Análisis y diseño

3.1. Metodología de desarrollo

3.2. Análisis de requisitos

3.3. Solución propuesta

4. Implementación

4.1. Introducción

4.2. Cómo usarlo

4.3. Documentación del código

5. Evaluación

5.1. Metodología

5.2. Corpus

5.3. Gold Standard

5.4. Experimentos

6. Conclusiones y vías futuras

7. Apéndices

7.1. Apéndice A: Bibliografía

Referencias

Tweetsc web. `https://tweetsc.github.io`.

Iñaki Alegria, Nora Aranberri, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. Introducción a

la tarea compartida tweet-norm 2013: Normalización léxica de tuits en español, 2013.

Iñaki Alegria, Nora Aranberri, Pere R. Comas, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. Tweet-norm: a benchmark for lexical normalization of spanish tweets, 2015.

Antonio Fernández Anta, Luis Núñez Chiroque, Philippe Morere, and Agustín Santos. Sentiment analysis and topic detection of spanish tweets: A comparative study of nlp techniques, 2013.

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. A phrase-based statistical model for sms text normalization, 2009.

Richard Beaufort, Sophie Roekhaut, Louise-Amelie Cougnon, and Cedrick Fairon. A hybrid rule/model-based finite-state framework for normalizing sms messages, 2002.

Eric Brill and Robert C. Moore. An improved error model for noisy channel spelling correction, 2000.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. Investigation and modeling of the structure of texting language, 2007.

Paul Cook and Suzanne Stevenson. An unsupervised model for text message normalization, 2009.

J.M. Cotelo, F.L. Cruz, J.A. Troyano, and F.J. Ortega. A modular approach for lexical normalization applied to spanish tweets, 2015.

Jacob Eisenstein. What to do about bad language on the internet, 2013.

Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. Annotating named entities in twitter data with crowd-sourcing, 2010.

Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. From news to comment: Resources and benchmarks for parsing the language of web 2.0, 2011.

Freeling. Freeling. `http://nlp.lsi.upc.edu/freeling/`.

Pablo Gamallo, Marcos García, and Santiago Fernández-Lanza. Tass: A naive-bayes strategy for sentiment analysis on spanish tweets, 2013a.

Pablo Gamallo, Marcos García, and Santiago Fernández-Lanza. Word normalization in twitter using finite-state transducers, 2013b.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael, Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for twitter: annotation, features, and experiments, 2011.

Jose M. Gomez-Hidalgo, Andrés A. Caurcel-Díaz, and Yovan Iñiguez del Rio. Un método de análisis de lenguaje tipo sms para el castellano, 2013.

Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: Makn sens a twitter, 2011.

Joseph Kaufmann and Jugal Kalita. Syntactic normalization of twitter messages, 2010.

Catherine Kobus, Franois Yvon, and Graldine Damnati. Transcrire les sms comme on reconnat la parole, 2008.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. Recognizing named entities in tweets, 2011.

Mosquera, Alejandro, Elena Lloret, and Paloma Moreda. Towards facilitating the accessibility of web 2.0 texts through text normalisation, 2012.

Forsyth Eric N. and Craig H. Martell. Lexical and discourse analysis of online chat dialog, 2007.

Jesús Oliva, José I. Serrano, María D. Del Castillo, and Angel Iglesias. Sms normalization: combining phonetics, morphology and semantics, 2011.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters, 2013.

James L. Peterson. Computer programs for detecting and correcting spelling errors., 1980.

Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations, 2010.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: an experimental study, 2011.

Pablo Ruiz, Montse Cuadros, and Thierry Etchegoyhen. Lexical normalization of spanish tweets with preprocessing rules, domain-specific edit distances, and language models, 2013.

Pablo Ruiz, Montse Cuadros, and Thierry Etchegoyhen. Lexical normalization of spanish tweets with rule-based components and language models, 2014.

Claude Elwood Shannon. A mathematical theory of communication, 1948.

Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. Summarizing microblogs automatically, 2010.

Kristina Toutanova and Robert C. Moore. Pronunciation modeling for improved spelling correction, 2002.

Tweet-Norm. Tweet-norm. `http://komunitatea.elhuyar.eus/tweet-norm/`.

Sistema Vicomtech. Sistema vicomtech. `https://github.com/pruizf/tweet-norm-es`.

7.2. Apéndice B: Glosario de Términos