Making #Sense of #Unstructured Text Data

L. Li, W. M. Campbell, C. Dagli, J. P. Campbell
MIT Lincoln Laboratory
Lexington, MA 01740
lin.li@ll.mit.edu, wcampbell@ll.mit.edu

1 Introduction

Automatic extraction of intelligent and useful information from data is one of the main goals in data science. Traditional approaches have focused on learning from structured features, i.e., information in a relational database. However, most of the data encountered in practice are unstructured (i.e., social media posts, forums, emails and web logs); they do not have a predefined schema or format. In this work, we examine unsupervised methods for processing unstructured text data, extracting relevant information, and transforming it into structured information that can then be leveraged in various applications such as graph analysis and matching entities across different platforms.

Various efforts have been proposed to develop algorithms for processing unstructured text data. At a top level, text can be either summarized by document level features (i.e., language, topic, genre, etc.) or analyzed at a word or sub-word level. Text analytics can be unsupervised, semi-supervised, or supervised. In this work, we focus on word analysis and unsupervised methods. Unsupervised (or semi-supervised) methods require less human annotation and can easily fulfill the role of automatic analysis. For text analysis, we focus on methods for finding relevant words in the text. Specifically, we look at social media data and attempt to predict hashtags for users' posts. The resulting hashtags can be used for downstream processing such as graph analysis. Automatic hashtag annotation is closely related to automatic tag extraction and keyword extraction. Techniques for hashtags extraction include topic analysis, supervised classifiers, machine translation methods, and collaborative filtering [1, 2]. Methods for keyword extraction include graph-based and topical analysis of text [3, 4].

2 Structuring Unstructured Text Data

Our approach differs from the previous methods by offering an end-to-end process of analyzing unstructured text data via automatic hashtag annotation, building graphs of entities and hashtags, and performing graph analysis that are directly applicable to downstream applications.

For hashtag annotations, we explore two different but related strategies: topic-based method and community-based graph method. Our pipeline for the topic-based method is as follows. First, we extract word counts from users' posts and perform PLSA topic analysis [5] at the topic level. Second, we extract the top-M most relevant words from each topic. Each of these words is then annotated as a hashtag in the posts. For the community-based graph method, we first construct a word co-occurrence graph, where vertices in the graph represent words and edges represent co-occurrence of words in users' posts. We then perform community detection on the co-occurrence graph. From each community, we extract words with the highest PageRank values and then annotate them as hashtags.

Graph construction is performed by designating both users and the annotated hashtags as vertices and representing edges as four types of interactions: user-to-user posts, user mentions of hashtags, reposts and co-occurrence of users or hashtags. Edge weights are given by the counts across different edge types. The resulting content + context graph has been shown to be useful in a number of applications. In the experiment, we showcase its usefulness through an important application, entity resolution [6].

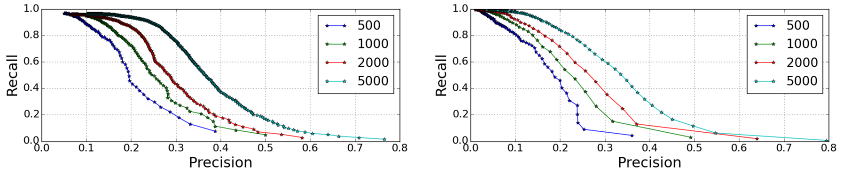
[*] This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

3 Experiments

We perform experiments on Twitter and Instagram corpora; see [7, 6] for details about the data. First, we remove user-annotated hashtags and then use topic- and community-based methods for automatic hashtags annotation. Fig. 1 shows precision and recall of recovering the hashtag vocabulary.

Second, we examine the application of cross-domain entity resolution. We use the automatically-extracted hashtags to construct Twitter & Instagram graphs, and then apply cross-domain community detection [7] and entity-resolution analysis [6] to extract graph features for matching accounts across Twitter and Instagram. The graph features are then combined with users' profile features (i.e., username and full name match) to obtain a fused entity resolution system. The performance of cross-domain entity resolution system is measured by standard miss and false alarm rates. The resulting equal error rates (EER) for all trials and non-trivial trials (NT) are shown in Table 1. Non-trivial trials are trials where the usernames are not an exact string match. Observe that systems using automatic hashtag analysis perform significantly better than the profile-only system. They also show comparable performance (if not better) to the system using the original user-generated hashtags.



(a) Topic-based method

(b) Community-based method

Figure 1: Precision and recall curves of the automatically-annotated hashtags against top 500, 1000, 2000 and 5000 user-generated hashtags.

Table 1: Summary of entity resolution results. 'P' denotes the profile feature, 'N' denotes the content + context graph feature, and 'P+N' denotes the fusion of the two. The interpolated value is given by *.

Fusion	**Hashtags**	**Method**	**EER ALL (%)**	**EER NT (%)**
P			1.54	5.74*
P+N	Original Hashtags		1.16	3.79
P+N	4K Automatic Hashtags	Topic	**1.17**	3.48
P+N	4K Automatic Hashtags	Community	1.19	**3.39**
P+N	10K Automatic Hashtags	Topic	1.17	3.51
P+N	10K Automatic Hashtags	Community	**1.1**	**3.24**

4 Conclusion

We present an end-to-end system for analyzing unstructured text data and transforming the data into structured graphs that can be directly applied to various downstream applications. For text analysis, we have presented two unsupervised methods for automatic hashtag annotation: topic-based method and community-based method. Both methods show promising results in predicting relevant words from the text. We also build content + context graphs using the automatically-annotated hashtags and apply them to the cross-domain Twitter/Instagram entity resolution problem. The performance of the resulting system is comparable to the system using the original user-generated hashtags.

References

[1] R. Krestel, P. Fankhauser, and W. Nejdl. Latent dirichlet allocation for tag recommendation. In *Proc. of the 3rd ACM conf. on Recommender sys.*, pages 61–68. ACM, 2009.

[2] Z. Liu, X. Chen, and M. Sun. A simple word trigger method for social tag suggestion. In *Proc. of the Conf. on Empirical Met. in Nat. Lang. Proc.*, pages 1577–1588. ACL, 2011.

[3] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. ACL, 2004.

[4] W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim, and X. Li. Topical keyphrase extraction from twitter. In *Proc. of the 49th An. Meet. of ACL: Human Lang. Tech.-Vol. 1*, pages 379–388. ACL, 2011.

[5] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of the 22nd an. int. ACM SIGIR conf. on res. and devel. in inf. retrieval*, pages 50–57. ACM, 1999.

[6] W. M. Campbell, L. Li, C. Dagli, J. Acevedo-Aviles, K. Geyer, J. P. Campbell, and C. Priebe. Cross-domain entity resolution in social media. In *The 4th Int. Workshop on Nat. Lang. Proc. for Social Media*, 2016.

[7] L. Li and W. M. Campbell. Matching community structure across online social networks. In *Proc. of the NIPS Workshop: Networks in the Social and Information Sciences*, 2015.