Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008)

Concept-Based Feature Generation and Selection for Information Retrieval

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

Abstract

Traditional information retrieval systems use query words to identify relevant documents. In difficult retrieval tasks, however, one needs access to a wealth of background knowledge. We present a method that uses Wikipedia-based feature generation to improve retrieval performance. Intuitively, we expect that using extensive world knowledge is likely to improve recall but may adversely affect precision. High quality feature selection is necessary to maintain high precision, but here we do not have the labeled training data for evaluating features, that we have in supervised learning. We present a new feature selection method that is inspired by pseudo-relevance feedback. We use the top-ranked and bottom-ranked documents retrieved by the bag-of-words method as representative sets of relevant and non-relevant documents. The generated features are then evaluated and filtered on the basis of these sets. Experiments on TREC data confirm the superior performance of our method compared to the previous state of the art.

Introduction

Information retrieval (IR) systems traditionally use the bag-of-words (BOW) representation for queries and documents, and retrieve results by finding occurrences of query terms in indexed documents. Such approaches, however, cannot retrieve relevant documents that do not mention query terms explicitly, in particular when users enter very short queries, such as in Web search. To demonstrate the problem, let us examine the query "cosmic events," an actual query from topic 405 in the TREC-8 test collection (Voorhees and Harman 2000). This short query was among the most difficult ones in TREC-8, with a meager median average precision of 0.06. Clearly, a simple search for these terms alone will not retrieve relevant documents such as FT911-3681, a very short news clip talking about the discovery of a very bright quasar, which does not mention any of the query keywords.

Finding relevant documents using additional knowledge has been the focus of many IR studies. Early approaches attempted to enrich query and document representation using lexical relations encoded in a thesaurus. In our example, a resource such as Roget's Thesaurus offers synonyms like "planetary" or "universal" for "cosmic," and "happening" or "occasion" for "event." However, augmenting the original query with alternatives such as "universal happenings" or "planetary occasions" will only cause the query focus to drift away and degrade system performance. Indeed, previous research has shown improvement due to this approach to be inconsistent (Voorhees 1994), and successful only when applied manually (Gonzalo et al. 1998). Other researchers attempted to extract the additional knowledge from the target corpus itself using *local context analysis* (Xu and Croft 2000), from *structured* external knowledge resources such as Web directories (Norasetsathaporn and Rungsawang 2001; Ravindran and Gauch 2004), or from feedback generated from massive *unstructured* data such as query logs (Cui et al. 2003). The main obstacle to the accuracy of these methods was the low granularity of the structured knowledge on the one hand, and the amount of noise introduced by the unstructured data on the other hand.

In this work, we present MORAG[1], a new IR methodology based on *Explicit Semantic Analysis* (ESA), recently proposed by Gabrilovich and Markovitch (2006). ESA leverages extensive encyclopedic knowledge to enhance the classic BOW text representation with conceptual and semantically rich features. ESA is constructed from information sources that are comprehensive enough to be useful for information retrieval in numerous domains, yet granular enough to offer high specificity of representation.

The main contributions of this paper are twofold. First, we propose a new retrieval model that can enhance any existing BOW-based IR system with concept-based features. Second, we propose a new unsupervised feature selection algorithm that leverages the combined BOW-concepts structure to bootstrap a feature selection process that performs well without any training examples.

Background: Explicit Semantic Analysis

A basic element in the design of IR systems is the representation chosen for documents and queries. Traditionally, two main approaches were available for enhancing the basic

*Second author's current affiliation: Yahoo! Research, 2821 Mission College Blvd, Santa Clara, CA 95054. Email: gabr@yahoo-inc.com

[1]Morag is the Hebrew word for *flail*, an agricultural tool used to separate grain from husks.

BOW representation with semantic features:

- Construct a taxonomy of semantic concepts and relations, manually or automatically, and map documents and queries onto them (Norasetsathaporn and Rungsawang 2001; Ravindran and Gauch 2004). This method produces coherent concepts that humans can "understand" and reason about, but the representation granularity is limited by the taxonomy, which in turn is limited by the work and complexity required to build and maintain it.
- Extract semantic meaning from documents and queries by analyzing the latent relationships between text words (Deerwester et al. 1990). Such methods can uncover unlimited vocabularies of concepts in the texts, but the resulting concepts are very difficult for humans to interpret.

Recently, a third approach was proposed, which strives to achieve the best of both worlds. The novelty of this method, called ESA (Gabrilovich and Markovitch 2006), lies in its approach to concept encoding. Instead of mapping a text to a node (or a small group of nodes) in a taxonomy, it maps the text to the *entire* collection of available concepts, by computing the degree of affinity of each concept to the input text. The mapping function is automatically derived from the knowledge repository, and can treat with equal ease both very short and very long text fragments. This simple yet powerful idea breaks the barrier of representation granularity, while using real-life human-defined concepts. ESA was successfully implemented using repositories as comprehensive as the Open Directory Project (`dmoz.org`) and Wikipedia (`wikipedia.org`), and was shown to be effective in tasks such as text categorization (Gabrilovich and Markovitch 2006) and computing semantic similarity (Gabrilovich and Markovitch 2007).

Concepts identified by ESA can be used as additional features of texts in information retrieval. However, ESA is not without its limitations, and in particular ones that are harmful to the IR task. Consider, for example, the query "law enforcement, dogs" (TREC topic 426). The top 10 Wikipedia-based ESA features generated for this query are (in order of significance):

CONTRACT
DOG FIGHTING
POLICE DOG
LAW ENFORCEMENT IN AUSTRALIA
BREED-SPECIFIC LEGISLATION
CRUELTY TO ANIMALS
SEATTLE POLICE DEPARTMENT
LOUISIANA
AIR FORCE SECURITY FORCES
ROYAL CANADIAN MOUNTED POLICE

While some of the concepts generated, such as POLICE DOG, are strongly related to the query, others, such as CONTRACT or LOUISIANA seem unrelated. If we examine the Wikipedia articles from which the ESA classifier was built, we see several incidental mentions of the word "dog," which together with a strong relation to law enforcement were sufficient to trigger these concepts. If used to augment information retrieval, such "noisy" concepts will likely lead to documents that are completely unrelated to the query. Other concepts, such as DOG FIGHTING or BREED-SPECIFIC LEGISLATION are related not to police dogs, as was the intent of the TREC topic, but rather to the enforcement of various laws related to dogs. Without additional context, this ambiguity is passed from the query text to the ESA features, as well as to the BOW query. Yet, an ambiguous BOW query will still require the ambiguous term to appear in retrieved documents, whereas concept-based ambiguity further expands the potential error to a broader range of documents, with a far more detrimental effect. For example, the DOG FIGHTING feature might also be generated for documents that discuss illegal gambling or combat sports, even if dogs were not mentioned anywhere in the text.

Previous research applied ESA to text categorization, which is inherently a supervised learning task. Consequently, features generated by ESA could be filtered on the basis of their information gain on training examples (or any other feature selection metric) (Gabrilovich and Markovitch 2006). The information retrieval task offers no labeled training examples; automatically generated features might thus adversely affect precision because they can shift the retrieval focus away from the very specific information need. Indeed, as we will show later, system performance without feature selection is significantly inferior to that when feature selection is employed.

Semantically-Enhanced Information Retrieval

In this section we present a new methodology for enhancing existing IR systems with concept-based features. We begin with the general solution architecture, and then focus on its core, the feature selection module.

Solution Overview

MORAG uses the ESA implementation as described by Gabrilovich and Markovitch (2006), which constructs a feature generator from a Wikipedia dump, using Wikipedia articles as concepts. The ESA feature generator receives a fragment of text as input, and outputs a vector in the space of all Wikipedia concepts, where each element represents the relatedness of the corresponding concept to the input text.

The architecture of MORAG is presented in Figure 1. Each document is indexed using a BOW method. Additionally, the document is fed into the ESA feature generator, which produces a concept-based representation. We index the corpus both as entire documents, and in overlapping, fixed-size (50 words) passages. Indexing documents by passages is a common IR approach, but in our case it was particularly crucial: ESA works best when the text to be analyzed is focused on a single topic; otherwise, ESA tries to "average" several topics and the result is less coherent.

At retrieval time, we first rank the documents and passages based on the bag of words in the query. We then generate ESA features for the query, and perform feature selection as described below. The selected features are then used to produce a concept-based rank for the passages and documents. For both BOW-based ranking and concept-based ranking, we found that the best performance is obtained by computing a document's score as a sum of (1) the

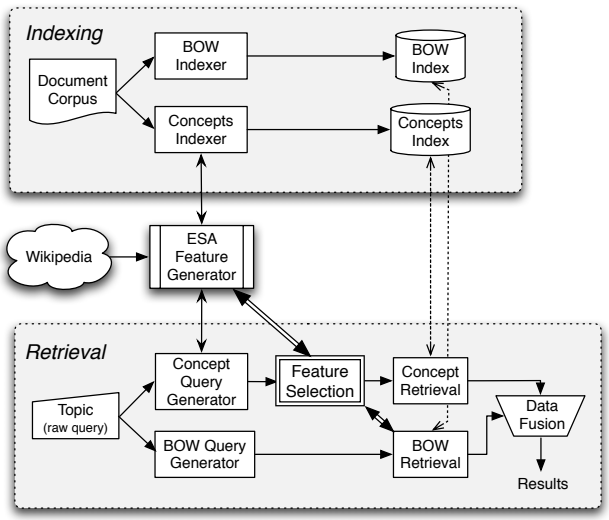

Figure 1: Solution architecture

score of the best passage in it, and (2) the score of the entire document. Explicitly considering the best-scoring passage works well because documents are often labeled relevant on the basis of only a few relevant sentences, even if the rest of the document text is irrelevant to the topic. Finally, we produce the merged ranking by first normalizing the document scores in these two ranked lists using the *fitting* method (Wu, Crestani, and Bi 2006), and then fusing them using linear combination of document scores (Vogt and Cottrell 1999). The document scores are calculated as $w \cdot ConceptsScore + (1-w) \cdot BowScore$ (w being a system parameter).

Feature Selection with Pseudo-Relevance Feedback

In supervised learning, features are usually evaluated and selected according to how well they separate the classes in the training set. Such a method is not applicable to general IR, where labeled examples are not available. There is, though, a supervised method in IR, called *relevance feedback* (Salton and Buckley 1990), where the user makes relevance judgments on initial results, and this feedback is used to get an improved set of results. Relevance feedback can be extended to the unsupervised case, by assuming that the top ranked documents are relevant. This method is called *pseudo-relevance feedback* (PRF).

The MORAG method for feature selection uses the same principle to evaluate features. The process begins with retrieving, using the base BOW method, a sequence of m documents, $D_Q = \langle d_1, \ldots, d_m \rangle$, that are sorted according to their relevance score. The top-k documents in D_Q will be used as a set of relevant examples $D_r = \langle d_1, \ldots, d_k \rangle$, and the bottom-$k$ documents in D_Q as a set of non-relevant examples $D_{nr} = \langle d_{m-k+1} \ \ldots \ d_m \rangle$.

The next step calculates how well each feature separates D_r from D_{nr}, by computing the entropy in two subsets induced by a threshold on the feature's strength in each of the example documents. We filter out features whose frequency is *negatively* correlated with relevancy, since they will later be used as query terms. A more formal description of the algorithm is given in Figure 2.

Procedure SELECT-FEATURES(Q, k, θ)
  $D_Q \leftarrow$ RETRIEVE-WITH-BOW(Q)
  $D_r \leftarrow$ first k documents in D_Q
  $D_{nr} \leftarrow$ last k documents in D_Q
  $F \leftarrow$ GENERATE-FEATURES(Q)
  Foreach $f \in F$
    $IG[f] \leftarrow 0$
    $Thresholds \leftarrow \{weight(f,d) | d \in D_r \cup D_{nr}\}$
    Foreach $t \in Thresholds$
      $S_\oplus \leftarrow \{d \in D_r \cup D_{nr} | weight(f,d) > t\}$
      $S_\ominus \leftarrow \{d \in D_r \cup D_{nr} | weight(f,d) \leq t\}$
      $IG_t \leftarrow 1 - \text{ENTROPY}(S_\oplus) * \frac{|S_\oplus|}{|D_r \cup D_{nr}|}$
        $-\text{ENTROPY}(S_\ominus) * \frac{|S_\ominus|}{|D_r \cup D_{nr}|}$
      $IG[f] \leftarrow \max(IG[f], IG_t)$
  Return fraction θ from F with highest IG value

Figure 2: Feature selection based on pseudo-relevance, with 2k examples and θ selection aggressiveness

Empirical Evaluation

MORAG was implemented using an open source probabilistic IR library (Xapian, `xapian.org`), and evaluated using the TREC-8 test collection (Voorhees and Harman 2000). This dataset is composed of several document sets (mainly newswire), and contains about 528,000 documents, 50 topics (information need statements) and human relevance judgments for each topic. TREC topics are composed of several formulations of the information need (title, description, narrative). We chose to use the shortest (title), as it better reflects common uses of IR systems today (e.g. Web search), and because we believe MORAG shows the most improvement when information is scarce.

According to TREC methodology, the output of a system is an ordered list D of 1000 documents for each query. Performance is measured by Mean Average Precision (MAP), defined as $\frac{1}{|Q|} \sum_{q \in Q} \sum_{k=1}^{|D|} rel(d_k) \cdot Prec@k / R_q$, where rel is a boolean relevance function, $Prec@k$ is precision at cutoff k, and R_q is the total number of relevant documents in the corpus for query q. We compared performance of MORAG to our baseline BOW retrieval (Xapian), and to three top systems in TREC-8: Okapi (Robertson and Walker 1999), AT&T (Singhal et al. 1999), and PIRCS (Kwok, Grunfeld, and Chan 1999), all BOW retrieval systems.

Before testing on TREC-8, we performed parameter tuning using the TREC-7 dataset to determine, for each system, the values for θ (selection aggressiveness), k (PRF group size), and w (the linear combination weight). The obtained values for (k, θ, w), used as defaults for the rest of the experiments were: Xapian (15,20,0.5), Okapi (25,20,0.3), AT&T (10,30,0.4), and PIRCS (25,20,0.4).

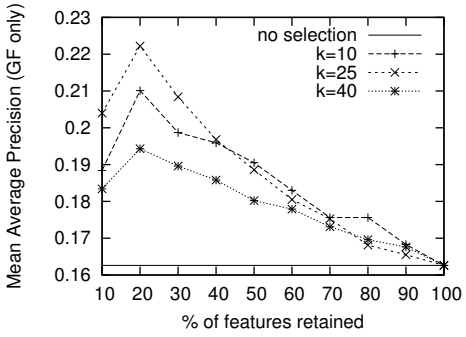

Figure 3: Concept-based performance as a function of a fraction of the concepts selected (θ)

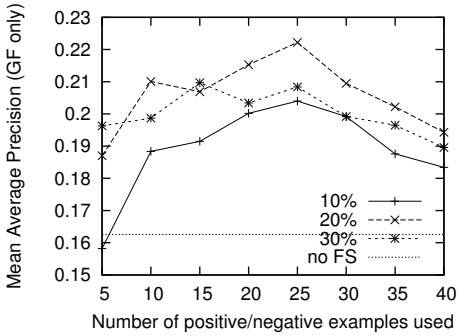

Figure 4: Concept-based performance as a function of the number of pseudo-relevance examples (k)

Effectiveness of Feature Selection

Our experiments show that the suggested feature selection method performs well. Figure 3 shows the retrieval effectiveness (using generated features only) for TREC-8 topics, as a function of the selection aggressiveness θ, for several values of k. For practical reasons of storage and computational resources, we used only the 50 most prominent features generated by the ESA method (hence 20% implies 10 features were used). The graph shows that IR performance steadily improves as more features are removed, and peaks around 20%-30% features remaining, for any value of k.

Figure 4 shows that, as expected, IR performance improves as k increases, but beyond a certain value it starts to decrease. This decrease can be attributed to the reduction in relevance of the lower ranked documents. Note that our selection method does not require all top-k documents to be perfectly relevant (i.e., judged as relevant in the dataset). For our purposes, it is sufficient that the density of relevant documents at the top of the ranking is higher than at the bottom. This density gap decreases as k increases.

The Effect of Concept-Based Retrieval

We evaluated concept-based retrieval in MORAG by comparing performance with and without the generated features for our baseline and for 3 other top TREC-8 participants. The results are presented in Table 1. For each system, we list its BOW-only performance (2nd column), followed by several results of enhancing it with MORAG. The 3rd column shows MORAG results *without* feature selection, the 4th column shows MORAG results with feature selection using the default parameter values, and the 5th column shows MORAG results using optimal parameter values. We used paired t-test to assess the statistical significance of results; significant improvements ($p < 0.05$) are shown in bold. We also conducted experiments on the TREC Robust 2004 dataset (Voorhees 2005), and achieved similarly significant improvements (10.7% over the BOW baseline on the set of 50 new queries, and 29.4% over baseline for the set of 50 hard queries). We omit additional details of this experiment owing to lack of space.

As Table 1 shows, there is less improvement gained when better-performing BOW methods are enhanced. This can be explained if we recall the relatively low performance of concept-based retrieval on its own, and the fact that successful data fusion (Lee 1997) requires the merged methods to have similar performance. We therefore believe that improving the basic concept-based performance in future work will lead to even better performance of fused results.

One question that may be posed in light of these positive results is to what extent these results may be attributed to fusion itself. Data fusion (and meta-search) initially originated from the observation that merging results from several sources has a synergetic potential to improve results, with formal reasoning later suggested by Lee (1997). Is it possible that the improvement is mostly, if not completely, attributable to the actual merging of an additional source of results, regardless of the conceptual nature of this source?

We investigated this important concern by using the same framework to fuse all pairs of BOW results mentioned. None of these pairs achieved results as high as our optimal results, and most of them were even lower than non-optimal results. The superiority of MORAG is even more impressive considering that the success of fusion greatly depends on the performance of the individual systems to be fused, and our concept-only retrieval currently achieves a relatively low MAP of 0.22. To demonstrate this point, we chose another TREC-8 run with a similarly low MAP, the RMIT/CSIRO system title-only run, and fused it with all tested systems. As the 6th column in Table 1 shows, results are significantly lower than those achieved by concept-based retrieval.

Finally, we note that evaluation of MORAG may suffer from a certain bias in TREC methodology that arises from its use of "pooling," which rather than scan the entire corpus for relevancy, only examines the top documents in submitted runs. Zobel (1998) showed that a substantial portion of relevant documents may not be discovered in this process, but relative ranking of evaluated systems would still be reliable. However, Zobel also pointed out that pooling could discriminate against methods based on novel principles, and recommended that researchers consider the number of non-judged documents fetched, as indication that performance is probably underestimated. Following Zobel's advice, we found that our concept-based run retrieved over 30% more non-judged documents than the evaluated BOW systems (which all produced similar numbers of non-judged documents). Hence, there is reason to assume that MORAG is indeed, to some extent, undervalued by TREC methodology.

Method	Baseline	+ Concepts (no selection)	+ Concepts	+ Concepts (optimal)	+ RMIT (optimal)
Xapian	0.2481	0.2526 (+1.8%)	**0.2864 (+15.4%)**	**0.2947 (+18.8%)**	0.2503 (+0.9%)
Okapi	0.2787	0.2844 (+2.0%)	**0.3042 (+9.1%)**	**0.3065 (+10.0%)**	**0.2921 (+4.8%)**
AT&T	0.2853	0.2832 (0.0%)	0.2977 (+4.3%)	**0.3088 (+8.2%)**	0.2943 (+3.2%)
PIRCS	0.3063	0.2965 (-3.2%)	0.3211 (+4.8%)	**0.3239 (+5.7%)**	0.3086 (+0.7%)

Table 1: Fusing concept-based retrieval, with various selection settings, vs. fusing BOW retrieval

Qualitative Analysis

To better understand the advantages of using conceptual features, let us revisit the introduction example. The first column in Table 2 shows the top 10 concepts generated from the query text "cosmic events," while the second column shows the top 10 features after feature selection. Examining the differences, we note that irrelevant concepts such as COSMIC ERA and MARVEL UNIVERSE (both comics-related) were pruned, as they were not triggered by any of the top results. Average precision for this topic in MORAG is 0.18, well above the 0.06 median of TREC-8 participants. One of the reasons for this success is the relatively high *recall* rate of 0.6, one of the highest obtained by TREC-8 participants for this topic. As an example of the causes for this high recall, consider the relevant document FT911-3681, of the following short content: "*A mysterious quasar far brighter than everything else in the universe has been discovered by British astronomers in the Canary Islands. It makes the sun appear pale.*" Standard BOW systems will have no evidence to match "cosmic events" to this text, and indeed it was not retrieved in the top 1000 documents by any of the discussed BOW systems. However, MORAG retrieves it using relevant features such as GAMMA RAY BURST and BIG BANG.

Another reason for the success of MORAG is the filtering of irrelevant documents that mention "cosmic" and other related words in unrelated contexts. For example, document LA012289-0002, titled "Cosmic Christ Comes to Mother Earth," discusses a book review and mentions the words "cosmic," "event," and potential query-expansion terms such as "Earth." This document is retrieved by Okapi and RMIT systems in the top 5 results, and by PIRCS and AT&T in the top 50. MORAG does not include it at all in its top 1000, as the features generated for it clearly do not match the astronomy-related query concepts.

Top-10 Raw Features	Top-10 Selected Features
COSMIC ERA	SOLAR COSMIC RAY
UNIVERSE	COSMIC RAY
SOLAR COSMIC RAY	NEUTRINO
COSMIC RAY	SOLAR VARIATION
NEUTRINO	COSMIC INFLATION
OH-MY-GOD PARTICLE	COSMIC MICROWAVE BACKGROUND RADIATION
MARVEL UNIVERSE	GAMMA RAY BURST
SOLAR VARIATION	COSMIC DUST
COSMIC INFLATION	GALACTIC COSMIC RAY
COSMIC MICROWAVE BACKGROUND RADIATION	WMAP

Table 2: Features generated and selected for TREC topic 405 ("cosmic events")

Related Work

Semantic feature generation for IR has traditionally revolved around using language thesauri such as WordNet to perform indexing using synsets (Gonzalo et al. 1998) and lexical query expansion (Voorhees 1994). In recent years, researchers have turned to larger-scale knowledge bases, such as Wikipedia and the Open Directory Project (ODP), as sources for enhancing IR. In some works, the extracted knowledge was used to expand textual queries with related terms (Milne, Witten, and Nichols 2007; Ratinov, Roth, and Srikumar 2008). Others used it to propose methods that replace BOW retrieval altogether, such as ESA-only retrieval (Ratinov, Roth, and Srikumar 2008) and retrieval based on semantic-relatedness between query and document terms (Gurevych, Muller, and Zesch 2007). A few works suggested a combined BOW-Concepts framework but were limited either by requiring additional information for constructing the concept-based query (Ravindran and Gauch 2004) or relying on structured entity recognition, which is limited by ontology coverage (Bast et al. 2007). MORAG uses extracted knowledge and ESA representation for concept-based retrieval integrated with classic BOW retrieval methods, thus allowing the previous state of the art to be enhanced rather than replaced. In addition, MORAG incorporates a novel method of unsupervised feature selection to optimize conceptual query representation.

Feature selection is very common in text categorization and other classification tasks, but much less so in text retrieval. Retrieval of non-textual items, such as music or images (Dy et al. 2003), often involves feature extraction, and therefore a selection step is occasionally applied as well. In mainstream text retrieval, the term "feature selection" is mostly used only in the context of choosing characteristics of an IR system that best fit a specific problem (Fan, Gordon, and Pathak 2004; Metzler 2007).

PRF in IR can be considered a variant of feature selection, where the candidate features are the words appearing in pseudo-relevant documents, and the selection process is extracting those that co-occur with the query words (Rocchio 1971). Pseudo-relevance information was extended to also use external resources such as the Web (Kwok et al. 2005) or query logs (Cui et al. 2003). PRF was also applied to statistical language modeling in IR, with recent work further extending it to iterative methods (Kurland, Lee, and Domshlak

2005) and to multi-word expansion concepts (Metzler and Croft 2007). The differences between our method and those based on PRF go beyond such trivial issues as using the examples to extract features rather than *measuring* utility of features generated from another source. More importantly, our method strives to *select* features according to their potential for separating relevant documents from non-relevant ones, rather than merely tweak feature weights, and the latter approach may be more influenced by outliers in the term weights in the documents.

Discussion

We described MORAG, a framework for enhancing traditional BOW-based IR systems with concept-based features derived from Wikipedia, and demonstrated that it leads to substantial performance gains. We proposed to alleviate the noise introduced by generated features (especially detrimental to IR) by a novel unsupervised learning algorithm that filters the generated features to match the specific query-corpus context, using the BOW results as examples.

The architecture of the proposed solution is modular, so our approach can be applied to any existing IR system; indeed, we show improvement on several other baselines besides ours. Improvement was most significant when baseline results were low, and decreased as baselines improved. We attribute this to the relatively low performance of concept-based retrieval alone, which can be partly attributed to an inherent bias in the TREC "pooling" evaluation method.

In light of the encouraging results, we believe that MORAG, and concept-based retrieval based on massive knowledge bases in general, add the world knowledge that is vital in processing many information requests, and pave the way to potential major improvements in IR performance.

Acknowledgements. We thank Don Metzler for discussions that helped us improve the paper.

References

Bast, H.; Chitea, A.; Suchanek, F.; and Weber, I. 2007. Ester: Efficient search on text, entities, and relations. In *SIGIR*, 671–678.

Cui, H.; Wen, J.-R.; Nie, J.-Y.; and Ma, W.-Y. 2003. Query expansion by mining user logs. *IEEE Transactions on Knowledge and Data Engineering* 15(4):829–839.

Deerwester, S. C.; Dumais, S. T.; Landauer, T. K.; Furnas, G. W.; and Harshman, R. A. 1990. Indexing by latent semantic analysis. *JASIS* 41(6):391–407.

Dy, J. G.; Brodley, C. E.; Kak, A.; Broderick, L. S.; and Aisen, A. M. 2003. Unsupervised feature selection applied to content-based retrieval of lung images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 25(3):373–378.

Fan, W.; Gordon, M. D.; and Pathak, P. 2004. Discovery of context-specific ranking functions for effective information retrieval using genetic programming. *IEEE Transactions on Knowledge and Data Engineering* 16(4):523–527.

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

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

Gonzalo, J.; Verdejo, F.; Chugur, I.; and Cigarrin, J. 1998. Indexing with wordnet synsets can improve text retrieval. In *COLING/ACL Workshop on Usage of WordNet for NLP*.

Gurevych, I.; Muller, C.; and Zesch, T. 2007. What to be? - electronic career guidance based on semantic relatedness. In *ACL*, 1032–1039.

Kurland, O.; Lee, L.; and Domshlak, C. 2005. Better than the real thing? iterative pseudo-query processing using cluster-based language models. In *SIGIR*, 19–26.

Kwok, K.; Grunfeld, L.; Sun, H.; and Deng, P. 2005. Robust track experiments using PIRCS. In *TREC-13*.

Kwok, K.; Grunfeld, L.; and Chan, M. 1999. Trec-8 ad-hoc, query and filtering track experiments using pircs. In *TREC*, 217–228.

Lee, J.-H. 1997. Analyses of multiple evidence combination. In *SIGIR*, 267–276.

Metzler, D., and Croft, W. B. 2007. Latent concept expansion using markov random fields. In *SIGIR*, 311–318.

Metzler, D. A. 2007. Automatic feature selection in the markov random field model for information retrieval. In *CIKM*, 253–262.

Milne, D. N.; Witten, I. H.; and Nichols, D. M. 2007. A knowledge-based search engine powered by wikipedia. In *CIKM*, 445–454.

Norasetsathaporn, P., and Rungsawang, A. 2001. Kasetsart university TREC-10 experiments. In *TREC-10*, 339–346.

Ratinov, L.; Roth, D.; and Srikumar, V. 2008. Conceptual search and text categorization. Technical Report UIUCDCS-R-2008-2932, UIUC, CS Dept.

Ravindran, D., and Gauch, S. 2004. Exploiting hierarchical relationships in conceptual search. In *CIKM*, 238–239.

Robertson, S. E., and Walker, S. 1999. Okapi/keenbow at trec-8. In *TREC*, 151–162.

Rocchio, J. J. 1971. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall. 313–323.

Salton, G., and Buckley, C. 1990. Improving retrieval performance by relevance feedback. *JASIS* 41(4):288–297.

Singhal, A.; Abney, S.; Bacchiani, M.; Collins, M.; Hindle, D.; and Pereira, F. 1999. At&t at trec-8. In *TREC*, 317–330.

Vogt, C. C., and Cottrell, G. W. 1999. Fusion via a linear combination of scores. *Inform. Retrieval* 1(3):151–173.

Voorhees, E. M., and Harman, D. 2000. Overview of the eighth text retrieval conference (trec-8). In *TREC*, 1–24.

Voorhees, E. M. 1994. Query expansion using lexical-semantic relations. In *SIGIR*, 61–69.

Voorhees, E. M. 2005. Overview of the TREC 2004 robust retrieval track. In *TREC*, 70–79.

Wu, S.; Crestani, F.; and Bi, Y. 2006. Evaluating score normalization methods in data fusion. In *AIRS*, 642–648.

Xu, J., and Croft, W. B. 2000. Improving the effectiveness of information retrieval with local context analysis. *ACM TOIS* 18(1):79–112.

Zobel, J. 1998. How reliable are the results of large-scale information retrieval experiments? In *SIGIR*, 307–314.