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`

20 de mayo de 2018

Índice

1. Introducción

1.1. Motivación

1.2. Objetivos

1.3. Resumen del documento

2. Estado del arte

En la actualidad, la normalización lingüística de tuits (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 tuit, y cómo la comunidad científica ha respondido por dos caminos: normalización y adaptación de herramientas. El artículo (Han and Baldwin, 2011) 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. En adaptación de herramientas es interesante el trabajo [9] que replantea el tema de reconocimiento de entidades nombradas en corpus de tuits. Combina un clasificador KNN con CRF (Conditional Random Fields).

Una introducción a la normalización de tuits en español es (Alegria et al., 2013)[6]. Utiliza la herramienta Freeling [10] 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), 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)[11]. Una mejora a este trabajo por los mismos autores es (Ruiz et al., 2014). 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 nor-

malización de tuits y en concreto para el idioma español.

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 [12]. El trabajo [13] se centra en una técnica Naive-Bayes para el análisis de sentimientos en tuits en español.

Sistemas que participaron en la tarea Tweet-Norm 2013 y que son públicos: Vicomtech (Ruiz et al., 2013) (Ruiz et al., 2014) [14] RAE (Mejores resultados) [15]

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. Readme Github

4.3. Javadoc

5. Evaluación

5.1. Metodología

5.2. Corpus

5.3. Goal Standard

5.4. Experimentos

6. Apéndices

6.1. Apéndice A: Bibliografía

Referencias

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.

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.

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

6.2. Apéndice B: Glosario de Términos