Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, Noah A. Smith 2011

Aditya Vadrevu

I. Description of the Study

In this paper the main goal for the authors is to tag part of speech for users tweets. The authors collected a total of 2217 tweets but cut it down by 390 to only tag tweets that are in English. The authors first created a POS tag system for the tweets and then the authors manually tagged 1827 tweets. Tagging took about 200 person-hours which was split up by 17 people. This overall project took about two months overall. This problem is more difficult than part of speech tagging a news article or a book because of the various ways a person can convey a message in 140 characters.

II. Methods and Design

This problem that the authors are trying to solve is very difficult not only because of spelling differences but twitter specific grammar and symbols. For example words like LOL or ILY and also shortening words like ima for I'm going to make tagging part of speech more difficult than pure English. The authors also had to take into account the twitter specific characters like the @ symbol and Hashtags(#). First they came up with a tagset which includes tags for punctuation, urls, emojis, as well as normal part of speech tags like nouns, verbs, adjectives, etc. They collected the tweets and ran a WSJ-trained Stanford part of speech tagger to speed up the annotations that they would eventually had to do themselves. Next they pruned the tweets that were not in English and manually tagged the rest of the tweets. After they refined the POS tagger and added hashtags and at-mention functionality. Next They used other features to expand their data. They used TwOrth, Names, TagDict, DistSim, and Metaph.

III. Analysis

The authors randomly split the data into 3 parts, training data of 14,500 tokens, development data with 4,700, and testing data that contains 7,124 tokens. The full feature set achieves an accuracy of 89.37 for the test set. The system predicts verbs, pronouns,prepositions very well and predicts urls and @ with almost 100% accuracy. the system struggles with proper nouns and also the tag G. Even though they were unsuccessful getting the miscellaneous token to get an accuracy above 50% the authors are happy with the overall system.

IV. Results

The basic Stanford training was 85.85% which was better than the basic tagger which they had created without other features. Out of the features Names had the highest test accuracy of 89.39% and the lowest was DistSim and TagDict. The tags that most accurate were the Url, punctuation, coordinating conjunction and pronouns. All of these achieved an accuracy of 97% or higher. The lowest accuracy tag, which was 26%, was the G tag that looks at abbreviations of words.

V. Limitations

The main limitation that they have for this project is that they can not properly identify tags such as ILY and LOL. It is a difficult task but they created a new part of speech tagging system specifically for twitter. The authors could have figured out a better way to split the G tag into smaller more indepth tags. Another hindering limitation is the Metaph system. The Metaph system takes in the phonetic pronunciation of the word and maps it to a simpler word. Words that are almost the same but have different meaning are also roped into one word. For example, War worry and were all sound a bit similar so this system will automatically tag them as the same word.

VI. Conclusion

I think overall this project was a success but not incredibly ground breaking. Even though their work was done well, I believe that a major part of twitter is people who write solely with acronyms and emojis. Not being able to get a decent accuracy for those tags is disappointing but overall getting hastags, URLs, Emojis and @ are a great start and with more learning they can improve upon their original accuracy when predicting a miscellaneous word tag.

References

[1] Gimpel, Kevin, et al. Part of speech tagging for Twitter: annotation, features, experiments. June 2011.