CS246: Twitter Purification

Wen Shi, Zijun Xue, Jennifer Zhang

June 2014

Abstract

Introduction

Motivation

Overview

Implementation

Dictionary

In a standard spellchecker, a word would be searched against an existing dictionary, and skipped over if it exists and corrected if it doesn't. This is, unfortunately, not an acceptable method of attack for correcting tweets. The major issue with this approach is that it assumes a perfect dictionary; given the amount of slang, proper nouns, etc. in tweets, this is not only exceptionally difficult, but constitutes a moving target as new slang and names are propagated daily.

In the case where a given word already exists in the dictionary, it could easily be a word misuse; aside from common grammatical mistakes such as confounding *your* or *you're*, it is common to intentionally misspell *then* as *den*. Since *den* is a legitimate word in the English language, a normal spellchecker would fail to attempt to correct it.

In the case where a word is encountered that does not exist in a dictionary, it is often more appropriate to not apply a correction. Slang and proper nouns abound in Twitter; while the line between a legitimate word versus a misspelling can be blurry, e.g. *want to* vs *wanna*, but a great deal of slang have no 'correct' equivalent in a dictionary, e.g. Twitter specific words such as *retweet* as well as many more vulgar examples.

Single word correction

Abbreviations: single word

TODO: Jennifer

Squeeze

TODO: Zijun

Edit distance

TODO: Zijun

Phonetic candidates: Soundex

TODO: Shi Wen

Phonetic candidates: Metaphone

TODO: Shi Wen

Letter similarity: Viterbi

Several hundred correction candidates may be found for a word, based on searching within a particular edit distance or the same Soundex/Metaphone class. To trim this list, we first make some intuitive assumptions about the manner in which Twitter users typically misspell their words.

- Abbreviation: Twitter words are usually shorter than their correct counterparts
- Letter similarity: the same important letters are usually present in the Twitter word as the correct counterpart. There are a few phonetic exceptions to this, e.g. substituting *d* for *th* as in *dere* vs *there*.
- Transposition of remaining significant letters is rare, which makes sense if other letters have already been omitted for brevity.

With these observations in mind, we set out to find a reasonable scoring algorithm to measure the similarity between a Twitter word versus its correction candidate.

The Viterbi algorithm is a dynamic programming algorithm typically used in hidden Markov models, such that it finds the optimal path of hidden states given a set of observations.

Word frequency scaling

TODO: Jennifer

Abbreviations: phrase

TODO: Jennifer

Bigram correction

TODO: add some subsections here?

Discussion and Evaluation

Single word results

Bigram results

Future work

TODO: listing any more aspects where it didn't work well

Unconventional word tokenization

Obscure vocabulary or slang

Proper nouns and acronyms

Abbreviations

Dictionary