Trabajo de Fin de Máster

TweetSC: Twitter Spell Checker

Supervisors

Óscar Corcho García

Víctor Rodríguez Doncel

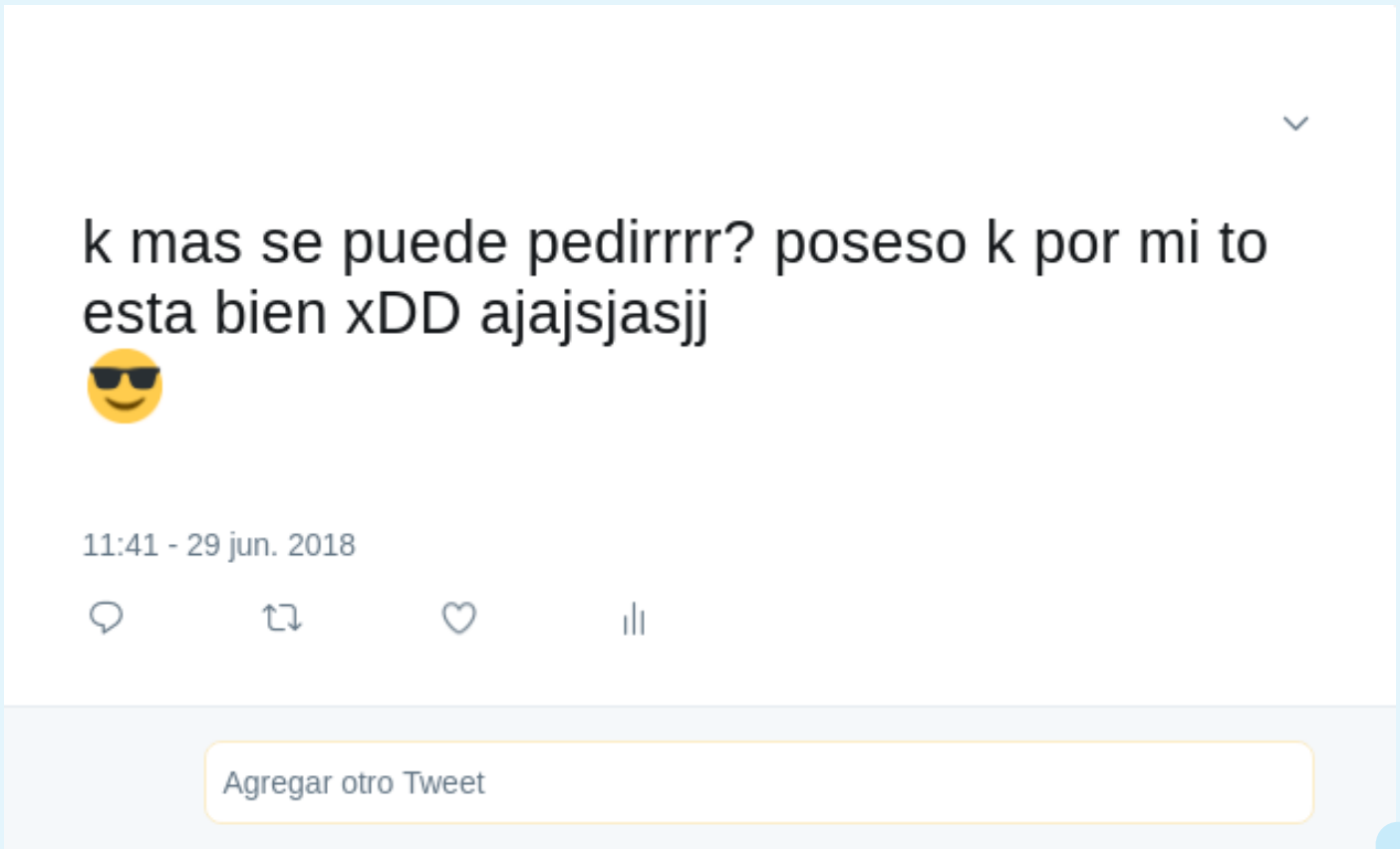
Autor

Javier Moreno Vega

July 18, 2018

Motivación

Tweet mal escrito



Motivación (II)

- Nuevos sistemas de comunicación (mensajería instantánea, chats, redes sociales, ...)
- Uso diferente de los idiomas
- Problemas a la hora de analizar textos
- Red social donde predominan: emoticonos, repetición de vocales, uso abusivo de mayúsculas,...

Motivación (II)

- Nuevos sistemas de comunicación (mensajería instantánea, chats, redes sociales, ...)
- Uso diferente de los idiomas
- Problemas a la hora de analizar textos
- Red social donde predominan: emoticonos, repetición de vocales, uso abusivo de mayúsculas,...
- Twitter

Motivación (III): Twitter

- Ofrece una gran cantidad de datos de libre acceso
- Uso de palabras propias de la red social (hashtags, RT, etc.)
- Limitación de 280 caractéres

Motivación (III): Twitter

- Ofrece una gran cantidad de datos de libre acceso
- Uso de palabras propias de la red social (hashtags, RT, etc.)
- Limitación de 280 caractéres
- Abreviaturas, omisión de palabras, abreviaciones, sustituciones fonéticas o estructuras no gramaticales

Motivación (III): Twitter

- Ofrece una gran cantidad de datos de libre acceso
- Uso de palabras propias de la red social (hashtags, RT, etc.)
- Limitación de 280 caractéres
- Abreviaturas, omisión de palabras, abreviaciones, sustituciones fonéticas o estructuras no gramaticales
- Palabras mal formadas o OOV (*Out-Of-Vocabulary*)

Objetivos

Objetivo principal

Creación de un corrector que “normalice” tweets en español.

Objetivos

Objetivo principal

Creación de un corrector que "normalice" tweets en español.

Subobjetivos

- Acceso a la API de Twitter para obtener tweets
- Tokenización de tweets
- Detección entre los tokens las palabras fuera del vocabulario (OOV)
- Anotar el tipo de palabra OOV
- Corrección de palabras OOV

Estado del arte

- Campo de gran interés
- Mayoría de trabajos sobre textos en inglés
- Introducción al tema de la normalización de tweets es el trabajo de Eisenstein, 2013
- Normalización y adaptación de herramientas

Estado del arte (II): Normalización

- Modelo del canal ruidoso (Shannon 1948), se definen modelo del lenguaje y modelo de error
- Brill y Moore (2000) caracterizan el modelo de error
- Toutanova y Moore (2002) añaden información de la pronunciación
- Choudhury et al. (2007) normalización SMS usando el modelo de Markov (HMM)
- Cook y Stevenson (2009) expanden el modelo de error

Estado del arte (II): Normalización

- Modelo del canal ruidoso (Shannon 1948), se definen modelo del lenguaje y modelo de error
- Brill y Moore (2000) caracterizan el modelo de error
- Toutanova y Moore (2002) añaden información de la pronunciación
- Choudhury et al. (2007) normalización SMS usando el modelo de Markov (HMM)
- Cook y Stevenson (2009) expanden el modelo de error
- Pero modelo del canal ruidoso ignora el contexto del OOV

Estado del arte (III): Normalización

- Surgen alternativas que consideran el contexto
- Beaufort et al. (2002) métodos de estados finitos combinando las ventajas del modelo del canal ruidoso y el SMS
- Kobus et al. (2008) reconocimiento de voz
- Aw et al. (2009) traducción automática estadística (SMT)
- Kaufmann y Kalita (2010) SMT con preprocesado para normalizacin sintáctica

Estado del arte (III): Normalización

- Surgen alternativas que consideran el contexto
- Beaufort et al. (2002) métodos de estados finitos combinando las ventajas del modelo del canal ruidoso y el SMS
- Kobus et al. (2008) reconocimiento de voz
- Aw et al. (2009) traducción automática estadística (SMT)
- Kaufmann y Kalita (2010) SMT con preprocesado para normalizacin sintáctica
- Pero todos estos trabajos requieren datos de entrenamiento

Estado del arte (IV): Adaptación de herramientas

- En vez de adaptar el texto a las herramientas de análisis
- Adaptar las herramientas al texto
- Reconocimiento de voz: Gimpel et al. (2011) y Owoputi et al. (2013)
- Reconocimiento de entidades: Finin et al. (2010), Ritter et al. (2011) y Liu et al. (2011)
- Análisis gramatical: Foster et al. (2011)
- Modelización de diálogos: Ritter et al. (2010)
- Resumen automtico de textos: Sharifi et al. (2010)
- Reconocimiento de entidades nombradas (NER): Enron (Minkov, 2005) y CoNLL03 (Tjong, 2003)
- NER sobre tweets: Finin et al. (2010)

Estado del arte (V): Adaptación de herramientas

- Desambiguación léxica o etiquetado gramatical (POST)
- Palabras que pueden ser asignadas a más de una clase morfológica a más de un *part-of-speech* (PoS)
- Trabajo ms importante y para español: SWPoST de Sánchez-Villamil et al. (2004)

Estado del arte (VI): Normalización en español

- Tweet-Norm 2013
- Dos categoras de candidatos:
  - Generación de candidatos + LM
  - Transductores o FSTs (*Finite State Transducers*)
- Mejor participante: Sistema RAE por los autores Gamallo et al. (2013) con un accurancy de 0.781
- Sistema basado en FSTs
- Otros trabajos: Mosquera et al. (2012) candidatos con indexación fonética y Olivia et al. (2011)
- Sobre mensajes SMS
- Tokenización en español: Gomez-Hidalgo et al. (2013)

Estado del arte (VI): Word2Vec

- Palabras como unidades atómicas no hay noción de similaridad entre palabras
- Aparecen las representaciones continuas de palabras
- Muchos tipos diferentes de modelos: *Latent Semantic Analysis* (LSA) y *Latent Dirichlet Allocation* (LDA)
- Distribuciones de palabras aprendidas por redes neuronales.
- Las redes neuronales basadad en LM mejoran los modelos N-grama
- Dos modelos de redes neuronales destacan basados en LM: *Feed Forward Neural Net Language Model* (NNLM) y *Recurrent Neural Net Language Model* (RNNLM)
- Costosos con grandes candidades de datos

Estado del arte (VI): Word2Vec

- Palabras como unidades atómicas no hay noción de similaridad entre palabras
- Aparecen las representaciones continuas de palabras
- Muchos tipos diferentes de modelos: *Latent Semantic Analysis* (LSA) y *Latent Dirichlet Allocation* (LDA)
- Distribuciones de palabras aprendidas por redes neuronales.
- Las redes neuronales basadad en LM mejoran los modelos N-grama
- Dos modelos de redes neuronales destacan basados en LM: *Feed Forward Neural Net Language Model* (NNLM) y *Recurrent Neural Net Language Model* (RNNLM)
- Costosos con grandes candidades de datos
- Word2Vec, dos modelos: CBOW y Skip-gram

Estado del arte (VII): Word2Vec

- Técnicas de representación continua de vectores representan cada palabra como vector distinto
- Se ignora la estructura iterna de las palabras

Estado del arte (VII): Word2Vec

- Técnicas de representación continua de vectores representan cada palabra como vector distinto
- Se ignora la estructura iterna de las palabras
- Bojanowski et al. (2017) desarrollado por Facebook y denominado fastText

Estado del arte (VII): Word2Vec

- Técnicas de representación continua de vectores representan cada palabra como vector distinto
- Se ignora la estructura iterna de las palabras
- Bojanowski et al. (2017) desarrollado por Facebook y denominado fastText
- Nuevo enfoque donde cada palabra se representa como una bolsa de caracteres n-gramas.

Solución propuesta

TweetSC (Tweet Spell Checker)

Proceso iterativo sobre el tweet a normalizar que se puede dividir en seis fases: tokenización, reglas de preproceso, detección de OOVs, generación de candidatos para cada OOV, ranking de candidatos y postproceso.

Solución propuesta

TweetSC (Tweet Spell Checker)

Proceso iterativo sobre el tweet a normalizar que se puede dividir en seis fases: tokenización, reglas de preproceso, detección de OOVs, generación de candidatos para cada OOV, ranking de candidatos y postproceso.

TweetSC. Sistema web

Aplicación web con acceso a la API de Twitter para obtener los tweets mediante consultas introducidas en un formulario de la aplicación web.

Solución propuesta (II)

Tokenización

- Primera fase también de los analizadores léxicos en los compiladores
- Genera una lista de tokens que pasan a la siguiente fase

Solución propuesta (II)

Tokenización

- Primera fase también de los analizadores léxicos en los compiladores
- Genera una lista de tokens que pasan a la siguiente fase

Reglas de preprocesado

- Se aplican las reglas una a una a los tokens de entrada
- Los tokens que acepten alguna regla se eliminan, se crea un OOV anotado como variación y pasa a la lista final

Solución propuesta (III)

Detección de OOV

- Tokens como entrada
- Aplicación de reglas (URL, Hashtag, etc.)
- Comparación con tres diccionarios: español, inglés y entidades
- Tokens detectados en diccionario español se descartan
- Tokens detectados en diccionario inglés se anotan como NoEs
- Tokens detectados en diccionario de entidades se anotan como Correct
- Los tokens no anotados se convierten en OOV y pasan a fase siguiente.

Solución propuesta (IV)

Generación de candidatos OOV

- Primera fase de la corrección
- Entrada: lista de OOV
- Para cada OOV se genera una lista de candidatos a partir de los métodos:
  - LevenshteinFST
  - Metaphone
  - L_L
  - FastText
  - Accented

Solución propuesta (V)

Ranking de candidatos

- Define la corrección de un candidato o se anota como NoEs si no tiene
- Uso de dos marcadores: N-Gram LM y distancia Damerau-Levenshtein
- Umbral mínimo

Solución propuesta (V)

Ranking de candidatos

- Define la corrección de un candidato o se anota como NoEs si no tiene
- Uso de dos marcadores: N-Gram LM y distancia Damerau-Levenshtein
- Umbral mínimo

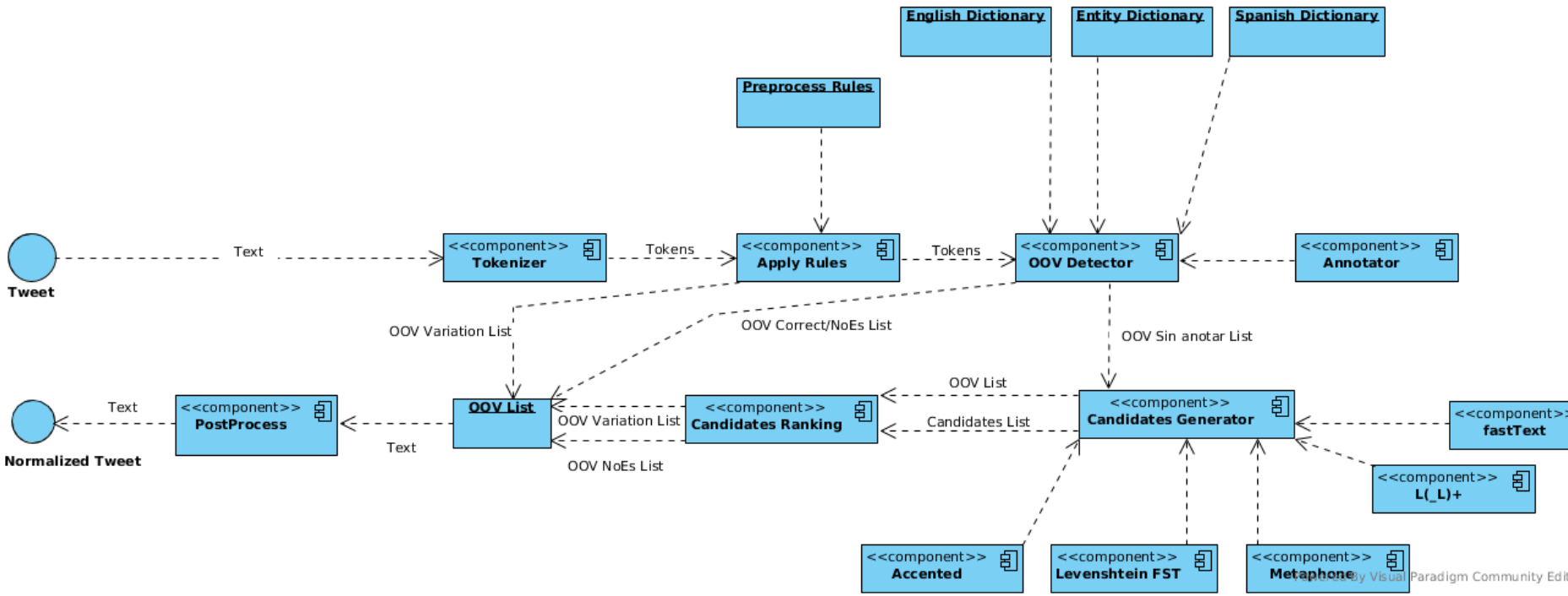
Postproceso

- Añadir mayúsculas, exclamaciones, interrogaciones o signos de puntiación

Implementación

- Tres módulos o componentes:
  - TweetSCCore
  - TweetSCExecutable
  - TweetSCWeb
- Java + Python
- Google Cloud Engine
- Sistema dinámico para que se puedan añadir nuevos métodos y eliminar otros de forma sencilla
- Dos métodos de normalización:
  - DictionaryMethod
  - TweetSCMethod
- Aplicación web: https://jmorenov.github.io/TweetSC/

Arquitectura del sistema



Evaluación

Metodología

- Tarea compartida Tweet-Norm 2013
- Medida de evaluación: corrección de errores

Evaluación

Metodología

- Tarea compartida Tweet-Norm 2013
- Medida de evaluación: corrección de errores

Corpus

- Tweet-Norm 2013
- Dos subconjuntos de tweets
- Desarrollo (500 tweets) y evaluación (600 tweets)

Evaluación

Metodología

- Tarea compartida Tweet-Norm 2013
- Medida de evaluación: corrección de errores

Corpus

- Tweet-Norm 2013
- Dos subconjuntos de tweets
- Desarrollo (500 tweets) y evaluación (600 tweets)

Gold standard

Sistema Rae con un resultado de 0.781

Evaluación (II)

Experimentos

Método	N	Positivos	Negativos	Errores	Accurancy (%)	Tiempo(s)
DictionaryM	10	6	10	4	31.578	1.889
DictionaryM	100	18	82	39	15.652	7.705
DictionaryM	500	131	405	162	21.510	18.926
TweetSCM	10	11	5	4	57.894	4.502
TweetSCM	100	23	77	39	20.0	45.118
TweetSCM	500	122	423	160	20.032	227.438
Sistema RAE	-	-	-	-	78.1	-

Conclusiones

- Diseñado e implementado un corrector de texto en español para Twitter
- Componente de investigación
- Más allá del estado del arte en algunos aspectos: Word2Vec, FST Levenshtein, algoritmo del metáfono y LM N-grama
- Desarrollados tres modulos: TweetSCCore, TweetSCExecutable, TweetSCWeb
- Resultados de los experimentos modestos
- Proyección para mejorar estos resultados

Conclusiones (II)

¿Preguntas?