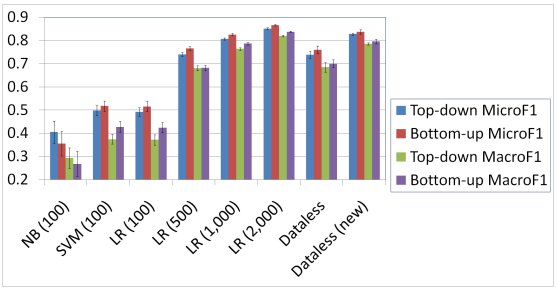

Figure 2: 20newsgroups: comparison of dataless hierarchical classification with supervised baselines. All methods are evaluated based on average of ten randomly sampled trials. "SVM (100)" represents SVM with 100 labeled data. "Dataless" means ESA (500) + bootstrapping. "Dataless (new)" means ESA (500) with new descriptions (in Table 2) + bootstrapping.

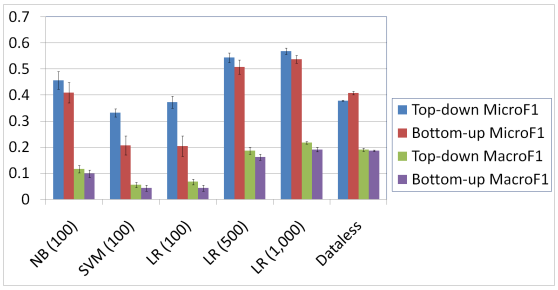

Figure 3: RCV: comparison of dataless hierarchical classification with supervised baselines. All methods are evaluated based on average of ten randomly sampled trials. "SVM (100)" represents SVM with 100 labeled data. "Dataless" means ESA (500) + bootstrapping.

this case, one-vs-rest approaches, i.e., NB and LR, perform better than SVM's pairwise approach. Moreover, NB, with binary features, seems to be more stable in this case of large number of classed with few examples each. It is also interesting to see that, for 20NG, bottom-up is better than top-down mechanism, but when the number of labels increases, as in RCV1, the top-down mechanism seems to be more stable and shows higher F_1 scores. The reason is that in higher levels of the hierarchy top-down has less class labels to work with, and thus has more examples for each class.

Dataless Classification Our key experiments make use of the two step datalass classification process: we choose ESA with 500 concepts as the initialization approach, and use LR as the base classifier for each node, since we find it to be more stable across datasets. In the bootstrapping process, when we find a labeled documents (via previous steps), we also add the label to all the ancestors' labeled sets. Therefore, for each iteration, there will be more than $N * |\mathcal{T}|$ documents added in the tree of classifiers. For 20NG data, we randomly sample 50% of the document set and allow the bootstrapping process to access it, and we use the rest as test data. For RCV1, bootstrapping can access 80% of the documents for training, for compatibility with the supervised methods. We empirically set $N = 20$ for both datasets. All the results are average of ten trials.

The bottom line is that the results of the overall dataless process (Figs. 2 and 3) is shown to be competitive with supervised training. Specifically, for 20NG, bootstrapping is competitive with supervised methods with 500 labeled documents for old description, and with 1,000 labeled documents for new label descriptions. For RCV1 data, although the performance of dataless classification is worse than supervised algorithms on $MicroF_1$ scores with top-down approaches, it is shown that dataless approaches are significantly better on $MacroF_1$. The reason is that for RCV1, there are several classes with no training data. While this does not impact the dataless approach, it does affect the average result over all labels of the supervised approaches. The results of dataless on $MacroF_1$ are also competitive with supervised methods with 500 labeled documents.

Discussion and Conclusion

In this section we discuss some further analysis we conducted to better understand our results and assess the practical use of dataless classification. The scenario we envision for dataless classification includes a collection of documents along with an ontology of possible categories that we want to assign to each document. While for evaluation purpose, we had to work with a small and closed set of category labels, we believe that this type of evaluation does not reflect the true ability of dataless classification. To validate this intuition, we performed two additional experiments.

First, we renamed the categories in the 20NG dataset to better reflect the content of the collection (shown in Table 2). Given the new descriptions, we tested some of the semantic representations and compared them with the previous performance. The key observation is that the dataless classification given by both ESA and $\mathrm{WE}_{\mathrm{Mikolov}}$ are significantly improved. As a consequence, the bootstrapping results are also improved (results of ESA are shown in Table 3 and Fig. 2).

In our second experiment we used the Yahoo! Directory's categories as the dataless labels. We used 661 unique categories which are the leaves in our hierarchy, taken from the first, second and (some of the) third level of the hierarchy. Once we classified the 20NG documents into this large hierarchy, we analyzed the result by comparing the labels given by the dataless algorithm to the gold labels. The results are very satisfying. For example, the documents in the "rec.autos" newsgroup are mostly classified to Yahoo! categories "news and media: traffic and road conditions" and "sports: wheelchair racing." Moreover, documents in newsgroup "talk. politics.misc" that are known to contains document on social issues are classified mostly into Yahoo! categories "news and media: cultures and groups," "social science: lesbian gay bisexual and transgendered studies," "health: long term care," etc. Finally, we observe that coherent groups are classified as such – most of documents classified in "science: aeronautics and aerospace" are from "sci.space" newsgroup. Our conclusion is that given a large label hierarchy such as the Yahoo! Directory, our dataless method allows for robust organization of the documents by their content.

Overall, we proposed a dataless hierarchical classification approach for text categorization. Hierarchical classification is a more general and realistic protocol for text classification. We studied both top-down and bottom-up mechanisms and showed that "bottom-up" approache is more useful in the dataless setting. We systematically compared the ESA approach to other "modern" representations, i.e., Brown clusters, word embedding, and OHLDA topics, thus demonstrating the importance of representation for dataless classification. Not surprisingly, ESA is found to be better suited for this task. Finally, our experiments indicate that dataless hierarchical classification is a promising and practical direction.

Acknowledgements

This work is supported by the Army Research Laboratory (ARL) under agreement W911NF-09-2-0053, by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20155, and by DARPA under agreement number FA8750-13-2-0008. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied by these agencies or the U.S. Government.

References

Agrawal, R.; Gupta, A.; Prabhu, Y.; and Varma, M. 2013. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *WWW*, 13–24.

Brown, P. F.; Pietra, V. J. D.; DeSouza, P. V.; Lai, J. C.; and Mercer, R. L. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18(4):467–479.

Cai, L., and Hofmann, T. 2004. Hierarchical document categorization with support vector machines. In *CIKM*, 78–87.

Chang, M.; Ratinov, L.; Roth, D.; and Srikumar, V. 2008. Importance of semantic representation: Dataless classification. In *AAAI*, 830–835.

Chen, H., and Dumais, S. 2000. Bringing order to the web: Automatically categorizing search results. In *CHI*, 145–152.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. P. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* 12:2493–2537.

Dagan, I.; Karov, Y.; and Roth, D. 1997. Mistake-driven learning in text categorization. In *EMNLP*, 55–63.

Dumais, S., and Chen, H. 2000. Hierarchical classification of web content. In *SIGIR*, 256–263.

Elhoseiny, M.; Saleh, B.; ; and A.Elgammal. 2013. Write a classifier: Zero shot learning using purely textual descriptions. In *ICCV*, 1433–1441.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.* 9:1871–1874.

Gabrilovich, E., and Markovitch, S. 2006. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *AAAI*, 1301–1306.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *IJCAI*, 1606–1611.

Gopal, S., and Yang, Y. 2013. Recursive regularization for large-scale classification with hierarchical and graphical dependencies. In *KDD*, 257–265.

Ha-Thuc, V., and Renders, J.-M. 2011. Large-scale hierarchical text classification without labelled data. In *WSDM*, 685–694.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *ECML*, 137–142.

Lang, K. 1995. Newsweeder: Learning to filter netnews. In *ICML*, 331–339.

Lewis, D. D.; Yang, Y.; Rose, T. G.; and Li, F. 2004. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* 5:361–397.

Liang, P. 2005. Semi-supervised learning for natural language. Master's thesis, Massachusetts Institute of Technology.

Liu, T.-Y.; Yang, Y.; Wan, H.; Zeng, H.-J.; Chen, Z.; and Ma, W.-Y. 2005. Support vector machines classification with a very large-scale taxonomy. *SIGKDD Explor. Newsl.* 7(1):36–43.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*. 3111–3119.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, 746–751.

Murphy, G. L. 2002. *The Big Book of Concepts*. MIT Press.

Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In *NIPS*, 1410–1418.

Ratinov, L., and Roth, D. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*, 147–155.

Rizzolo, N., and Roth, D. 2010. Learning based java for rapid development of NLP systems. In *LREC*.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. Y. 2013. Zero-shot learning through cross-modal transfer. In *NIPS*, 935–943.

Sun, A., and Lim, E.-P. 2001. Hierarchical text classification and evaluation. In *ICDM*, 521–528.

Turian, J.; Ratinov, L.; and Bengio, Y. 2010. Word representations: A simple and general method for semi-supervised learning. In *ACL*, 384–394.

Xiao, L.; Zhou, D.; and Wu, M. 2011. Hierarchical classification via orthogonal transfer. In *ICML*, 801–808.

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. *Inf. Retr.* 1(1-2):69–90.