Automatic Detection of Tags for Political Blogs

Khairun-nisa Hassanali
Human Language Technology Institute
The University of Texas at Dallas
Richardson, TX 75080, USA
nisa@hlt.utdallas.edu

Vasileios Hatzivassiloglou
Human Language Technology Institute
The University of Texas at Dallas
Richardson, TX 75080, USA
vh@hlt.utdallas.edu

Abstract

This paper describes a technique for automatically tagging political blog posts using SVM's and named entity recognition. We compare the quality of the tags detected by this approach to earlier approaches in other domains, observing effects from the political domain and benefits from NLP techniques complementary to the core SVM method.

1 Introduction

Political blogs are a particular type of communication platform that combines analyses provided by the blog owner or a team of regular contributors with shorter, but far more numerous, entries by visitors. Given the enthusiasm that activities for or against a particular politician or party can generate, political blogs are a vibrant part of the blogosphere: more than 38,500 blogs specifically dedicated to politics exist in the US alone according to Technorati, and some of the more active ones attract more than 30 million unique visitors each month (double that number just before major elections).

Political blogs provide a wealth of factual information about political events and activities, but also by their nature are colored by strong opinions. They are therefore a particularly attractive target for semantic analysis methods using natural language processing technology. In fact, the past two years have brought an increased number of collaborations between NLP researchers and political scientists using data from political sources, including two special issues of leading political science journals on such topics (see (Cardie and Wilkerson, 2008) for an overview). Our motivation for working with this kind of data is the construction of a system that collates information across blog posts, combines evidence to numerically rate attitudes of blogs on different topics, and traces the evolution of these attitudes across time and in response to events. To enable these tasks, we first identify the major topics that each blog post covers. In the present paper, we describe our recognizer of blog post topics. We show that, perhaps because of the richness of political blogs in named entities, an SVM-based keyword learning approach can be complemented with named entity recognition and co-reference detection to achieve precision and recall scores higher than those reported by earlier topic recognition systems in other domains.

2 Related Work

In our approach, as in earlier published work, we take *tags* assigned by many blogs to individual blog posts as a reference list of the topics covered by that post. Tags are single words or short phrases, most often noun phrases, and are usually chosen by each post's authors without a controlled vocabulary; examples include "Michigan", "George Bush", "democracy", and "health care". Earlier work in predicting tags includes (Mishne, 2006), who adopts a collaborative filtering approach; in contrast, we rely on training classifiers from earlier posts in each blog. Our approach is more similar to (Sood et al., 2007) and (Wang and Davison, 2008) who use different machine learning techniques applied to a training set. We differ from the last two approaches in our addition of proper noun and named entity recognition methods to our core SVM classifiers, in our exploration of specifically political data, and in our subsequent use of

Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, pages 21–22,
Los Angeles, California, June 2010.

the predicted tags (for semantic analysis rather than tag set compression or query expansion).

3 Data

We collected data from two major political blogs, Daily Kos (www.dailykos.com) and Red State (www.redstate.com). Red State is a conservative political blog whereas Daily Kos is a liberal political blog. Both these blogs are widely read and tag each of their blog entries. We collected data from both these blogs over a period of two years (January 2008 – February 2010). We collected a total of 100,000 blog posts from Daily Kos and 70,000 blog posts from Red State and a total of 787,780 tags across both blogs (an average of 4.63 tags per post).

4 Methods

We used SVM Light (Joachims, 2002) to predict the tags for a given blog post. We constructed one classifier for each of the tags present in the training set. The features used were counts of each word encountered in the title or the body of a post (two counts per word), further subdivided by whether the word appears in any tags in the training data or not, and whether it is a synonym of known tag words. We extract the top five proposed tags for each post, corresponding to the five highest scoring SVM classifiers.

We also attempt to detect the main entities being talked about. We perform shallow parsing and extract noun phrases and then proper nouns. The most frequent proper NPs are probable tags. We also added named entity recognition and co-reference resolution using the OpenNLP toolkit (maxent.sourceforge.net). We found that named entity recognition proposes additional useful tags while the effect of co-reference resolution is marginal, mostly because of limited success in actually matching co-referent entities.

5 Results and Evaluation

For evaluating our methods, we used 2,681 posts from Daily Kos and 571 posts from Red State. We compared the tags assigned by our tagger to the original tags of the blog post, using an automated method (Figure 1). A tag was considered a match if it exactly matched the original tag or was a word super set – for example "health care system" is considered a match to "health care". We also manually evaluated the relevance of the proposed tags on a small portion of our test set (100 posts).

Method	Precision	Recall	F-Score
Single word SVM	27.3%	60.3%	37.6%
+ Stemming	26.1%	59.5%	36.3%
+ Proper Nouns	36.5%	56.8%	44.4%
Named Entities	48.4%	49.1%	48.7%
All Combined	21.1%	65.0%	31.9%
Manual Scoring	**67.0%**	**75.0%**	**70.8%**
Single word SVM	19.0%	30.0%	23.3%
+ Stemming	22.0%	30.2%	25.5%
+ Proper Nouns	46.3%	54.0%	49.9%
Named Entities	**60.1%**	41.5%	49.1%
All Combined	20.3%	**65.7%**	31.0%
Manual Scoring	47.0%	62.0%	**53.5%**

Figure 1: Results on Daily Kos (top) and Red State (bottom) data. Best scores in bold.

6 Conclusion

We described and evaluated a tool for automatically tagging political blog posts. Political blogs differ from other blogs as they often involve named entities (politicians, organizations, and places). Therefore, tagging of political blog posts benefits from using basic name entity recognition to improve the tagging. The recall in particular exceeds the score obtained by earlier techniques applied to other domains (Sood et al. (2007) report precision of 13% and recall of 23%; Wang and Davison (2008) report precision of 45% and recall of 23%).

References

Claire Cardie and John Wilkerson (editors). "Special Volume: Text Annotation for Political Science Research". *Journal of Information Technology and Politics*, 5(1):1-6, 2008.

Thorsten Joachims. SVM-Light. 2002. http://www.svmlight.joachims.org.

Gilad Mishne. "AutoTag: A Collaborative Approach to Automated Tag Assignment for Weblog Posts". In *Proceedings of WWW*, 2006.

Sanjay C. Sood, Sara H. Owsley, Kristian J. Hammond, and Larry Birnbaum. "TagAssist: Automatic Tag Suggestion for Blog Posts". In *Proceedings of ICWSM*, 2007.

Jian Wang and Brian D. Davison. "Explorations in Tag Suggestion and Query Expansion". In *Proceedings of SSM '08*, 2008.