BioKI, a general literature navigation system at TREC Genomics 2006

Sabine Bergler, Jonathan Schuman, Julien Dubuc, Alexandr Lebedev
Department of Computer Science and Software Engineering
Concordia University
1455 de Maisonneuve Blvd West
Montreal, Quebec, H3G 1M8

Abstract

We present the passage reranking component of a general literature navigation system. Based on weighted keyword scoring without automatic enhancements such as term expansion, the system performed slightly above average on all three tasks, with a strong performance on aspect retrieval.

1 Introduction

TREC Genomics 2006 took on the important tasks of document, passage, and aspect retrieval from a large corpus of full-text articles. Mining the full text of scientific articles rather than just abstracts opens up the possibility to access information that goes beyond the topical description in the abstract, be it the more technical rendition of that same topical information that often occurs in the body of the text or information that gives technical background, just as valuable to the researcher.

The user of this information is an expert, if not in the details of the requested information, so in the field at large. Thus search engines have to rely on the expertise of the user to supply appropriate keyphrases for retrieval. The interface to the results supplied traditionally by PubMed Central[1] is becoming difficult to navigate with hundreds of results displayed in linear order with only title, author, and journal information displayed at first glance. This is suitable when the requested information is apparent from title or authorship, but when this background information is not at hand, scanning more than the first 20 results is not feasible. Google Scholar[2] provides the familiar text snippets with search terms highlighted in bold, which provides some context of the occurrence of the search terms but still displays results in linear order. Other search engines have taken the additional step of clustering the results according to inherent terms, such as Vivisimo's BioMetaCluster[3] or according to a fixed organizing hierarchy, such as GoPubMed[4], which cluster's (abstracts only) according to Gene Ontology terms and displays the abstract with not only search terms, but also relevant GO terms highlighted.

But at the heart of the matter is still: which passage in a large corpus answers the query best? When the overall task is not known, enhancing query terms with additional terms (from on-line databases) is dangerous, as it can lead to a drift in meaning away from the intention of the user. PubMed Central (PMC) search uses MeSH[5] terms and MeSH term annotations, thus selecting for topical information. The system we describe here forms the second phase in a two-step process, that begins with collecting a corpus for reranking using PMC search, and thus its MeSH term expansion on a first set of keyphrases.

When no systematic bias towards a task or even subdomain can be assumed, as in our work outside TREC Genomics, a very simple, general strategy seems best. So for our participation at TREC Ge-

[1] http://www.pubmedcentral.nih.gov/

[2] http://scholar.google.com/

[3] http://vivisimo.com/html/biometacluster

[4] http://www.gopubmed.org/

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

nomics we did not employ any domain knowledge sources (such as PubMed's MeSH term expansion, UMLS, or database access), but worked under the assumption that the system should be as general as possible, and that the domain expertise would be supplied by the user. We developed a baseline system that allows the user to specify keyphrases with weights (to identify more important terms). The system scores each paragraph of the given document collection according to how close the keyphrases occur. It was our goal to assess how viable such a very general approach is in the specific subdomain of genomics.

The results are encouraging. TREC Genomics 2006 scored three different tasks: document and passage ranking, and aspect retrieval. TREC Genomics participants submitted a ranked list of up to 1000 text extracts from the document collection, where the text extracts could not exceed a paragraph in length. For the document score, each extract is linked to the document it comes from. The document rank is set equal to the highest ranking extract, lower ranked extracts from the same document are ignored. For the passage score, a character overlap measure was defined, that penalized characters not part of the Gold standard passage. For the aspect score, judges identified MeSH terms with the pooled extracts. The aspect score correlates text extracts with the MeSH terms associated by the judges and attributes these MeSH terms to the system. All scores used the mean average precision measure (MAP).

Passage ranking is closest to the native task for BioKI, except that BioKI would always return the full paragraph, not a minimal span that contains the answer, as expected for TREC. BioKI2 models that behavior and always returns the full paragraph. For BioKI 1 and BioKI3 we submitted the smallest set of contiguous sentences that contained all scoring keywords. While returning full sentences incurs penalties for extra text outside the Gold standard passage, this is in keeping with the spirit of BioKI, where we assume that the user requires a certain amount of context to validate whether the result is actually adequate. We feel that complete sentences provide more context in this sense.

The three runs submitted performed very similarly, with BioKI2 performing the best overall. BioKI showed average performance across queries for the document retrieval task, but above average performance for the passage and aspect retrieval tasks, and BioKI2 was or tied the best performing system three times and came within 0.0004 on a fourth for aspect retrieval. The aspect score favored systems that covered more of the aspect terms the judges had identified for a passage. It thus seems that the very general nature of BioKI held it in the average range for document retrieval, but covered more diverse aspects, making it a useful tool in the hands of experts.

2 BioKI for TREC Genomics 2006

2.1 Data and Preprocessing

The TREC Genomics 2006 data consisted of 162,259 documents from 49 journals that are published electronically by Highwire Press. Workshop organizers provided all "legal spans" for all documents in the collection. Legal spans are defined as any text between HTML paragraph tags for a total of 12,641,127 legal spans in the collection. Legal spans were used to define allowed passages in the pooling and evaluation process. No returned passage was allowed to cross paragraph boundaries, and the legal spans provided a means of verifying this requirement.

BioKI did not apply any coarse-grained filtering of the corpus, as it usually does through its first stage of PMC search. Instead, each query was run over the entire collection. The articles were preprocessed to remove all HTML markup and to separate each article into paragraphs corresponding to the legal spans.

BioKI scores text segments independently. In its general mode of operation, BioKI combines the scores of highly ranked segments within a document to obtain a score for document-level ranking. For TREC, this functionality was left out, and a ranking of all relevant segments in the corpus was returned for each query.

In addition to the legal spans for paragraph output, we also used the sentence boundaries provided by Martijn Schuemie, Erasmus University Medical Center Rotterdam to all TREC Genomics 2006 participants. BioKI scores only paragraphs, but for the passage retrieval we submitted that subpart of the paragraph, that spanned the scoring keywords within their surrounding sentences. We used these semi-official sentence boundaries to streamline our output and avoid penalties from improper byte offsets[6]

.

2.2 Scoring Method

BioKI assigns scores to paragraphs based on the keyphrases and weights supplied by the user. The scoring method used for TREC Genomics 2006 is based on the following principles:

- the closer the keyphrases occur together, the higher the rank
- the more keyphrases are found in the text segment, the higher the rank

The main components of this scoring measure are thus (weighted) keyphrase coverage and keyphrase proximity. The function is similar to one used in (Lawrence and Giles, 1998), which considers the number of keyphrases found, their proximity, and their frequency. In early experiments, we observed that scoring multiple occurrences of the same keyphrase lead to less relevant rankings, and so term frequency is not considered in our scoring function. Also, to allow for unequal weighting of keyphrases, coverage is calculated relative to assigned weights. The scoring function is calculated as:

$$T(1+p)(2c) \tag{1}$$

$$p = \frac{t-w}{t} \tag{2}$$

$$c = \frac{\sum_{\text{matched}} k_i}{\sum_{\text{queried}} k_i} \tag{3}$$

where T is a scoring threshold, p is a proximity factor, and c is a relative keyphrase coverage factor. The proximity and coverage factors are defined in terms of w, the number of characters in the smallest span containing all matched terms; t, the number of characters in the entire tile; and k_i the weight of the i^{th} keyphrase.

The impact of weights on the scoring function is highlighted in an experiment with TREC 2005 data. Over 40 queries, we compared the Mean Average Precision (MAP) score of equal weight settings and of giving the genes a special weight. The equal weight setting only produced the highest MAP for one query, where it was very close to the MAP of the favorite weighted score. For all other queries, a weight on genes improved performance, the optimal value was 25. Since we do not have a domain expert to generate the manual input to our system, we used the 2005 data and scoring as an indication. We submitted three runs, BioKI1, BioKI2, and BioKI3, see Figure 1. All three were manual and interactive, because we took recourse to the 2005 data to emulate our expert. BioKI3 was a system that approximated automatic input by using the nouns from the query at equal weight except for genes, enzymes, and proteins, which received a uniform weight of 25. BioKI1&2 both used the same input, namely manually optimized, varying weights instead of the uniform weight of 25 and possibly additional key terms that had been spotted when analyzing the 2005 results. This was done by the same person, we did not apply any automated techniques to optimize the input, to insure its plausibility as expert input.

Name:	BioKI1
Run type:	interactive
Description:	Weighted keyphrases interactively optimized over 2005 data for each query. Output limited to sentence boundaries
Run type:	interactive
Name:	BioKI2
Description:	Weighted keyphrases interactively optimized over 2005 data for each query. Output limited to paragraph boundaries
Run type:	interactive
Name:	BioKI3
Description:	Weighted keyphrases (weight fixed at 25) interactively optimized over 2005 data for all queries. Output limited to sentence boundaries

Figure 1: Description of BioKI runs

2.3 Example

BioKI's input for question 179, where it has best system's score in passage and aspect retrieval is shown in Figure 2. Here, no weights were assigned for BioKI1&2, all keyphrases were weighted equally. In contrast, BioKI3 assigned the usual weights of 25 to the two most specific terms. This small change makes a difference in the document and passage scores. Throughout the TREC runs, however, the greater difference in behavior between

[6] We incurred a 10% loss of submitted paragraphs or passages due to output formatting problems.

Query <179>	*How do interactions between HNF4 and COUP-TF1 suppress liver function?*
BioKI1&2:	”HNF4” ”COUP-TF+I” ”suppression” ”liver” Document Retrieval score: 0,0833 Passage Retrieval score: 0.0281 Aspect Retrieval score: 0.7143
BioKI3:	”HNF4::25” ”COUP-TF+I::25” ”suppression” ”liver” Document Retrieval score: 0.0851 Passage Retrieval score: 0.018 Aspect Retrieval score: 0.7143

Figure 2: Weighted keywords and their performance

BioKI1 and BioKI3 (both return sentence spans and are thus closest) stems from additional key phrases added to the BioKI1 input.

3 Results

The results confirm certain expectations: Since BioKI1&2 differ only in how much text was returned, they perform near identically on the document retrieval level, where there was no penalty for extraneous text. BioKI1 returns only sentences and outperforms BioKI2, which returns paragraphs, at the passage retrieval level, where systems get penalized for words returned that do not overlap the Gold standard passage. And BioKI2 slightly outperforms BioKI1 at aspect retrieval.

BioKI3, using a fixed weight of 25 for genes, is almost uniformly outperformed by BioKI1&2. The difference is, however, surprisingly small for all three systems.

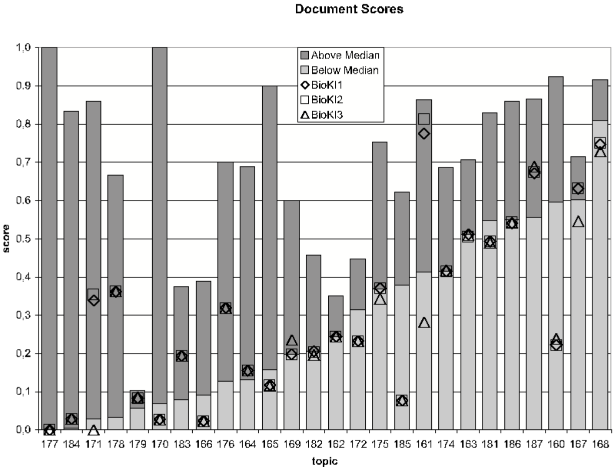
3.1 Document Retrieval Task

Overall performance on the document retrieval task is just barely above average for BioKI. The MAP score of BioKI2 for the document retrieval task is 0.3093. BioKI closely follows the median, BioKI2 is outperformed by the median 11/26 times and exceeds the median by 0.021 on average.

We observe that the distance of BioKI2 to the mean correlates with the number of relevant passages for a topic. In Figure 3 we observe most of the topics with many relevant passages have negative distance to the mean (Δ). Where BioKI2 exceeds the median, the number of relevant passages (# RPass.) is drastically smaller. This demonstrates BioKI's emphasis on low-frequency information (Bergler et al., 2006).

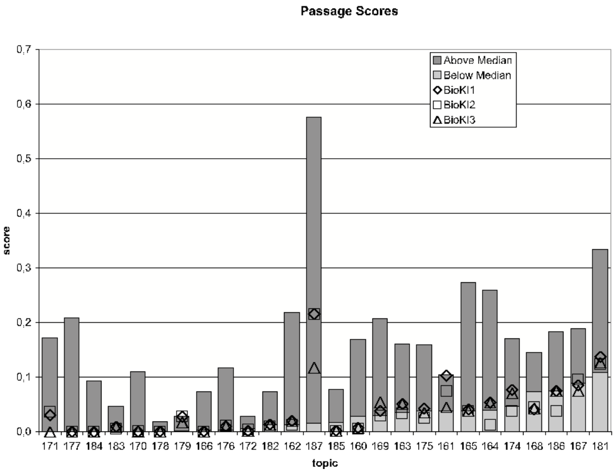
Topic	# RPass.	# Asp.	BioKI2	Median	Δ
160	527	32	0,2219	0,5968	-0,3749
185	25	55	0,0769	0,3791	-0,3022
172	593	78	0,2311	0,3137	-0,0826
166	34	19	0,0237	0,0917	-0,0680
168	243	35	0,7514	0,8094	-0,0580
181	589	96	0,4932	0,5472	-0,0540
170	36	23	0,0260	0,0692	-0,0432
165	17	11	0,1172	0,1582	-0,0410
163	262	35	0,5050	0,5117	-0,0067
186	388	32	0,5444	0,5498	-0,0054
182	144	35	0,2048	0,2053	-0,0005
174	36	12	0,4161	0,4161	0,0000
177	9	12	0,0000	0,0000	0,0000
169	103	32	0,1976	0,1976	0,0000
162	18	20	0,2441	0,2441	0,0000
175	33	27	0,3714	0,3684	0,0030
164	7	14	0,1546	0,1309	0,0237
184	5	10	0,0294	0,0038	0,0256
179	13	7	0,0833	0,0563	0,0270
167	208	35	0,6319	0,6031	0,0288
183	19	11	0,1938	0,0798	0,1140
187	3	13	0,6740	0,5556	0,1184
176	14	9	0,3192	0,1278	0,1914
171	50	13	0,3552	0,0294	0,3258
178	7	21	0,3623	0,0319	0,3304
161	68	94	0,8129	0,4136	0,3993

Figure 3: Document MAP ordered by increasing distance to mean



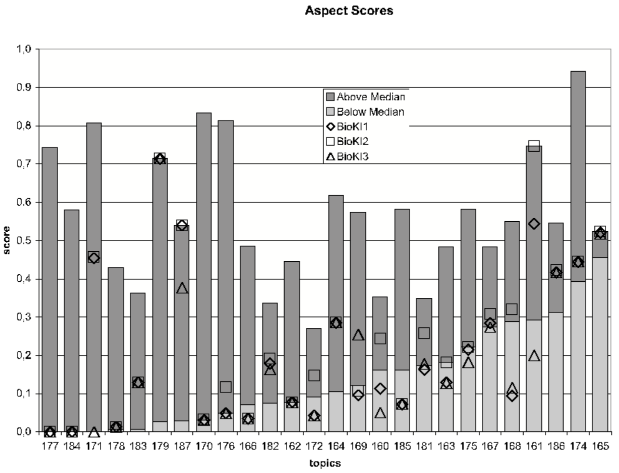
3.2 Passage Retrieval Task

As predicted, BioKI1 outperformed BioKI2&3 at the passage retrieval task. This is because it returns smaller units than paragraphs, namely only the sentences that contain the scoring keywords, and thus accumulates fewer penalties. Its MAP score is 0.0419, the average offset from the median is 0.0132. BioKI2 ranked 44/92 in the passage retrieval task.



3.3 Aspect Retrieval Task

Ranking 17/92, BioKI2 was our best run for aspect, as predicted: BioKI2 returns the entire paragraph in which the scoring keywords were found and thus has more chances to include additional aspects than BioKI1. The difference in score is noticeable and suggests that interesting additional aspects are covered close to, but not necessarily within the span of the scoring keywords. Thus returning slightly larger text segments is an advantage.



BioKI made no special efforts to improve aspect retrieval. We believe that the very generality of our method increased aspect coverage. The MAP score for BioKI2 was 0.2537, an average of 0.1067 above the median.

4 Conclusion

BioKI (Bergler et al., 2006) is designed as an interactive system that capitalizes on the expertise of the user by giving him explicit control and different views of the context of scoring keywords. The reranking component of BioKI participated in the TREC Genomics 2006 shared task as a baseline system. Performance of BioKI was slightly above the median for all tasks. Aspect retrieval showed excellent performance for five topics and was situated in the top 35%. This is thus a viable approach for general information retrieval for user-formulated targeted keyword queries on low-frequency information.

Acknowledgments

We thank Kevin B. Cohen and Greg Caporaso from the Center for Computational Pharmacology, University of Colorado for helpful comments.

References

S. Bergler, J. Schuman, J. Dubuc, and A. Lebedev. 2006. BioKI:Enzymes - an adaptable system to locate low-frequency information in full-text proteomics articles. In *Proceedings of BioNLP'06, a workshop associated with HLT-NAACL 2006*, Brooklyn, NY. Poster abstract.

S. Lawrence and C. Lee Giles. 1998. Inquirus, the NECi meta search engine. In *Seventh International World Wide Web Conference*.