A review on :

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, and Noah A. Smith
School of Computer Science, Carnegie Mellon Univeristy, Pittsburgh, PA 15213, USA

Amin Majdi
Student ID: 01924731

• Problems and contributions: (5%) Summarize the problem of interest and its motivations, and (theoretical, methodological, algorithmic, empirical, etc.,) contributions of the paper.

Usually, Informal writings, people follow some formal conventions. On the other hand, sometimes, the content on social media does not follow these conventions and is very close to speaking. This issue makes POS tagging of content in social media very hard. The motivation of this experiment is to find some features to better do the POS tagging in areas where people do not formally talk. The main contributions of this work are developing a tweeter-specific POS tagger, annotating a data set of tweets and making it publicly available for future works, and analyzing different combinations of features to evaluate the performance of each set of features.

• Method: (10%) Explain the key aspects of the proposed methodology.

Annotation:

First, they did tokenize and pre_tagged using available tokenizers and PBD tagger (Formal tagger) to speed up manual tagging. Then they tried to annotate the tweets manually. In this stage, they found all the tweeter-specific challenges and areas that they can improve in the annotation. Based on these findings, they revised the tokenization and tagging guidelines in the next stage. They did multi annotator analysis, and the inter-annotator agreement rate was 92.2%. Finally, another annotator did the last check to correct some possible mistakes in the process.

Tagset:

They developed a tag-set in addition to some formal tags like a verb, noun, etc., it has some other social media-related tags like hashtags, at-mentions, emotions, etc. The other innovation in this work was the tagging for traditional POS categories. They also introduced four new tags for combined forms: {nominal, proper noun} × {verb, possessive} . The final tag-set contains 25 tags denoted with a single ASCII character.

System:

After defining a tweeter-related tag set, They developed some new features to find and classify the tags in each tweet. The Base features for this experiment consist of a feature for each word type,
a set of features that check whether the word contains digits or hyphens, suffix features up to length three, and features looking at capitalization patterns in the word. Then in each evaluation, they added some other own-developed features consisting of TWORTH, NAMES, TAGDICT, DISTSIM, and METAPH.

Experiments:

In this part, they analyzed the influence of each feature on the accuracy of the model by eliminating that feature from the whole set of features that applied for POS tagging. The final results show that they reached the highest performance by eliminating the NAMES feature from all features.

They also calculated the accuracy rate for all specific tags and showed which false tag was the main cause of the confusion. The other analysis in this paper covers the average position of each tag relative to the middle word in the tweet, which shows the probability of finding the tags in each location. Position analysis can help improve some ambiguous tokens' accuracy by tagging the tokens based on their location in the tweets.

• Strengths: (5%) Describe the strengths of the paper in terms of theoretical or empirical soundness as well as significance and novelty of the contributions.

One of the most significant strengths points of this experiment is the precision in the annotation part. They followed well-developed guidelines to reach a 92.2 % agreement rate.

A significant improvement in accuracy rate (~4%) compared to previous work (Stanford tagger) is another major accomplishment of this work. They did that by developing new features targeting some notations that are only related to the tweeter. They focused more on the features that can find the conversational tags.

• Limitations: (20%) Explain the limitations of the paper; provide specific and detailed points, and avoid vague or subjective complaints.

- The explanation about feature ablation experiments is not convincing. It is not clear why eliminating DISTSIM , TAGDICT, and TWORTH are more costly than eliminating the NAMES.

- Another bold limitation of this experiment is the data-set size. They provided 1,827 annotated tweets for train, development, and test set, which is insufficient for several reasons. Social media like tweeter contains a vast user range that can talk about various topics. Each topic has its orthography, so this work's trained model probably will not work very well on another test set.
- It is unclear how sophisticated the tweets are in terms of grammar and word usage. To have a better data set for training, they could eliminate simple tweets and only train the model with some sophisticated annotated tweets. A good measure of tweet sophistication is the ratio introduced in the Tweet Complexity assignment, Part b.
- The train set limitation can also lead to poor performance in tagging unusual tokens such as split words, hyphenated words, and words with missing or spurious spaces. More training and manual annotation can cover this gap and improve accuracy.