Specificity Helps Text Classification

Lucas Bouma Maarten de Rijke

ISLA, University of Amsterdam,
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
lbouma,mdr@science.uva.nl

Abstract. We examine the impact on classification effectiveness of semantic differences in categories. Specifically, we measure broadness and narrowness of categories in terms of their distance to the root of a hierarchically organized thesaurus. Using categories of four different levels degrees of broadness, we show that classifying documents into narrow categories gives better scores than classifying them into broad terms, which we attribute to the fact that more specific categories are associated with terms with a higher discriminatory power.

1 Introduction

While text categorization has a long history [7], the increased availability of large scale semantically rich thesauri and ontologies, raises a number of challenging scientific questions. If we classify text documents into categories that are organized in such a semantic structure, how can we exploit the structure? How does the position of a category in such a hierarchy impact a classifier's performance?

Specifically, in this paper we aim to find out whether classification accuracy is influenced by the level of "broadness" (or "narrowness") of a category. A priori, one may entertain one of two clear intuitions here. One is that classification into broader classes is more effective than into narrow categories due to more training examples [9]. The competing intuition is that classification into more narrow categories is more effective because the terms associated with such categories tend to be more discriminating. Our experiments show that the latter is the case.

The rest of the paper is organized as follows. In Section 2 we describe our experimental set-up. We follow with our results and a discussion in Section 3, and conclude in Section 4.

2 Experimental Set-up

We addressed our research question by working with data provided by TREC as part of the classification task for the 2004 edition of the Genomics track [8]. Here, Medline documents need to be classified in categories that correspond to term descriptions in the MeSH thesaurus [6]. Categories are organized in levels, from broad to narrow, depending on the length of the shortest path to the root of the thesaurus. A total of eleven levels are found in MeSH.

Level 1		Level 3	
6847	Pharmaceutical_Preparations	8383	Bladder
3937	Eye_Diseases	8186	Education,_Medical
2472	Parasitic_Diseases	4203	Malondialdehyde
1421	Archaea	1990	Philosophy,_Medical
1110	Organic_Chemicals	1365	Product_Surveillance,_Postmarketing
947	Animal_Diseases	1118	Work_Schedule_Tolerance
910	Endocrine_System	844	Disasters
Level 8		**Level 10**	
8409	Xenopus_laevis	8360	Macaca_mulatta
7226	Mice,_Mutant_Strains	7943	Cercopithecus_aethiops
4216	Motor_Cortex	2396	Trypanosoma_cruzi
2376	Receptors,_Antigen,_T-Cell,_gamma-delta	1530	Trypanosoma_brucei_brucei
1421	Medroxyprogesterone_17-Acetate	1183	Entamoeba_histolytica
1162	Goldfish	4190	Macaca_fascicularis
1024	Receptors,_Kainic_Acid	981	Leishmania_donovani

Table 1: Categories chosen for our experiments, grouped by level, together with the number of examples per selected category.

From the eleven levels found in MeSH, we selected four for our experiments—1, 3, 8 and 10—, and from each we selected seven categories, which we hoped would allows us to demonstrate differences in classification effectiveness across levels. Level 10 had the smallest number of categories (32); we selected the seven categories with the most examples. Level 3 had the most categories (2525). For levels 1, 3, and 8 we selected seven categories with roughly the same number of examples as the selected categories at level 10. Table 1 shows the chosen categories and the number of positive examples used in the experiments. To rule out other possible semantic influences we made sure that the selected categories are all unambiguous (that is, they have one, and only one, level in the MeSH thesaurus).

To build the training material for our experiments, we took a sample of documents from the Medline corpus used at TREC. One hundred categories were randomly selected from MeSH. We used around 40 thousands documents that are classified with these categories, these were used as negative instances. We made sure that the term distributions in the different MeSH levels in the sample were statistically the same as in the entire corpus. In the experiments the positive instances of the chosen category were merged with this sample.

For text representation, we employed Weka [10]. Following standard practice, documents were turned into word vectors, each consisting of one thousand most significant words after eliminating stopwords; here, significance was measured by using TF.IDF. Stemming was not used.

Finally, we carried out single-label classification experiments using the SVM-Light [5] and BBR [2] classifiers for each of the 28 categories chosen. Both classifiers have been shown to perform well on the classification task at the TREC Genomics track [4]. The classification effectiveness is measured in Precision, Recall, and F-score, all averaged over all categories per level.

	SVM			BBR		
Level	Precision	Recall	F-score	Precision	Recall	F-score
1	90.33	61.98	72.89	67.50	65.21	66.22
3	90.21	73.93	80.76	75.66	74.64	75.07
8	94.80	85.41	89.78	86.84	85.19	85.94
10	96.80	87.48	91.80	92.71	88.51	90.52

Table 2: Average scores per level (SVM and BBR)

3 Results

Classification into narrow categories was found to be significantly more effective than into broad categories: for each level considered, the F-scores for that level were higher (in many cases significantly so) than the F-scores for all broader levels.

Specifically, Table 2 shows the averaged Precision, Recall, and F-scores for each of the levels. Observe that the F-scores increase, for both classifiers, as the category level increases. The two classifiers behave quite differently, however. For SVM the precision is high for all levels, even for the broadest categories (level 1); for BBR precision and recall increase almost in sync.

A significant ($\alpha = 0.1$) difference of 10 points in F-score was found between level one and level ten. Level one compared with level three and level three compared with level eight both gave a significant difference of 5 points in F-score, but with weaker evidence ($\alpha = 0.25$). No significant difference was found between levels eight and ten.

For finding a possible explanation for the observed differences in classification effecitiveness, we carried out an analysis of the TF.IDF scores in the word vectors used to represent documents. For every category, we ranked the features according to their TF.IDF score, and found no differences between the TF.IDF scores of the most discriminating terms for levels 1 and 3, while the scores for the most discriminating terms at levels 8 and 10 as much as 50% higher—supporting the intuition that more specific categories are associated with terms with a higher discriminatory power.

4 Conclusion

Our findings refute claims by Wibowo and Williams [9] that classification into broader categories is more accurate than into narrow categories. We explain the different findings in terms of the fact that 80 of narrow categories used by Wibowo and Williams [9] had only one training example. In our study the number of positive examples for the narrow categories ranged from 981 to 8360. The larger amount of narrow category examples can be seen as a positive influence on the discriminatory power of the features. Also the specific domain of the MeSH thesaurus should help in that matter.

As to future work, in our research so far we ignored the fact that many category labels are ambiguous, in the sense that they may occur at different levels

in the thesaurus: we did not investigate whether the ambiguity of a category label impacts categorization accuracy. Additionally, for a broad category like *Animal diseases* the singular and plural form of the words *animal* and *disease* are both scored separately. Scoring according to the same morphological root could increase their influence, and we conjecture that multiple word representations [1] will probably have a positive effect on classification effectiveness here. Finally, Granitzer [3] uses the hierarchy as a path for classification. More attention could be advised for top level decisions, also since they are propagated downwards.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 220-80-001, 264-70-050, 365-20-005, 612.000.106, 612.000.207, 612.013.001, 612.066.302, 612.069.006, 640.001.501, and 640.002.501.

References

[1] S. Bloehdorn and A. Hotho. Boosting for text classification with semantic features. In *Proceedings of the Workshop on Mining for and from the Semantic Web at the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 70–87, 2004. http://www.aifb.uni-karlsruhe.de/WBS/sbl/publications/2004-08-ws-msw-bloehdorn-hotho_boosting-semantic-features.pdf.

[2] A. Dayanik, D. Fradkin, A. Genkin, P. Kantor, D. Madigan, D. Lewis, and V. Menkov. Dimacs at the TREC 2004 genomics track. In *The Thirteenth Text Retrieval, Conference (TREC 2004)*, 2005.

[3] M. Granitzer. Hierarchical Text Classification using Methods from Machine Learning. Master's thesis, Graz University of Technology, 2003.

[4] W. Hersh, R. Bhuptiraju, L. Ross, P. Johnson, A. Cohen, and D. Kraemer. TREC 2004 genomics track overview. In *The Thirteenth Text Retrieval, Conference (TREC 2004)*, 2005.

[5] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods*. 1999.

[6] MeSH. National library of medicine, medical subject headings (MeSH), 2005. URL: http://www.nlm.nih.gov/mesh/MBrowser.html.

[7] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

[8] TREC Genomics Track. Trec genomics 2004 ad hoc task documents, 2005. URL: http://ir.ohsu.edu/genomics/.

[9] W. Wibowo and H. Williams. On using hierarchies for document classification. In *Proceedings of the Fourth Australasian Document Computing Symposium*, Coffs Harbour, Australia, 1999.

[10] I. H. Witten and E. Frank. *Data mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000. ISBN 1-55860-552-5.