Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments Gimpel 2011

Heman Baral

I. Description of the Study

The purpose of this research is about creating POS tagging by using tweeter data. Creating parts of speech tagging on tweeter data was difficult then Standard English text because of the text on tweeter is informal and 140 character limit(280 character limit now) of each message. By using metaphone algorithm and adding some future on the parts of speech tagger the researcher team develop a tag set that provide 90 percent accuracy.

II. Methods and Design

There were 17 people works together for this project and develop a tree banks by categories twitter using URL and hash tag and then they tokenized random sample of American English tweets data by using twitter tokenizer. In order to speed up the annotation researcher team pre-tagged them by using the WSJ-trained Stanford POS Tagger. A twitter word was categories in tokens by taking precedence over the Stanford tags.They distributed tweets to the annotators and classified non-English word and removed them. The annotation process wasnt cover every situation on the tweet text and for that reason the tagset, annotation guidelines, and tokenization rules were poor or ambiguous. First two annotators examined and fixed all of the English tweets tagged. By reading the annotation instructions the third annotator estimating inter-annotator agreement to create tweets and compare it with tagged tokens. Another annotator correct errors and improve consistency of tagging decisions across the corpus and those data gives the output.

III. Analysis

Twitter hashtags, twitter at-mentions, re-tweet,URLs, emoticons, hashtags and ellipsis dot makes harder to separating words. They used regular expression to separating word and comparing those word with sentence and find the most frequency word and provide the right parts of speech token. For the short from words they combine the preprocess nominal, pronoun, verb and possessive. Partial words, artifacts of to kenization errors, miscellaneous symbols, possessive endings multiword abbreviations and arrows that are not used as discourse markers that do not fit in any of the other they classified in G tagger. Suffix and capitalization patterns unpredictable in the word therefore they added those words in gazetteers tokens. They build a tagger to checks each of the word up to length 3 and separate mostly tags in the words and see what they project. Because of variety word in twitter they use methaphone algorithm which contains of 19 rules to match similar words and names in sentence and provide same key for them.To provide same key they rewrite consonants and delete vowel. They also provide token for most frequent tag for PTB words. Then they evaluated the the parts of speech tagger system that they created by metaphone algorithm and compare with Standford tagger.

IV. Results

They chose randomly twietter data for parts of speech tagging from 1.9 million token from 134,00 unlabled tweet and divided 1,827 annotated tweets into a training set of 1,000 and develop a set of 327 tokens , and a test set of 500 and compare our system against the Stanford tagger and the found out tagging result nearly 90 percent correct.

V. Limitations

The part-of-speech tagging data was limited training and the tagger struggles to identify proper nouns with nonstandard capitalization.The annotation process doesnot covered all the situations for tagset and tokenization rules wasn't sufficent.They tokenize word shape in a 3-word window because of that the 5-word shape word wasnt feet on the tagger set as a result they couldnt compare data with pre trained Stanford tagger to their data.

VI. Significance

The tagger with full feature wasn't able to reduce error 100 percent but was able to reduce 25 percent on standford tagger.The underline tokens were also incorrect in some specific condition and elects, governor and next were providing wrong tagger for those word in twitter.Within was also misclassified with ohh.Additionally,shoutout was appearance one time and identify as verb and rare token that provided in G token also provide errors.

VII. Conclusion

The success of this approach demonstrates that with careful design, supervised machine learning can be applied to rapidly produce effective language technology in new domains. The data and tools have been made available to the research community with the goal of enabling richer text analysis of Twitter and related social media data sets.

References

[1] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010a. From tweets to polls: Linking text sentiment to public opinion time series. In Proc. of ICWSM.

[2] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. An- notating named entities in Twitter data with crowd- sourcing. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.

[3] Grady Ward. 1996. Moby lexicon. http://icon. shef.ac.uk/Moby.