Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2006-054
CBCL-263

August 9, 2006

Human Document Classification Using Bags of Words

Florian Wolf, Tomaso Poggio, and Pawan Sinha

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Human Document Classification Using Bags of Words

Florian Wolf, Tomaso Poggio and Pawan Sinha

Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139

Humans are remarkably adept at classifying text documents into categories. For instance, while reading a news story, we are rapidly able to assess whether it belongs to the domain of finance, politics or sports. Automating this task would have applications for content-based search or filtering of digital documents. To this end, it is interesting to investigate the nature of information humans use to classify documents. Here we report experimental results suggesting that this information might, in fact, be quite simple. Using a paradigm of progressive revealing, we determined classification performance as a function of number of words. We found that subjects are able to achieve similar classification accuracy with or without syntactic information across a range of passage sizes. These results have implications for models of human text-understanding and also allow us to estimate what level of performance we can expect, in principle, from a system without requiring a prior step of complex natural language processing.

Introspection suggests that for text understanding, humans use a form of representation that takes into account structural and layout information in addition to word-level information. This is particularly likely in a typical document classification task scenario where only limited time is available to choose, from a set of documents, those that are relevant to a certain query or interest. In such a scenario, it seems plausible that syntax and layout information, such as headlines or paragraph boundaries, may be important for performing the classification task. However, there is little systematic experimental work that directly tests this expectation; how would classification performance suffer if syntax and layout information were removed from text passages?

Following standard practice in the field (Mitchell 1997), we refer to a syntax and layout free representation as a 'bag of words' (BOW). A BOW is a feature vector where each element in the vector indicates the presence (or absence) of a word. Certain words are excluded from the BOW, such as function words and words whose meaning is context-dependent (*and, but, because, additionally, the, a(n), they, while, where, who, of, in, on,* etc). Since a BOW is an unordered vector, it lacks structural and layout information (such as sentence structures, paragraph outlines, text formatting, etc).

Our goal was to test whether human classification performance is compromised with a BOW based representation relative to normally structured text and whether increased time pressure on the task increases the need for a structured representation. Furthermore, we wanted to assess how classification performance with the two representations changes as a function of the number of words included in the passages. In order to probe these questions, we experimentally compared human document classification performance on fully structured documents with performance on a BOW representation, with and without time constraints on the task and with and without progressive revealing of words in a passage.

The first set of investigations comprised three similar experiments, which differed only in the source of their stimuli. The sources were The New York Times, e-mails from newsgroups on the internet, and CNN news articles. Participants were drawn from the M.I.T. community. In each experiment, the passages were grouped into five categories (For NYT passages: business, politics, sports, science, technology; For newsgroups: computer, miscellaneous sales, sports, science/technology, politics; For CNN passages: science, business, sports, politics, entertainment). For NYT and CNN, we conflated international and national politics into one category.

In order to create the BOW representations, we used a Perl script (Wall *et al*., 2000) to remove function words and words whose meaning is context-dependent. The remaining words were placed in an array that was subjected to Fisher-Yates shuffling (Fisher and Yates, 1938; Durstenfeld, 1964) in order to create an unordered array. The unordered array was then printed out in three columns of words (The Perl script we used is available as Supplementary Material). In order not to make the task trivially easy, we excluded the title/subject-line of each passage. The BOW and formatted texts were presented with five levels of presentation time per document: one, two, three, and four seconds, and unlimited. The conditions with limited presentation time did not allow complete reading of the full documents, but forced participants to skim over the texts. Additional experimental details are provided in the Methods section under the heading ‘Experiment 1’.

If layout information plays a role in human document classification, performance on the BOW texts should be worse than performance on the formatted texts. In addition, if participants rely on layout information more under time pressure than if they have unlimited time to perform the categorization task, performance on the BOW texts should decrease with decreasing presentation time more than for the formatted texts, showing an interaction of presentation format and presentation time.

For both the NYT and the CNN stimuli, we found no significant differences across the randomized BOW and the intact text and no interactions between presentation format and presentation time. Absolute performance was high (>90%) for both of these experiments in all conditions. For the newsgroups, we found a main effect ($p < 0.003$) of presentation format (BOW or normal) but the difference in performance was small, averaging 5%. The newsgroup data are shown in figure 1.

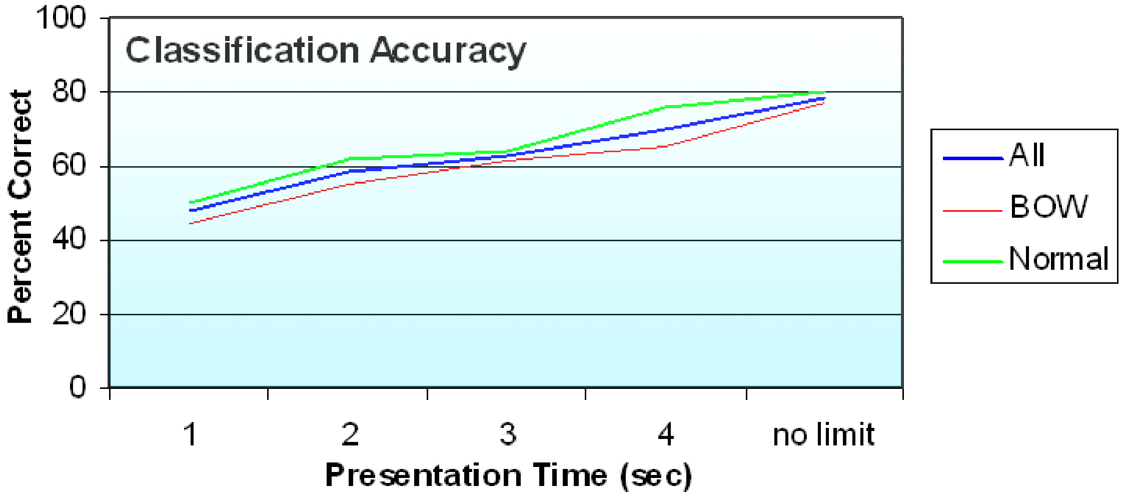

Figure 1. Experiment 1 results: Newsgroup text classification accuracy of human readers across different presentation times.

Although these data provide preliminary evidence that word order may not be crucial for basic text classification, they suffer from an important limitation. The reason that we are unable to see large differences in accuracy across the conditions could simply be that ceiling effects mask out any distinctions. Although the newsgroup data argue against this account (since performance is well below ceiling), this is a valid concern for results with the NYT and CNN datasets.

In order to gain greater sensitivity in our comparison of intact and BOW text representations, we conducted a second set of experiments involving progressive revealing of words. These were inspired in part by sequential sampling process models of decision making (Laming, 1968; Link and Heath, 1975; Ratcliff, 1978, Vickers, 1979); their goal was to determine how the probability of correct classification changed as a function of the number of words shown in the two representations. This set of experiments was performed with a subject population distinct from the population for experiment set 1. Subjects were instructed to provide not only a class label for the text, but also a numerical rating of their level of confidence in their classification. Figure 2 shows the plots of accuracy and confidence ratings as a function of amount of text revealed (number of ‘regions’, with each region corresponding to two content words). The data are averaged across 20 participants.

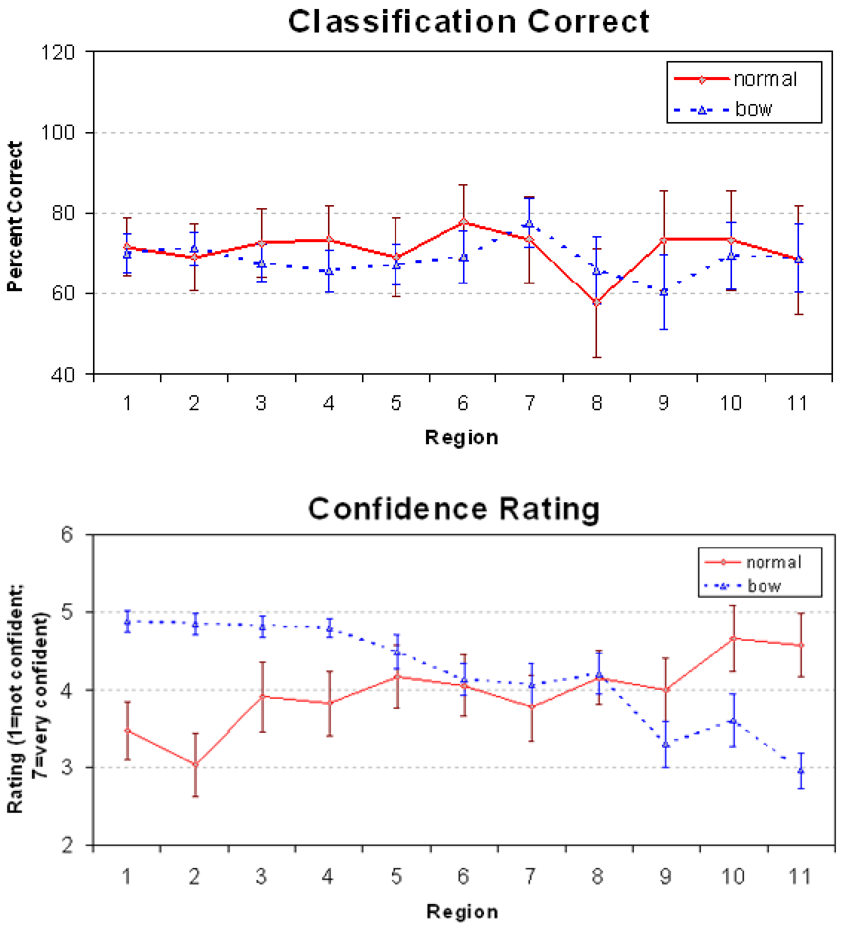

Figure 2. Results from text-classification experiments using progressive revealing of passages. The top and bottom graphs show accuracy and confidence ratings, respectively, as a function of amount of text seen across the ‘normal format’ and ‘Bag of Words’ conditions. The data are averaged across 20 participants.

Again, the classification-accuracy data show no statistically significant differences across the raw text and randomized bag-of-words conditions. The confidence ratings data, however, present a familiar paradox: as expected, subjects’ confidence ratings improve as more of the text is revealed in the unscrambled condition. In the BOW condition, on the other hand, the confidence ratings exhibit a decline, even though classification performance stays statistically unaffected. Such dissociations between confidence and performance have previously been reported in the visual domain (Marcel, 1983; He *et al*., 1996; Koch and Braun, 1996).

Setting aside the issue of confidence ratings, these results suggest that the randomized bag of words representation is comparable in its ability to support basic text classification to the original ordered text.

The strategies for text classification we have considered here bear some interesting parallels to a fundamental issue in the domain of visual object recognition. There, researchers have debated whether objects are recognized on the basis of overall configuration or isolated features. This is somewhat analogous to text classification on the basis of overall layout and syntactic structure versus individual words. The emerging consensus in the physiology of vision is that feature-based processing with features at a range of complexity levels is important for object recognition (Riesenhuber and Poggio, 1999). A similar account appears to hold for text-classification tasks as well. The contribution of our study lies in demonstrating that for the kinds of classification tasks considered here, an unordered word based representation is sufficient to account for the remarkable performance that humans exhibit. This finding also helps explain the efficacy of speed-reading techniques and readers' ability to understand seemingly complex passages 'at a glance' (Potter, 1997). From the pragmatic perspective, this finding has interesting implications regarding the need, or lack thereof, of syntactic information for text classification. It suggests that machine based document classification systems can, in principle, achieve significant performance without needing to perform natural language processing – a task that has proven to be rather complex despite the significant research attention it has attracted (Charniak, 1993; Manning and Schuetze, 1999). The recent success of web search engines, such as Google, that use word frequency statistics for page classification (Henzinger *et al*., 2003), lends support to the results we have presented here.

Methods

Experiment 1: The first screen explained the five categories to the participants. Then, participants saw examples of each category, one formatted and one BOW per category. In the actual experiment, participants saw the texts in a pseudo-randomized order so that no two examples of one category appeared immediately after each other. Each participant saw a different random order of texts. After each text, subjects indicated under which category they classified the text they just saw. The names of the categories were displayed after each text.

For the NYT experiment, there were 8 documents per category. Four texts in each category were presented in a normal format including page layout; the other four texts were presented in BOW format. Thus, each participant saw 100 texts in total (5 categories x 4 BOW x 4 formatted x 5 presentation times). For the newsgroups experiment too, the total number of passages presented was 100, which included 20 passages for each of the five classes. 10 of the passages per class were presented as BOW, and the remaining 10 were normally formatted.

The CNN experiment comprised a total of 50 passages, 10 per class (5 BOW and 5 normal). 10 subjects participated in the NYT experiment, and 5 in both newsgroups and CNN experiments.

Acknowledgments

The authors wish to thank Ethan Meyers for his help in data collection. TP is supported in part by the Eugene McDermott Chair in the Dept. of Brain & Cognitive Sciences. PS is supported by a John Merck Scholars Award.

References

Argamon, S., Koppel, M. and Avneri, G. (1998). Routing documents according to style. In *First International Workshop on Innovative Information Systems*.

Charniak, E. (1993). *Statistical Language Learning*, MIT Press:Cambridge, MA.

Durstenfeld**,** R. (1964). *Algorithm 235: Random permutation*, CACM 7(7):420.

Fisher, R. A. and Yates**,** F. (1938). Example 12, Statistical Tables, London.

He, S., Cavanagh, P. and Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature*, **383**: 334-337.

Henzinger, M., Chang, B. W., Milch, B. and Brin, S. (2003). Query –Free news search. *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary.

Koch, C. and Braun, J. (1996). On the functional anatomy of visual awareness. *Cold Spring Harbor Symposium on Quantitative Biology*, **61**: 49-57.

Laming, D. R. J. (1968). *Information theory and choice-reaction time*. London: Academic Press.

Link, S. W. and Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, **40**, 77-105.

Manning, C. D. and Schuetze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marcel, A. J. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology*, **15**, 197-237.

Mitchell, T. M. (1997). *Machine learning.* New York: McGraw-Hill.

Potter, M. C. (1997). Understanding sentences and scenes: The role of conceptual short term memory. In: *Fleeting memories*, V Coltheart (Ed.). Cambridge, MA: MIT Press.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, **85**, 59-108.

Rauber, A. and Muller-Kogler, A. (2001). Integrating automatic genre analysis into digital libraries. In *First ACM-IEEE Joint Conf on Digital Libraries*.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, **2**, 1019-1025.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1):1–47.

Vickers, D. (1979). *Decision Processes in Visual Perception.* New York, NY: Academic Press.

Wall, L., Christiansen, T., Orwant, J. (2000). Programming Perl, Third Edition, O'Reilly Publishers.