Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge

Evgeniy Gabrilovich and **Shaul Markovitch**
Department of Computer Science
Technion—Israel Institute of Technology, 32000 Haifa, Israel
{gabr,shaulm}@cs.technion.ac.il

Abstract

When humans approach the task of text categorization, they interpret the specific wording of the document in the much larger context of their background knowledge and experience. On the other hand, state-of-the-art information retrieval systems are quite *brittle*—they traditionally represent documents as bags of words, and are restricted to learning from individual word occurrences in the (necessarily limited) training set. For instance, given the sentence "Wal-Mart supply chain goes real time", how can a text categorization system know that Wal-Mart manages its stock with RFID technology? And having read that "Ciprofloxacin belongs to the quinolones group", how on earth can a machine know that the drug mentioned is an antibiotic produced by Bayer? In this paper we present algorithms that can do just that. We propose to enrich document representation through automatic use of a vast compendium of human knowledge—an encyclopedia. We apply machine learning techniques to Wikipedia, the largest encyclopedia to date, which surpasses in scope many conventional encyclopedias and provides a cornucopia of world knowledge. Each Wikipedia article represents a *concept*, and documents to be categorized are represented in the rich feature space of words and relevant Wikipedia concepts. Empirical results confirm that this knowledge-intensive representation brings text categorization to a qualitatively new level of performance across a diverse collection of datasets.

Introduction

From time immemorial, the human race strived to organize its collective knowledge in a single literary work. From "Naturalis Historiae" by Pliny the Elder to the contemporary mammoth "Encyclopaedia Britannica", encyclopedias have been major undertakings to systematically assemble all the knowledge available to the mankind.

Back in the early years of AI research, Buchanan & Feigenbaum (1982) formulated the *knowledge as power hypothesis*, which postulated that "The power of an intelligent program to perform its task well depends primarily on the quantity and quality of knowledge it has about that task." Lenat *et al.* (1990) argued that without world knowledge computer programs are very *brittle*, and can only carry out tasks that have been fully foreseen by their designers.

When computer programs face tasks that require human-level intelligence, it is only natural to use an encyclopedia to endow the machine with the breadth of knowledge available to humans. There are, however, several obstacles on the way to using encyclopedic knowledge. First, such knowledge is available in textual form, and using it requires natural language understanding, a major problem in its own right. Furthermore, language understanding may not be enough, as texts written *for* humans normally assume the reader possesses a large amount of common-sense knowledge, which is omitted even from most detailed encyclopedia articles (Lenat 1997). To address this situation, Lenat and his colleagues launched the CYC project, which aims to explicitly catalog the common sense knowledge of the humankind.

In this work we propose and evaluate a way to render text categorization systems with true encyclopedic knowledge, based on the largest encyclopedia available to date—Wikipedia (www.wikipedia.org). *Text categorization* deals with automatic assignment of category labels to natural language documents. The majority of existing text classifiers use machine learning techniques and represent text as a *bag of words (BOW)*, that is, the features in document vectors represent weighted occurrence frequencies of individual words (Sebastiani 2002). The BOW method is very effective in easy to medium difficulty categorization tasks where the category of a document can be identified by several easily distinguishable keywords. However, its performance becomes quite limited for more demanding tasks, such as those dealing with small categories or short documents. Our aim is to empower machine learning techniques for text categorization with a substantially wider body of knowledge than that available to a human working on the same task. This abundance of knowledge will to some extent counterbalance the superior inference capabilities of humans.

To tap into the kind of knowledge we described above, we build an *auxiliary* text classifier that is capable of matching documents with the most relevant articles of Wikipedia. We then augment the conventional bag of words with new features, which correspond to the concepts represented by these articles. Representing documents for text categorization in this knowledge-rich space of words and constructed features leads to substantially greater categorization accuracy.

Let us illustrate the importance of external knowledge with a couple of examples. Given a very brief news title

"Bernanke takes charge", a casual observer can infer little information from it. However, using the algorithm we developed for consulting Wikipedia, we find out the following relevant concepts: BEN BERNANKE, FEDERAL RESERVE, CHAIRMAN OF THE FEDERAL RESERVE, ALAN GREENSPAN (Bernanke's predecessor), MONETARISM (an economic theory of money supply and central banking), INFLATION and DEFLATION. As another example, consider the title "Apple patents a Tablet Mac". Unless the reader is well-versed in the hi-tech industry and gadgets, she will likely find it hard to predict the contents of the news item. Using Wikipedia, we identify the following related concepts: MAC OS (the Macintosh operating system) LAPTOP (the general name for portable computers, of which Tablet Mac is a specific example), AQUA (the GUI of MAC OS X), IPOD (another prominent product by Apple), and APPLE NEWTON (the name of Apple's early personal digital assistant).

Observe that documents manipulated by a text categorization system are given *in the same form* as the encyclopedic knowledge we intend to use—plain text. Therefore, we can use text similarity algorithms to automatically identify encyclopedia articles relevant to each document, and then leverage the knowledge gained from these articles in subsequent processing. It is this key observation that allows us to circumvent the obstacles we enumerated above, and use encyclopedia directly, without the need for deep language understanding or pre-cataloged common-sense knowledge. Also, it is essential to note that we do *not* use encyclopedia to simply increase the amount of the training data for text categorization; neither do we use it as a text corpus to collect word cooccurrence statistics. Rather, we use the knowledge distilled from the encyclopedia to enrich the representation of documents, so that a text categorizer is induced in the augmented, knowledge-rich feature space.

Our approach belongs to the field of constructive induction—the study of methods that endow the learner with the ability to modify or enhance the representation language. Feature generation techniques, which search for new features that describe the target concept better than the ones supplied with the training instances, were found useful in a variety of machine learning tasks (Markovitch & Rosenstein 2002; Fawcett 1993; Matheus 1991). To date, few studies applied feature generation techniques to text processing, and those that did so using external knowledge, mostly used structured knowledge repositories, such as WordNet (Scott 1998; Urena-Lopez, Buenaga, & Gomez 2001) or the Open Directory Project (Gabrilovich & Markovitch 2005).

The contributions of this paper are twofold. First, we propose a way for machine learning techniques to tap into a major encyclopedic resource without the need for specially crafted understanding or inference rules. Importantly, our method does not require any ancillary common-sense knowledge in order to "read" the encyclopedia, as we do so by using standard statistical learning techniques. Second, we evaluate the benefit of empowering inductive learning techniques with extensive world knowledge, using the task of text categorization as our application domain. Empirical evaluation proves that using background knowledge results in notable improvements in a host of different datasets.

Feature Generation with Wikipedia

To extend the representation of documents for text categorization, we propose building a feature generator that augments the bag of words with knowledge concepts relevant to the document text. The feature generator operates prior to text categorization, and is completely independent of the latter; therefore, it can be built once and reused in many future text categorization tasks.

Wikipedia as a Knowledge Repository

What kind of knowledge repository should be used for feature generation? In our earlier work (Gabrilovich & Markovitch 2005), we assumed the external knowledge is available in the form of a generalization hierarchy, and used the Open Directory Project (ODP, `www.dmoz.org`) as an example. This method, however, had a number of drawbacks, which we propose to overcome in this paper.

First, the knowledge repository was required to define an "is-a" hierarchy, thus limiting the choice of appropriate repositories. Moreover, hierarchical organization embodies only one particular relation between the nodes (generalization), while numerous other relations, such as relatedness, meronymy/holonymy and chronology, are ignored. Second, large-scale hierarchies tend to be extremely unbalanced, so that the relative size of some branches is disproportionately large or small due to peculiar views of the editors. Such phenomena are indeed common in the ODP; for example, the TOP/SOCIETY branch is heavily dominated by one of its children—RELIGION AND SPIRITUALITY. Finally, to learn the scope of every ODP category, short textual descriptions of the categories were augmented by crawling the Web sites cataloged in the ODP. This procedure allowed us to accumulate many gigabytes worth of textual data, but at a price, as texts obtained from the Web are often plagued with noise.

In this work we propose to perform feature generation using Wikipedia, which is currently the largest knowledge repository on the Web. It contains 300+ million words in nearly one million articles, contributed by over 160,000 volunteer editors. Even though Wikipedia editors are not required to be established researchers or practitioners, the open editing approach yields remarkable quality. A recent study (Giles 2005) found Wikipedia accuracy to rival that of Encyclopaedia Britannica. Another benefit of this openness is scalability—Britannica is about an order of magnitude smaller, with 44 million words in 65,000 articles (`http://store.britannica.com`, visited on 10.02.06).

Wikipedia has several advantages over the ODP. First, its articles are much cleaner than typical Web pages, and mostly qualify as standard written English. Although Wikipedia offers several orthogonal browsing interfaces, their structure is fairly shallow, and we propose to treat Wikipedia as having essentially no hierarchy. This way, mapping documents onto relevant Wikipedia concepts yields truly multi-faceted classification of the document text, and avoids the problem of unbalanced hierarchy branches. Moreover, by not requiring the knowledge repository to be hierarchically organized, our approach is suitable for new domains, for which no ontology is available. Finally, Wikipedia articles are heavily cross-linked, in a way reminiscent of linking on the Web.

We believe that these links encode many interesting relations between the concepts, and constitute an important source of information in addition to the article texts.

Feature Construction

The feature generator acts similar to a text classifier: it receives a text fragment, and maps it to the most relevant Wikipedia articles. For example, feeding the title of this paper as input to the feature generator yields the following relevant concepts: ENCYCLOPEDIA, WIKIPEDIA, ENTERPRISE CONTENT MANAGEMENT, BOTTLENECK, PERFORMANCE PROBLEM, and HERMENEUTICS. It is crucial to note that the concepts produced by the feature generator are orthogonal to the labels assigned in text categorization, and their sole purpose is to enrich the representation language.

We now enhance the standard text categorization process with feature construction. Training documents are first passed to the feature generator, which produces relevant Wikipedia concepts that augment the bag of words. Feature selection subsequently decimates the augmented feature set. Testing documents undergo a similar feature construction process, which ultimately allows to categorize them more reliably using the model induced from augmented training documents. Naturally, feature selection statistics is only computed using training documents.

We believe that considering the entire document as a single unit for feature generation can often be misleading, as its text might be too diverse to be readily mapped into a coherent set of concepts, while notions mentioned only briefly may be overlooked. Therefore, we perform feature generation using a *multi-resolution* approach. Features are generated for each document first at the level of individual words, followed by sentences, paragraphs, and finally the entire document. Working on individual contexts implicitly performs *word sense disambiguation*, and addresses one of the most important problems in natural language processing—*polysemy*. A context that contains polysemous words is mapped to the concepts that correspond to the sense *shared* by the context words. Thus, the correct sense of each word is determined with the help of its neighbors.

As an example, consider two sample phrases that contain the same ambiguous word in two different senses. For the input "jaguar car models", the Wikipedia-based feature generator returns the following concepts: JAGUAR (CAR), DAIMLER and BRITISH LEYLAND MOTOR CORPORATION (companies merged with Jaguar), V12 (Jaguar's engine), as well as concepts corresponding to specific Jaguar models such as JAGUAR E-TYPE and JAGUAR XJ. However, given the phrase "jaguar Panthera onca", we only get animal-related concepts, including JAGUAR, FELIDAE (feline species family), related felines such as LEOPARD, PUMA and BLACK PANTHER, as well as KINKAJOU (another endangered carnivore species).

Under the multi-resolution scheme, the feature generator produces a large number of features for each document. We rely on the subsequent feature selection step to eliminate extraneous features, keeping only those with high discriminative capacity for the ultimate categorization task. Furthermore, it is feature selection that allows the feature generator to be less than a perfect classifier. When some of the concepts assigned to the document are correct, feature selection can identify them and eliminate the spurious ones.

Feature generator design

Although Wikipedia has almost a million articles, not all of them are equally useful for feature generation. Some articles correspond to overly specific concepts (e.g., METNAL, the ninth level of the Mayan underworld), or are otherwise unlikely to be useful for subsequent text categorization (e.g., a list of events in a particular year). Other articles are just too short, so we cannot reliably classify texts onto the corresponding concepts. We developed a set of simple heuristics for pruning the set of concepts, by discarding articles that have fewer than 100 non stop words or fewer than 5 incoming and outgoing links. We also discard articles that describe specific dates, as well as Wikipedia disambiguation pages.

The feature generator performs classification of texts onto Wikipedia concepts. However, this is a very peculiar classification problem with hundreds of thousands of classes, each having a single positive example—the article text. Conventional induction techniques can hardly be applied in these settings, so we opted to use a simple and efficient centroid classifier (Han & Karypis 2000), which represents each concept with an attribute[1] vector of the article text.

When using a centroid classifier, it is essential to perform attribute selection to reduce noise. However, since we only have a single article for each concept, standard attribute selection techniques cannot be applied, so we postpone noise control to the next step. Each concept is represented as an attribute vector, whose entries are assigned weights using a TF.IDF scheme. Then, we build an *inverted index*, which maps each attribute into a list of concepts in which it appears. The primary purpose of inverted index is to speed up vector matching. In addition to that, we use it to discard insignificant associations between attributes and concepts. This is done by removing those concepts whose weights for a given attribute are too low. This scheme allows us to circumvent the scarceness of text objects for each concept—we cast the problem of attribute selection per concept as *concept selection* per attribute.

Using the link structure

It is only natural for an electronic encyclopedia to provide cross-references in the form of hyperlinks. As a result, a typical Wikipedia article has many more links to other entries than articles in conventional printed encyclopedias.

This link structure can be used in several ways. Observe that each link is associated with an *anchor text* (clickable highlighted phrase). The anchor text is not always identical to the canonical name of the target article, and different anchor texts are used to refer to the same article in different contexts. Thus, anchor texts provide alternative names, variant spellings, and related phrases for the target concept, which we use to enrich the article text for each concept.

[1] We use the term "features" to denote entries of document vectors in text categorization. To avoid confusion, we use the term "attributes" to denote entries of vectors that represent concepts during feature generation.

Similarly to the WWW, incoming links contribute to the significance of an article. We use the number of incoming links to express a slight preference for more significant concepts in feature generation.

Given a concept, we would like to use related articles to enrich its text. However, indiscriminately taking all articles pointed from a concept is ill-advised, as this would collect a lot of weakly related material. We believe that using similarity analysis and focused crawling techniques can greatly enrich concept representation with that of strongly related concepts, and plan to pursue this direction in our future work.

Empirical Evaluation

We implemented the proposed methodology using a Wikipedia snapshot as of November 5, 2005.

Implementation Details

After parsing the Wikipedia XML dump, we obtained 1.8 Gb of text in 910,989 articles. Upon removing small and overly specific concepts, 171,332 articles were left for feature generation. We processed the text of these articles by removing stop words and rare words (occurring in fewer than 3 articles), and stemmed the remaining words; this yielded 296,157 distinct terms, which were used to represent concepts as attribute vectors. Features were generated from the 10 best-matching Wikipedia concepts for each context.

Experimental Methodology

The following test collections were used:

1. Reuters-21578 (Reuters 1997). Following common practice, we used the ModApte split (9603 training, 3299 testing documents) and two category sets, 10 largest categories and 90 categories with at least one training and testing example.

2. Reuters Corpus Volume I (RCV1) (Lewis *et al.* 2004) has over 800,000 documents. To speed up the experiments, we used a subset of RCV1 with 17808 training documents (dated 20–27/08/96) and 5341 testing ones (28–31/08/96). Following Brank *et al.* (2002), we used 16 Topic and 16 Industry categories that constitute representative samples of the full groups of 103 and 354 categories, respectively. We also randomly sampled the Topic and Industry categories into 5 sets of 10 categories each.[2]

3. OHSUMED (Hersh *et al.* 1994) is a subset of MEDLINE, which contains 348,566 medical documents. Each document contains a title, and about two-thirds (233,445) also contain an abstract. Each document is labeled with an average of 13 MeSH[3] categories (out of total 14,000). Following Joachims (1998), we used a subset of documents from 1991 that have abstracts, taking the first 10,000 documents for training and the next 10,000 for testing. To limit the number of categories for the experiments, we randomly generated 5 sets of 10 categories each.[2]

4. 20 Newsgroups (20NG) (Lang 1995) is a well-balanced dataset of 20 categories containing 1000 documents each.

5. Movie Reviews (Movies) (Pang, Lee, & Vaithyanathan 2002) defines a sentiment classification task, where reviews express either positive or negative opinion about the movies. The dataset has 1400 documents in two categories.

We used SVM[4] with a linear kernel as our learning algorithm, since it is considered to be the state of the art in the field (Sebastiani 2002). We measured text categorization performance using the precision-recall break-even point (BEP). For the Reuters and OHSUMED datasets we report both micro- and macro-averaged BEP, since their categories differ in size substantially.[5] Following established practice, we used a fixed data split for the Reuters and OHSUMED datasets, and consequently used macro sign test (S-test) (Yang & Liu 1999) to assess the statistical significance of differences in classifier performance. For 20NG and Movies we performed 4-fold cross-validation, and used paired t-test to assess the significance. We also used the Wilcoxon signed-ranks test to compare the baseline and the Wikipedia-based classifiers over multiple data sets (Demsar 2006).

The Effect of Feature Generation

We first demonstrate that the performance of basic text categorization in our implementation (column "Baseline" in Table 1) is consistent with other published studies (all using SVM). On Reuters-21578, Dumais *et al.* (1998) achieved micro-BEP of 0.920 for 10 categories and 0.870 for all categories. On 20NG, Bekkerman (2003) obtained BEP of 0.856. Pang *et al.* (2002) obtained accuracy of 0.829 on Movies. The minor variations in performance are due to differences in data preprocessing used in different systems. For RCV1 and OHSUMED, direct comparison with published results is more difficult, as we limited the category sets and the date span of documents to speed up experimentation.

Table 1 shows the results of using Wikipedia-based feature generation, with significant improvements ($p < 0.05$) shown in bold. We consistently observed larger improvements in macro-averaged BEP, which is dominated by categorization effectiveness on small categories. This goes in line with our expectations that the contribution of encyclopedic knowledge should be especially prominent for categories with few training examples. Categorization performance was improved for virtually all datasets, with notable improvements of up to 30.4% for RCV1 and 18% for OHSUMED. Using the Wilcoxon test, we found that the Wikipedia-based classifier is significantly superior to the baseline with $p < 10^{-5}$ in both micro- and macro-averaged cases. Given the performance plateau currently reached by the best text categorizers, these results clearly demonstrate the advantage of knowledge-based feature generation.

Classifying Short Documents

We conjectured that knowledge-based feature generation might be particularly useful for classifying short documents. To evaluate this hypothesis, we derived several datasets of short documents from the test collections described above. Recall that about one-third of OHSUMED documents have

[2]The full definition of the category sets we used is available at http://www.cs.technion.ac.il/~gabr/aaai2006-appendix.html.

[3]http://www.nlm.nih.gov/mesh

[4]SVM^{light} implementation (Joachims 1998).

[5]Micro-averaged BEP operates at the document level and is primarily affected by categorization performance on larger categories. Macro-averaged BEP averages results over categories, and thus small categories have large impact on the overall performance.

Dataset	Baseline		Wikipedia		Improvement	
	micro	macro	micro	macro	micro	macro
Reuters-21578 (10 cat.)	0.925	0.874	0.932	0.887	+0.8%	+1.5%
Reuters-21578 (90 cat.)	0.877	0.602	0.883	0.603	+0.7%	+0.2%
RCV1 Industry-16	0.642	0.595	0.645	**0.617**	+0.5%	**+3.7%**
RCV1 Industry-10A	0.421	0.335	**0.448**	**0.437**	**+6.4%**	**+30.4%**
RCV1 Industry-10B	0.489	0.528	**0.523**	**0.566**	**+7.0%**	**+7.2%**
RCV1 Industry-10C	0.443	0.414	**0.468**	**0.431**	**+5.6%**	**+4.1%**
RCV1 Industry-10D	0.587	0.466	0.595	0.459	+1.4%	-1.5%
RCV1 Industry-10E	0.648	0.605	0.641	0.612	-1.1%	+1.2%
RCV1 Topic-16	0.836	0.591	0.843	**0.661**	+0.8%	**+11.8%**
RCV1 Topic-10A	0.796	0.587	0.798	**0.682**	+0.3%	**+16.2%**
RCV1 Topic-10B	0.716	0.618	0.723	**0.656**	+1.0%	**+6.1%**
RCV1 Topic-10C	0.687	0.604	0.699	0.618	+1.7%	+2.3%
RCV1 Topic-10D	0.829	0.673	0.839	0.688	+1.2%	+2.2%
RCV1 Topic-10E	0.758	0.742	0.765	0.755	+0.9%	+1.8%
OHSUMED-10A	0.518	0.417	**0.538**	**0.492**	**+3.9%**	**+18.0%**
OHSUMED-10B	0.656	0.500	0.667	**0.534**	+1.7%	**+6.8%**
OHSUMED-10C	0.539	0.505	0.545	**0.522**	+1.1%	**+3.4%**
OHSUMED-10D	0.683	0.515	0.692	**0.546**	+1.3%	**+6.0%**
OHSUMED-10E	0.442	0.542	**0.462**	**0.575**	**+4.5%**	**+6.1%**
20NG	0.854		**0.862**		**+1.0%**	
Movies	0.813		**0.842**		**+3.6%**	

Table 1: The effect of feature generation

titles but no abstract, and can therefore be considered short documents "as-is." We used the same range of documents, but considered only those without abstracts; this yielded 4,714 training and 5,404 testing documents. For all other datasets, we created a short document from each original document by taking only the title of the latter (with the exception of Movie Reviews, where documents have no titles).

It should be noted, however, that substituting a title for the full document is a poor man's way to obtain a collection of classified short documents. When documents were first labeled with categories, the human labeller saw each document *in its entirety*. In particular, a category might have been assigned to a document on the basis of facts mentioned in its body, even though the information may well be missing from the (short) title. Thus, taking all the categories of the original documents to be "genuine" categories of the title is often misleading. However, because we know of no publicly available test collections of short documents, we decided to construct datasets as explained above. Importantly, OHSUMED documents without abstracts have been classified as such by humans; working with the OHSUMED-derived dataset can thus be considered a "pure" experiment.

Table 2 presents the results of this evaluation. In the majority of cases, feature generation yielded greater improvement on short documents than on regular documents. Notably, the improvements are particularly high for OHSUMED, where "pure" experimentation on short documents is possible. According to the Wilcoxon test, the Wikipedia-based classifier is significantly superior to the baseline with $p < 2 \cdot 10^{-6}$. These findings confirm our hypothesis that encyclopedic knowledge should be particularly useful when categorizing short documents, which are inadequately represented by the standard bag of words.

DATASET	Baseline		Wikipedia		Improvement	
	micro	macro	micro	macro	micro	macro
Reuters-21578 (10 cat.)	0.868	0.774	0.877	0.793	+1.0%	+2.5%
Reuters-21578 (90 cat.)	0.793	0.479	0.803	**0.506**	+1.3%	**+5.6%**
RCV1 Industry-16	0.454	0.400	**0.481**	**0.437**	**+5.9%**	**+9.2%**
RCV1 Industry-10A	0.249	0.199	**0.293**	**0.256**	**+17.7%**	**+28.6%**
RCV1 Industry-10B	0.273	0.292	**0.337**	**0.363**	**+23.4%**	**+24.3%**
RCV1 Industry-10C	0.209	0.199	**0.294**	**0.327**	**+40.7%**	**+64.3%**
RCV1 Industry-10D	0.408	0.361	**0.452**	**0.379**	**+10.8%**	**+5.0%**
RCV1 Industry-10E	0.450	0.410	**0.474**	**0.434**	**+5.3%**	**+5.9%**
RCV1 Topic-16	0.763	0.529	0.769	0.542	+0.8%	+2.5%
RCV1 Topic-10A	0.718	0.507	0.725	**0.544**	+1.0%	**+7.3%**
RCV1 Topic-10B	0.647	0.560	0.643	0.564	-0.6%	+0.7%
RCV1 Topic-10C	0.551	0.471	**0.573**	**0.507**	**+4.0%**	**+7.6%**
RCV1 Topic-10D	0.729	0.535	0.735	**0.563**	+0.8%	**+5.2%**
RCV1 Topic-10E	0.643	0.636	0.670	0.653	+4.2%	+2.7%
OHSUMED-10A	0.302	0.221	**0.405**	**0.299**	**+34.1%**	**+35.3%**
OHSUMED-10B	0.306	0.187	**0.383**	**0.256**	**+25.2%**	**+36.9%**
OHSUMED-10C	0.441	0.296	**0.528**	**0.413**	**+19.7%**	**+39.5%**
OHSUMED-10D	0.441	0.356	**0.460**	**0.402**	**+4.3%**	**+12.9%**
OHSUMED-10E	0.164	0.206	**0.219**	**0.280**	**+33.5%**	**+35.9%**
20NG	0.699		**0.749**		**+7.1%**	

Table 2: Feature generation for short documents

Conclusions and Future Work

We proposed a way to use extensive encyclopedic knowledge to improve document representation for text categorization. We do so by building a *feature generator*, which identifies the most relevant encyclopedia articles for each document, and uses the concepts corresponding to these articles to create new features that augment the bag of words. We implemented our methodology using Wikipedia, which is by far the largest encyclopedia in existence. Due to the vast amount of knowledge contained in Wikipedia, the enriched document representation contains information that could not be inferred from the training documents alone.

We succeeded to make use of an encyclopedia without deep language understanding and without relying on additional common-sense knowledge bases. This was made possible by applying standard text classification techniques to match document texts with relevant Wikipedia articles.

Empirical evaluation definitively confirmed the value of encyclopedic knowledge for text categorization across a range of datasets. Recently, the performance of the best text categorization systems became similar, as if a plateau has been reached, and previous work mostly achieved improvements of up to a few percentage points. Using Wikipedia allowed us to reap much greater benefits, with double-digit improvements observed on a number of datasets.

The proposed feature generation system is admittedly complex, as it makes use of huge amounts of world knowledge. We performed a series of ablation studies (not shown here for lack of space), which confirmed that all system components are indeed necessary. Notably, pruning the inverted index (concept selection) is vital in eliminating noise. Furthermore, the multi-resolution approach is essential, as it allows not to overlook important but briefly-mentioned aspects of the document, which might be lost when the document is only considered as a whole.

Putting our work in the context of earlier research, there

have been prior attempts to add semantics to conventional bag-of-words text processing. Deerwester *et al.* (1990) proposed Latent Semantic Indexing (LSI), which analyzes a large corpus of unlabelled text, and automatically identifies "concepts" using singular value decomposition. However, prior studies found that LSI can rarely improve the strong baseline established by SVM, and often even results in performance degradation (Wu & Gunopulos 2002; Liu *et al.* 2004). In contrast, our methodology relies on using concepts identified and described by humans.

Some studies used WordNet as a source of external knowledge (Scott 1998; Urena-Lopez, Buenaga, & Gomez 2001). Note, however, that WordNet was not originally designed to be a knowledge base, but rather a lexical database suitable for peculiar lexicographers' needs. Specifically, it has substantially smaller coverage than Wikipedia, while additional information about word senses (beyond their identity) is very limited. Consequently, using WordNet also rarely results in improvements over SVM performance.

In our earlier work (Gabrilovich & Markovitch 2005), we used the Open Directory as a structured knowledge base. Our new Wikipedia-based results are superior to that work on a number of datasets, and are comparable to it on others. Moreover, our new methodology imposes much fewer restrictions on suitable knowledge repositories, and does not assume the availability of an ontology.

This study is only the first step in automatically using Wikipedia as a knowledge resource. We believe that leveraging the high degree of cross-linking between Wikipedia articles will allow us to uncover important relations between concepts. In this work we capitalized on inter-article links by using anchor text and the number of incoming links to each article, and in our future work we intend to investigate more elaborate techniques for using the link structure.

The Wiki technology underlying the Wikipedia project is often used nowadays in a variety of open-editing initiatives. These include corporate intranets that use Wiki as a primary documentation tool, as well as numerous domain-specific encyclopedias on topics ranging from mathematics to Orthodox Christianity.[6] Therefore, we believe our methodology may be used for augmenting document representation in domains for which no ontologies exist. We further believe that our methodology may benefit many additional text processing tasks such as information retrieval.

Acknowledgments

This work was partially supported by funding from the EC-sponsored MUSCLE Network of Excellence (FP6-507752).

References

Bekkerman, R. 2003. Distributional clustering of words for text categorization. Master's thesis, Technion.

Brank, J.; Grobelnik, M.; Milic-Frayling, N.; and Mladenic, D. 2002. Interaction of feature selection methods and linear classification models. In *Workshop on Text Learning held at ICML-2002*.

Buchanan, B. G., and Feigenbaum, E. A. 1982. Forward. In Davis, R., and Lenat, D. B., eds., *Knowledge-Based Systems in Artificial Intelligence*. McGraw-Hill.

Deerwester, S.; Dumais, S.; Furnas, G.; Landauer, T.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391–407.

Demsar, J. 2006. Statistical comparison of classifiers over multiple data sets. *JMLR* 7:1–30.

Dumais, S.; Platt, J.; Heckerman, D.; and Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. In *CIKM'98*, 148–155.

Fawcett, T. 1993. *Feature Discovery for Problem Solving Systems*. Ph.D. Dissertation, UMass.

Gabrilovich, E., and Markovitch, S. 2005. Feature generation for text categorization using world knowledge. *IJCAI*, 1048–1053.

Giles, J. 2005. Internet encyclopaedias go head to head. *Nature* 438:900–901.

Han, E.-H. S., and Karypis, G. 2000. Centroid-based document classification: Analysis and experimental results. In *PKDD'00*.

Hersh, W.; Buckley, C.; Leone, T.; and Hickam, D. 1994. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *SIGIR'94*, 192–201.

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

Lang, K. 1995. Newsweeder: Learning to filter netnews. In *ICML'95*, 331–339.

Lenat, D. B.; Guha, R. V.; Pittman, K.; Pratt, D.; and Shepherd, M. 1990. CYC: Towards programs with common sense. *Communications of the ACM* 33(8).

Lenat, D. B. 1997. From 2001 to 2001: Common sense and the mind of HAL. In *HAL's Legacy*. The MIT Press. 194–209.

Lewis, D. D.; Yang, Y.; Rose, T.; and Li, F. 2004. RCV1: A new benchmark collection for text categorization research. *JMLR* 5.

Liu, T.; Chen, Z.; Zhang, B.; Ma, W.-y.; and Wu, G. 2004. Improving text classification using local latent semantic indexing. In *ICDM'04*, 162–169.

Markovitch, S., and Rosenstein, D. 2002. Feature generation using general constructor functions. *Machine Learning* 49(1).

Matheus, C. J. 1991. The need for constructive induction. In Birnbaum, L., and Collins, G., eds., *8th Int'l Workshop on Machine Learning*, 173–177.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP'02*, 79–86.

Reuters. 1997. *Reuters-21578 text categorization test collection, Distribution 1.0*. Reuters. `daviddlewis.com/resources/testcollections/reuters21578`.

Scott, S. 1998. Feature engineering for a symbolic approach to text classification. Master's thesis, U. Ottawa.

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

Urena-Lopez, L. A.; Buenaga, M.; and Gomez, J. M. 2001. Integrating linguistic resources in TC through WSD. *Computers and the Humanities* 35:215–230.

Wu, H., and Gunopulos, D. 2002. Evaluating the utility of statistical phrases and latent semantic indexing for text classification. In *ICDM'02*, 713–716.

Yang, Y., and Liu, X. 1999. A re-examination of text categorization methods. In *SIGIR'99*, 42–49.

[6] See `http://en.wikipedia.org/wiki/Category:Online_encyclopedias` for a longer list of examples.