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

1 Introduction

One of the most fundamental parts of the linguistic pipeline is part-of-speech (POS) tagging, a basic form of syntactic analysis which has countless applications in NLP.

In this paper, we produce an English POS tagger that is designed especially for Twitter data. Our contributions are as follows:

• we developed a POS tagset for Twitter,

• we manually tagged 1,827 tweets,

• we developed features for Twitter POS tagging and conducted experiments to evaluate them, and

• we provide our annotated corpus and trained POS tagger to the research community.

2 Annotation

2.1 Tagset

3 System

Our tagger is a conditional random field (CRF; Lafferty et al., 2001), enabling the incorporation of arbitrary local features in a log-linear model. Our base features include: 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.

TWORTH: Twitter orthography.

NAMES: Frequently-capitalized tokens.

TAGDICT: Traditional tag dictionary.

DISTSIM: Distributional similarity.

METAPH: Phonetic normalization.

4 Experiments

5 Conclusion

We have developed a part-of-speech tagger for Twitter and have made our data and tools available to the research community at <http://www.ark.cs>. cmu.edu/TweetNLP. More generally, we believe that our approach can be applied to address other linguistic analysis needs as they continue to arise in the era of social media and its rapidly changing linguistic conventions. We also believe that the annotated data can be useful for research into domain adaptation and semi-supervised learning.