Gimple POS tagging Write Up

Jordan Dube

I. Description of the Study

The goal of this research was to analyze English natural language not just from formal writing but from conversational, informal writing, like that found from Twitter[1]. The authors wanted to explore methods of POS tagging to help future NLP research in the domain of user-produced text on web platforms. Another purpose was to develop a new tagset for tweets[1]. The problem is important because the corpus of social media and web-based language is growing rapidly, so having libraries that aid in the processing of such big data is vital to keeping the rate of processing even close to the rate of production of this language[1]. The authors also state that this study can be viewed as a case study to show the ease of development of a new NLP system for new domains[1].

The research question is something like ”Can we design a POS tagging system based on the domain of Twitter and electronic user-produced text, can we design a POS tagger with sufficient accuracy?” The objectives are described as: develop a tagset for Twitter, manually annotate data, develop features, and report tagging results[1]. The hypothesis is that they are able to design such a system.

II. Methods and Design

The rationale is justified by the work of previous authors in the field. They could logically extend the tree banks of other groups to include tags that fit their domain of Twitter messages[1]. They also go on to explain their choices for designing the tagger and their justification for including various tag sets and features[1].

The size of the sample used in the tests is described as 2217 tweets with 1827 of them being usable English. The key characteristics of these tweets are described by the features used by the tagger which included base features (a feature for each word type, a set of features that check whether the word contains digits or hyphens, suffix features up to length 3, and features looking at capitalization patterns in the word) and additional features that were developed specifically for the Twitter domain (TWORTH, NAMES, TAGDICT, DISTSIM, and METAPH).

The sample does not appear to be representative of the population of English Tweets. While they may have been randomly selected, they were all from 10/27/2010, which is way too specific of a date to be considered representative[1]. A better approach would be to spread collection along multiple dates (weeks or months).

The data was collected by the researchers by filtering tweets to 10/27/2010 on an English-localized server and randomly selecting tweets from the pool. The tweets were automatically tagged by a program and then processed by other researchers who manually corrected the tweets[1]. These annotations were used to create a tagset that was used by the novel tagging program of the researchers[1]. For each of the trials, the output accuracy was recorded. Various tags and their distributions were also recorded and presented to the reader[1].

The data collection is clearly described through a series of steps in the paper.

The authors do discuss the reliability and validity of their methods. They give intermediate metrics to build their case as to the reliability of the annotations such as the cohen κ value of 0.914 and the agreement rate between annotators of 92.2%[1]. The authors also provide the model and materials with the final release of the paper to encourage reproducibility and peer-reviewing.

III. Analysis

The data used to evaluate the method was the tags that the tagger predicted for the input dataset of tweets[1]. Accuracy was calculated based on these predictions and compared to the standard Stanford tagger[1]. If the accuracy was low, then improvements needed to be made to the methods and/or model.

The data was appropriate because it accurately reflected the efficacy of the model.

The metrics used to analyze the results were accuracies of the model runs of both the development and test sets with all the features used and each independent feature set used (ablations), as well as the recall of the more commonly missed predictions[1].

The metrics appear appropriate in that they clearly show the areas in which the model performs well and where it could be better.

IV. Results

The results of the trials are expressed in small tables which are easy to interpret. In addition to these tables, the authors provide explanations as to why some tag sets worked better than others on the input tweet sets[1]. These combine to form an understandable display of the results obtained by the authors.

The interpretations are consistent with the results in that they clearly walk through what the results mean in the context of the goal. The explanations use data shown in the tables to support their claims, such the model's difficulty in identifying proper names being shown in table 3[1].

The conclusions presented in the paper were relevant to the problem. They were trying to make a POS tagger for twitter and their data shows that they were successful in achieving their goal[1].

There were not many recommendations, but they did recommend putting more work into the identification of proper nouns with varying capitalization, which the data showed to be somewhat lacking in their model[1]. I found this to be appropriate because it is an achievable augmentation to their design.

The authors did compare their work to the Stanford pre-trained tagger throughout the paper by using it as a baseline to outperform[1]. They were able to outperform Stanford's model on their target domain of tweets.

V. Limitations

The study had a limitation of time and cost. Paying 17 researchers to manually annotate is expensive but not exclusive to this study. Another limitation is group G in the tagging bank. Just lumping what doesn't fit into a bucket is not robust because new words and phrases/abbreviations are constantly forming. Algorithms work better when the groups are well-defined, but G is a violation of this principle.

VI. Significance

The authors developed a tag set for Twitter, manually tagged 1827 tweets, developed a feature set for Twitter POS tagging, and evaluated features. They also provided the corpus and tagger to the community[1].

This proves to be a milestone in the analysis of modern, heavily abbreviated, and malformed English. Their work helps to guide other people in the community to design models that target specific domains that often operate outside of standard English rules and shows that this can be done in a reasonable amount of time.

This study raises additional questions ”how can proper nouns be better identified?” ”How can the miscellaneous category (G) be further broken down to reduce ambiguity?” ”What rules or what additions can be made to the training stage to reduce the general amount of confusion encountered by the tagger?”

VII. Conclusion

I thought this paper was pretty technical. I appreciate the work the authors did to detail their process, but I found the section detailing the additional features they selected to be lacking. There were a few confusing components about the DISTSIM and METPH features that I wish were further explained in the paper. I found the paper interesting because it uses data that I would find in my own life as opposed to novels or books. I find most social media studies to be pretty intriguing and insightful.

References

[1] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel P Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, 2011.