Exploring Traits of Adjectives to Predict Polarity Opinion in BLOGS and Semantic Filters in Genomics

Miguel E Ruiz [1], Ying Sun [2], Jianqiang Wang [2] and Hongfang Liu [3]

[1] University of North Texas

School of Library and Information Sciences

meruiz@unt.edu

[2] University at Buffalo

Department of Library and Information Studies

{sun3, jw254}@buffalo.edu

[3] Georgetown University Medical Center

Department of Biostatistics, Bioinformatics, and Biomathematics

hl224@georgetown.edu

Abstract:

This paper presents the results of our team in the Genomics and Blog tracks in TREC 2007. We used the language model implementation provided by Indri for both tracks. For the BLOG track we explored the use of adjectives with in a post as a way to predict opinion polarity. Our work in the Genomics track explores two approaches to generate queries from the original topics. The first approach performs automatic term expansion using UMLS to generate a structured query that can be submitted using Indri's query language. The second approach uses a query expansion and re-ranking method based on identification of semantic relatives. This approach tries to capture the semantic of the potential answer, key terms in the topic and detection of gene/protein terms mentioned in the topic.

1. Introduction

This paper presents the work of our team that was registered as the University at Buffalo, State University of New York team. This year we participated in two tracks BLOG and Genomics. For this year we decided to use indri as our search engine and tried to concentrate on ways to expand and construct rich queries. This paper is divided into 2 main sections. Section 2 presents our work in the blogs track while section 3 presents our work in the Genomics track.

2. BLOG Track

In traditional and most current information retrieval systems, topical relevance is the most popular and widely accepted property of texts that is used to retrieve and rank documents. These topic-only IR systems are built and evaluated upon the underlying assumption that information is relevant if it is about the topic expressed in the user's information need. Researchers have noticed and studied the various dimensions of users' information needs for a long period of time. Until recently, the tremendous amount information and various information needs, particularly in the World Wide Web, has provided the motivation of research on mechanisms to complement current topic-driven retrieval systems by using non-topical properties of text. The TREC BLOG track provides a platform to study one non-topical property: opinion orientation of texts.

In our first year BLOG track, the University at Buffalo team primarily focused on the discriminating power of various classes of adjectives. Two sets of adjective lists are tested: (1) the initial manually picked subjective adjectives in Wiebe's landmark study (2000a, 2000b) on the acquisition of subjective adjectives and (2) trait adjectives listed by Peabody and De Raad (2002) in their psychology study of adjectives as indicators of various dimensions of human personality. Wiebe et al. (1999) found that the mere presence of one or more adjectives in a sentence is one of several useful predictors of text's subjectivity. The Wiebe list includes adjectives that can be used to describe various types of objects. Our hypothesis is that a more specific adjective list for each type of target, for example person, will be more efficient as indicators of the polarity of documents. With the time and recourse limitation, we cannot develop a systematic classification of targets and identify adjectives for each category. Since person is a major type of target in the BLOG opinion tasks, we narrower our goal for this year to test if the trait adjectives for human personality will do better in identifying the polarity property of a document with regard to a person target than a general subjective adjective lists do.

2.1. Experiments

We used only the permalink data in the collection. We finish our task in two steps. For each run, we first use Lemur to retrieve 1000 documents for each query. All retrieved documents for the 50 queries are pooled together and indexed. This small collection is retrieved again with a set of two queries: positive query with positive adjectives and negative query with negative adjectives. A number is arbitrarily chosen to decide how many documents should be retrieved for the positive, as well as the negative query. The opinion score of each retrieved document is assigned as below:

If d_i is retrieved by both the positive query and negative query, $S(d_i) = 3$

If d_i is only retrieved by the positive query, $S(d_i) = 4$

If d_i is only retrieved by the negative query, $S(d_i) = 2$

If d_i is not retrieved by any of opinion query, $S(d_i) = 1$

d_i is a retrieved document in the pooled collection.

We used the Indri search model in Lemur because Indri provides a lot of operators that can be used to form Boolean queries, which is used to specify the "OR" relationship among adjectives in our opinion indicator lists.

2.2. Results and Discussion

The results of three submitted UB runs are summarized in table1. The title only run (UB2) is only performed as requested. No surprise that it is not as good as other runs using all fields of topic.

Table 1 Summary of Results in the BLOG tack

Runs	Query Type	Opinion Indicators	Opinion (MAP)	Polarity (Raccuracy)
UB1	Title, description, narrative	Wiebe manual lists	0.1501	0.0671
UB2	Title	Wiebe manual lists	0.1013	0.0418
UB3	Title, description, and narrative	Trait adjectives for human personality	0.1501	0.0663

We did not see the difference between two sets of adjective lists as expected. There are two possible reasons. First, our opinion assessments were ran against the pool of retrieved posts across all 50 topics, which makes the ranking not sensitive enough to adjust to each topic, especially the person topic we would like to investigate. We did not do any proximity control of the adjectives relative location to the target concept in the retrieved documents. Ideally we should count only the adjectives located near enough to the target concept. A personality adjective may contribute to the opinion score of a document, but not the opinion about a product mentioned in the documents.

2.3. Future Work

We understood from the very beginning that to identify and compare the discriminating power of these two sets of adjective lists, we needed to do more than using the standard search engine. We are planning to pursue this study further in three aspects. First, analyze subjectivity at the paragraph level, only process the paragraphs containing the target. Meanwhile, using additional natural language processing tool to improve the accuracy of identify target, recognize negation, etc. Second, examine the discriminating power of each individual adjective, selecting only powerful adjectives as indicators. Wiebe's manual list contains 683 negative and 660 adjectives. Even the personality trait lists are much shorter; there are still 234 positive and 252 negative adjectives. A query simply combining of all adjectives in the list equally with "OR" Boolean operator may not sensitive enough to tell difference level of subjectivity. The 2006 and 2007 topics and judgments will service as a good source to use machine

learning techniques to examine the discriminating power of individual adjective or any possible combination. In the future, we would like to develop a systematic way to classify targets, and develop special subjective indicators for each type of target. Third, compare two approaches of using the indicators: the count approach vs. the vector approach. The count approach will have only two indicators in the learning model, positive indicator and negative indicator. Each adjective is a member of either class. The vector approach will deal each individual adjective as a unique indicator. The count method is commonly used in classification tasks. However, Rittman (2007) found in his study that vector approach is more efficient than the count approach in using adjectives/adverbs to classify documents into different genres.

3. Genomics Track

For the Genomics track this year we concentrated on generating a semantic rich topic that can be used for retrieving relevant passages from full text documents. Two approaches were followed to generate those semantic representations. The first is a fully automated approach that used the categories present in the training topics to generate a list of candidate list of semantic types that will be used for expanding the query. The second approach used semantic relatives to generate terms for expansion and to re-rank the output of the retrieval system.

3.1. Document Collection Preparation

We used the preprocessed XML collection that was prepared by the NLM team (Demmer-Fushman, et al. 2007). The parsed documents provided by NLM divided the full text document into passages that correspond to the legal spans within each document (A legal span corresponds to a passage that does not cross the paragraph marks in the original HTML documents). HTML markup was removed to ease indexing for various systems that were used in the NLM runs. We used the "cleaned" version of the documents and used these passages as the unit of indexing. The passages included positional information in the original full text document. A total of 12,641,127 were indexed using Indri. We used the standard unigram model in Indri and the Krovetz stemmer.

3.2 Automatic Query Processing and Expansion

We used the 14 training topics to build a profile of the query that will be used in the test topics. For every topic we identified the type of answer that the topic was expecting. There are 14 possible entity types for a topic (for example: PROTEINS, "SIGNS OR SYMPTOMS", etc.). We also identified synonyms using UMLS and added them to the query.

One of the problems we faced during the training was that our system would rank higher bibliographic reference sections. In general these were very large passages that had many of the key terms from the query but might contain a relevant answer. To avoid this problem we used a simple filter that rejects passages with the term "Medline" in them. Although a very simple heuristic, this seems to be an effective way to identify paragraphs that mainly consist of bibliographic citations. We are aware that this can potentially discard relevant passages but this was probably the easiest way to solve the problem.

The original topics were preprocessed using MetaMap (Aronson, 2001) to identify UMLS terms. These terms were expanded using the list of synonyms of the terms (NLM, 2006). With this information we build a structured query that included the identified terms and the corresponding synonyms. Figure 1 shows a sample of the indri queries that were generated for the 2006 topic using this approach. Note that the system detected two terms that have synonyms "mad cow disease" and "PrnP" The synonyms are included in curly brackets. The operator *#filrej* filters passages that contain the term "Medline" from those retrieved by the *#combine* expression.

```
#filrej( Medline #combine( role PrnP mad cow disease  #1(mad cow disease)  #1(PrnP)
#1(mad cow disease)  #1(dsyn)
{#1(Mad Cow Disease)   #1(Bovine Spongiform Encephalopathy)}
{ #1(PrnP)  #1(prion protein)  #1(prion protein PrP)  #1(similar to Major prion protein
precursor)  #1(PrP27-30)  #1(PrP33-35C)  #1(ASCR)  #1(CD230 antigen)
#1(AA960666)  #1(AI325101)  #1(PrPC)  #1(PrPC)  #1(PrPSc)  #1(Prn-i)  #1(Prn-p)
#1(Sinc)  #1(CJD)  #1(GSS)  #1(MGC26679)  #1(PRIP) #1(PrP rel-2)  #1(PrP-2)
#1(PrP-like)  #1(PrPL-P1-like)  }
) )
```

Sample of a structured Indri query

3.3 Passage Retrieval Based on Semantic Relatives

We explored the use of semantic information for passage retrieval. The main idea behind this approach is to try to identify the semantic type of the expected answer to the query. This was achieved by conducting the following 5 steps:

Step 1: Obtaining semantic representations

We composed a dictionary consisting of names for all UMLS concepts and BioThesaurus concepts. A normalized dictionary lookup was then conducted and each passage was transformed into a list of concepts. We discarded stop words and words with less than three characters during the lookup. We also used BioTagger to mark up gene/protein names.

Step 2: Formulating the topics in structured representations.

For each topic, based on its semantic representations, we transformed it into a structured representation including the following fields:

- AnswerUMLSSem – the UMLS semantic categories for the expecting answers
- AnswerSem – the original answer semantic category
- EventRelationType – the event or relation types usually indicated using verbs.
- KeyConcepts – the key concepts and their corresponding normalized terms in the query topic.
- GPMarkup – the gene/protein mentions

For example, the training topic “What have been used to detect protein TLR4?” is transformed to

(T129, ANTIBODIES, detect, tlr_4:C1321919+C1336636, tlr_4, tlr_4)

Step 3: Obtaining conceptual relatives and matching unigrams and bigrams for query topics.

For each key concept, we obtained its conceptual relatives using the relationship table of the UMLS Metathesaurus. For simplicity, we only used one level relationship. Each passage was then transformed into the following:

- GPMarkup – the frequency of marked gene/protein names.
- MatchedAnswer – The frequency of the matched UMLS semantic categories.
- MatchedConcepts - The frequency of the matched key concepts
- MatchedRelatives- The weighted relative frequency of the matched key concepts where the weight was based on their semantic relation types (RN, SY, CHD, RL: 0.9, RB and PAR: 0.5, RO and SIB:0.2)
- MatchedNormalizedTerms – The frequency of matched normalized Terms
- MatchedUniGrams – The frequency of matched unigrams
- MatchedBigrams –The frequency of matched bigrams

For example, the passage 15784698_26_28296_2326 was transformed into

(24, 24, C1321919+C1336636:24, C1321919+C1336636:10.8, tlr_4:12, protein:5 tlr4:12,)

where the value for MatchedBigrams is null.

Step 4: Ranking passages according to a metric (based on subjective judgment, parameters were tuned based on training topics and their corresponding answers)

The metric was a weighted summation of the above vector and for fields with multiple entries such as MatchedUniGrams in the above example, we split them before summation. The score obtained for article-level as well as passage-level.

Step 5: Tuning the score by taking the overall scores of the article into consideration and pushing the rank of passages that are references and addresses to the bottom of the lists.

We tuned the passage-level score by adding a fractional (a parameter) of the article-level score. For passages that are references and addresses, we assigned a new score based on their original score.

3.4. Genomics Results:

We submitted two official runs (one for each of our approaches). Table x summarizes our results.

Table 2 Performance of our two official runs compared with the median

	UBexp1	Median system	UBHFmanual	Median system
Type of run	Automatic	Automatic	Manual	Manual
Document MAP	0.2209	0.1871	0.1799	0.1689
Passage MAP	0.0575	0.0486	0.0179	0.0322
Aspect MAP	0.1790	0.1078	0.1137	0.1246
Passage2 MAP	0.0698	0.0278	0.0189	0.0202

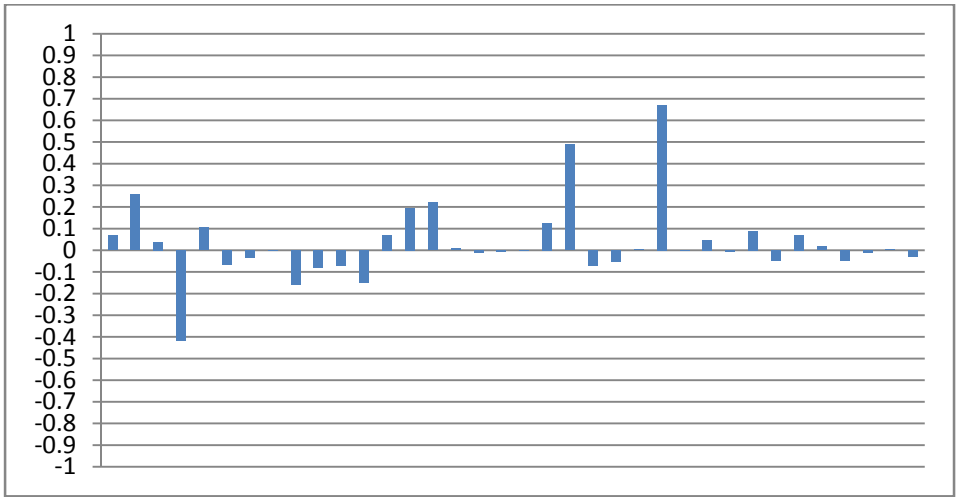

Figure 1 Difference of Document MAP between UBexp1 and the median (automatic runs)

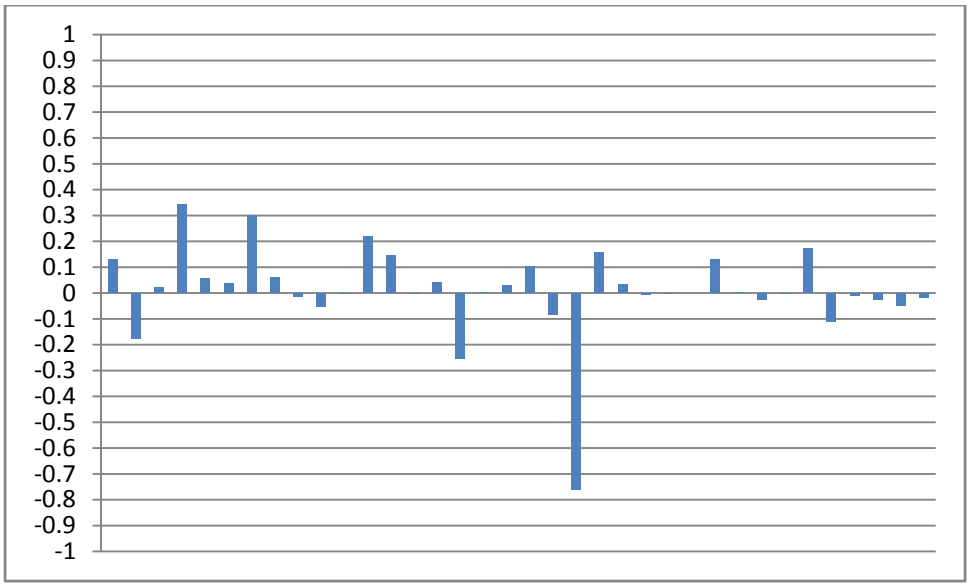

Figure 2 Difference of Document MAP between UBHFmanual and the median (manual runs)

Our results for the automatic expansion using UMLS synonyms performed above the level as the median system of the TREC Genomics. However, the query by query analysis shows that there are 13 topics in which our automatic run outperforms the median run and is outperformed by the median system in 11 topics. This indicates that the difference with the median is not statistically significant.

Our expansion using semantic relatives performed slightly below the median of the manual runs. It outperforms the median in 16 topics and is outperformed by the median in 10 topics. The difference is not statistically significant. Figure 3 shows that when compared to the automatic run topic by topic which shows that the automatic run outperforms the manual run in 17 topics and in 12 topics the manual run outperforms the automatic.

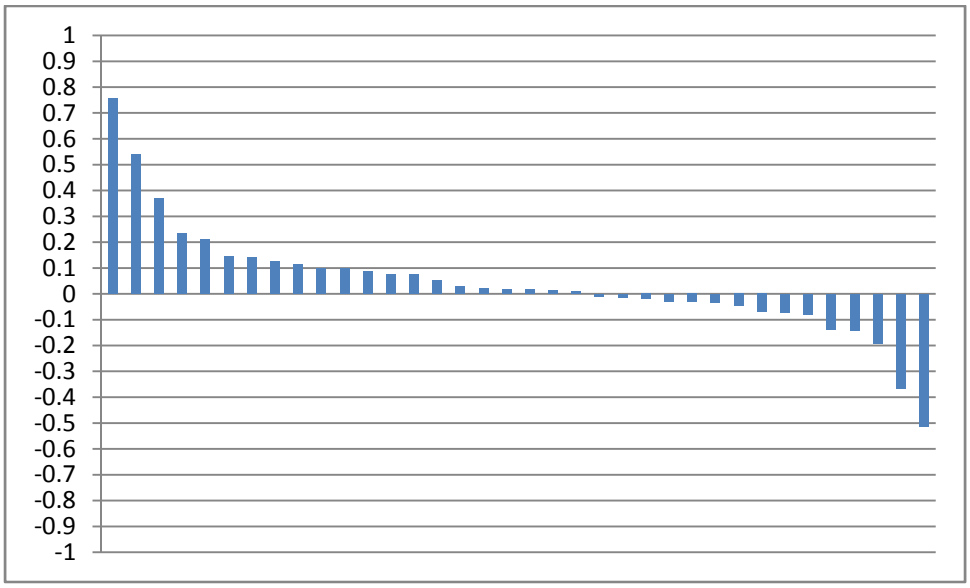

Figure 3 Comparison of Document MAP between UBexp1 and UBHFmanual

3.5. Conclusions and Future work

Or results indicate that expansion using UMLS concepts to build structured queries works well. Our results on using semantic relatives did not show positive improvements. However, we still need to do some more work on analyzing the results to find a better model that can consistently do a better job on re-ranking passages using semantic information.

References:

Aronson A. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. Paper presented at: American Medical Informatics Association Annual Symposium, 2001.

Demner-Fushman, D., Humphrey, S.M. Ide, N.C., Loane, R.F., Mork, J.G., Ruch, P. Ruiz, M.E., Smith, L.H., Wilbur, W.J and Aronson, A.R. (2007) Combining resources to find answers to biomedical questions. In Proceedings of TREC 2007 Conference. NIST.

Peabody, D., & De Raad, B. (2002). The substantive nature of psycholexical personality factors: A comparison across languages. *Journal of Personality and Social Psychology, 83* (4), 983-997.

Rittman, R. (2007). Automatic discrimination of genres: the role of adjectives and adverbs as suggested by linguistics and psychology. Unpublished doctoral dissertation, Rutgers, The State University of New Jersey.

NLM. (2006) U.S. National Library of Medicine: Unified Medical Language System (UMLS).

Wiebe, J., Bruce, R., & O'Hara, T. (1999). Development and use of a gold standard data set for subjectivity classifications. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park: University of Maryland*, 246-253. Retrieved December 4, 2003, from http://www.cs.pitt.edu/~wiebe/pubs/acl99

Wiebe, J. (2000a). Learning subjective adjectives from corpora. *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000), Austin.* Retrieved December 4, 2003, from http://www.cs.pitt.edu/~wiebe/pubs/aaai00

Wiebe, J. (2000b). [Learning subjective adjectives from corpora]. Unpublished list of subjective adjectives. Retrieved November 21, 2003, from http://www.cs.pitt.edu/~wiebe/pubs/aaai00/adjsMPQA