CS246 Progress Report: Twitter Purifier

Wen Shi, Zijun Xue, Jennifer Zhang

April 2014

Current Progress

We currently have harvested approximately 150,000 tweets via the Twitter sample API.

Building an abbreviation mapping list

Using a combination of UNIX's dictionary, Google Translate's list of most frequently used words[1], and the most common words from TV and movie scripts[2], the most frequent mismatched words from Twitter were found. Among these, around 50 commonly known abbreviations that were also phonetically far from their intended expansion were manually mapped. This process was made more difficult by proper nouns – particularly celebrities, names of companies, and sport teams. It may be desirable to improve this process by also crosschecking against a list of well-known public figures, but since this will not be affect the final product it may not be worthwhile.

Squeezing

Word frequency index

Inverse phonetic index

Changes

Ambiguous or rare abbreviations

In the course of mapping out abbreviations, it is clear that many abbreviations may have different intended expansions under different contexts. E.g. *hw* overwhelmingly represents *homework* as an abbreviation, but is also a typo for *how*. *kd* in an NBA related tweet likely refers to *Kevin Durant*. *kd* in a video game related tweet refers to *kill/death (ratio)*. Furthermore, Kevin Durant is a highly specific abbreviation – making a significance decision about what abbreviations to keep in a manually maintained list is causing some difficulty.

[1] google-10000-english

[2] Wiktionary Frequency Lists

Rather the original simplistic plan of plainly mapping abbreviations to their manually sorted expansions, it may be a better approach to add their expansions as possible candidates before a final decision is made about which one is appropriate, hopefully with some better semantic bigram evidence. This does not fully resolve whether to keep a *kd:Kevin Durant* mapping in the list, though.

Multiple abbreviations with the same meaning

The following abbreviations are all representative of laughter: *lol, lolz, lmao, lmfao, lls, rofl.* From a semantic perspective, it might be better to treat these all as the same token. Furthermore, it might be awkward to spell the full expansion out in the final result, e.g. is it desirable to purify *LOL I'll just ignore this* to *laughing out loud I'll just ignore this*? We are considering mapping all of these to a simple *(laughter)* token.

Semantic bigram frequency table