Automatic Highlighting of Bioscience Literature

Hudong Wang[*] and **Shannon Bradshaw**[†] and **Marc Light**[‡]
[*†‡]Computer Science Department
[†]Department of Management Sciences
[‡]Linguistics Department
[‡]School of Library and Information Science
University of Iowa
Iowa, USA 52242
hudwang@cs.uiowa.edu
{shannon-bradshaw, marc-light}@uiowa.edu

Abstract

We presented initial work on the task of automatic highlighting of bioscience literature. We discuss a small set of highlighted documents and queries, both of which were acquired from biology researchers. An automated system is presented along with a performance evaluation on the data set. A novel web-definition-based query expansion method is introduced and it produces an encouraging performance enhancement.

1 Introduction

Bioscience researchers often read articles with specific information needs in mind. Often articles are largely tangential to their needs. Researchers may skim an article and highlight or otherwise mark the relevant passages when they are found. If an application could automatically highlight their articles, it might allow researchers to more efficiently satisfy their information needs. Here we present preliminary results for this automatic highlighting task.

We treat the task as an information retrieval task: the sentence is the passage unit, each sentence is treated as document, the user provides a query, the system ranks the sentences of the article as to their relevance to the query. We use two methods of eliciting user queries: i) the standard query box where a user types in a few keywords and ii) the user highlights a passage to exemplify the information need. We evaluate performance using standard metrics such as mean average precision (MAP).

Since both the query **and** the passages to be retrieved are short, we hypothesize that query expansion will be crucial. To test this hypothesis, we experiment with the use of terms from definitions culled from web content. More specifically, we expand each query term, Q_i, with terms prevelant in definitions of Q_i. The definitions are found using Google's `define:`Q_i query language option.

Our results suggest that the highlighting task is non-trivial, since our baseline search system performs at 0.20 MAP. In addition, example highlighted passages produce better results than keyword queries. Finally query expansion using terms from web definitions improves performance. The main contributions of this research are i) the introduction of an initial highlighting corpus, ii) initial baselines results, and iii) a query expansion method based on web definitions.

2 Related Work

We know of no work on automatic highlighting of bioscience literature. However, the most related work is that on passage retrieval. Working in the area of bioscience O'Connor (1975; 1980) experimented with a number of approaches to passage retrieval. Salton, like O'Connor approached the problem as a means of finding relevant passages in documents that contain information on many different topics (Salton et al., 1993). Others have focused on the problem of identifying boundaries for passages on a specific subject. One example is Hearst's TextTiling approach (Hearst, 1997). The motivation for subject boundary identification is that some documents contain information on a variety of topics and therefore passages may be a more appropriate unit of retrieval. This motivation is different from ours. The community of users we are supporting are those for which many pieces of the same puzzle are scattered throughout the text of a single document.

Another related area is that of question answering (QA) (Voorhees, 1999) where a system searches a large collection of documents for a small set of passages that contain an answer to a specific question such as *who was Johnny Mathis' high school track coach?* In contrast to our highlighting task, QA has a focus on precision since only one or a small number of correct answers is needed.

Additional related work is found in the area of document summarization, particularly that in which sentence classification plays a significant role (Kupiec et al., 1995). The problem we address is similar to that of summarization; however, rather than looking for sentences that play a summarizing role, e.g., *in conclusion*, ..., a highlighting system must look for all relevant data. It should not necessarily reduce or compact the passages that it finds.

Finally, there is also related work on query expansion. In addition to relevance feedback approaches (Rocchio, 1971), many projects have employed a semantic network of terms such as UMLS to identify synonyms and other related terms with which to expand a query (Hersh et al., 2000).

3 Methods and Materials

Corpus: Our highlight corpus consists of 13 journal articles each highlighted by a biology graduate student. One student, whom we will refer to as annotator D, read 5 articles for the purpose of extending her doctoral work towards a post-doctoral position she had just accepted. D's articles have the following Pubmed ids: 11029064, 15568970, 15496557, 11497432, 12857643. The other student, whom we will refer to as annotator T, read 8 articles as part of background research she was doing on plastid protein targeting sequences. T's articles have the following Pubmed ids: 15032850, 11470820, 15078329, 14728677, 12758039, 12045287, 10631267, 12473690. Annotator D highlighted articles electronically using Acrobat Professional 6.0. Annotator T highlighted hard-copy of the articles. The highlighting was done **prior** to our request for highlighted materials. Ascii-encoded versions of the articles were obtained using Acrobat Profession 6.0 and the texts of highlighted regions were manually transcribed from the annotator's original versions.

Queries: The queries corresponding to the highlighting were constructed in retrospect. We asked both annotators to explain why they read the articles and to construct short queries that would correspond to their information need. For example, the query from Annotator T was *14-3-3 chloroplast protein targeting*. (This query was the query for all of annotator T's articles.) We also used the first highlighted region as a query. We also experimented with combining the keyword query and using the first highlighted region as feedback. The queries and text of the highlighted regions are available at `http://que.info-science.uiowa.edu/~light/`.

Document preparation, indexing, and retrieval: The text of articles was tokenized and broken into sentences using LingPipe (Baldwin and Carpenter, 2003). The data was indexed using the Zettair 0.6.1 open source information retrieval system (Zobel et al., 2004). Each article was treated as a separate document collection against which to search. Each sentence was treated as a document. Relevant sentences were those that were at least partially highlighted. Retrieval amounts to ranking the sentences with respect to their relevance to the query. Zettair was used to rank the sentences and we configured it to use the Okapi weighting scheme (Roberson and Walker, 1994) using k1=1.2, k3=1e10, and b=0.75.

Query expansion: Definitions were used to expand the queries. A set of definitions was found for each word of the query (except for stop words). And the words of these definitions were added to the original query. More specifically, a word, *w*, was submitted to the Google search engine as `define:w` and the returned definitions were used as the definitions set. We experimented with *tfidf* thresholds on the definition words. See `http://que.info-science.uiowa.edu:9006/defexp/definition.jsp` for an illustration of the words and weights resulting from such definition sets.

Evaluation metrics: We employ two standard metrics from the information retrieval domain for evaluation of our highlighting systems: MAP and

BEP. Both depend on the metric of precision which is the number of sentences correctly highlighted by the system divided by the number of sentences the system highlighted altogether both correct and incorrect. If there are are N relevant sentences for a query, the BEP, Break Even Precision, is the precision of the system after it has offered its top N highlighted sentences. We report the mean BEP where the BEP values for different queries are averaged. Precision values can also be taken after each relevant sentence is found and their average is the Average Precision. MAP, Mean Average Precision, is the Average Precision averaged over the queries.

4 Results

Distributional analyses: Each row of Table 1 contains information about one of the 13 annotated articles of this study. The articles are identified by their first authors. The left side of the table contains sentence counts and the right side contains query-related word token counts.

Rel.	Tot.	Art.	Ann.	Qlen.	1stHlen.
24	253	Collins	D	9	15
41	957	Lamb.	D	14	45
22	721	Lohne	D	15	18
9	356	Toor	D	10	23
13	483	Wiens	D	6	58
21	187	Ferl	T	4	26
13	358	Hilt	T	4	14
8	568	Mori	T	4	21
18	279	Nakrie	T	4	32
13	258	Roberts	T	4	3
9	749	Sehnke	T	4	2
28	175	SHenry	T	4	21
13	513	Smith	T	4	25

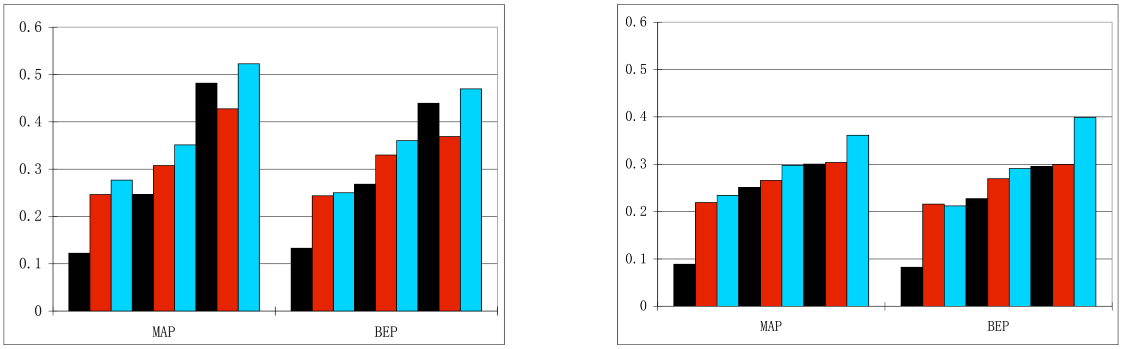
Table 1: Relevant sentence count, Total sentence count, Article first author, Annotator, Query length in word tokens, and First highlighted region length.

An interesting result of our query expansion method is that the queries become very long: The average length after one definition expansion was 255 with an average absolute deviation of 87 from this mean. The average length after using all definitions for expansion was 5446 with an average absolute deviation of 1751 from this mean. The first highlighted regions expanded similarly. We also experimented with short expansions where we only added in words with high *tfidf* weights were *tf* is the number times the word occurs in the web definitions and *idf* is the log base 2 of the inverse of the number of abstracts in Pubmed containing the word. It should also be mentioned that over 80% of the query word types had definitions in Google.

Performance analyses: Figure 1 displays the BEP and MAP values for a number of runs. The full range of the Y axis is 0 to 1. There are two dimensions explored in Figure 1. First, the type of base query used: the keyword query vs. the first highlighted region vs. their combination. The first three bars indicate these runs. The second dimension is the type of expansion used: none, vs. one web definition vs. all web definitions. Each set of three bars corresponds to one of these expansion methods. Thus, moving left to right, there are generally longer base queries, more expansion, and generally better performance.

5 Discussion

The number of queries is small however a number of trends can be noted. **First** the highlighting task is non-trivial. Simple keyword search with a standard weighting scheme performs poorly. The top BEP we obtained was 47% which entails finding roughly 50% the relevant regions along with an equal number of false positives. The **second** trend is that the first highlighted region performs better and the combination of a keyword query and the first highlighted region perform even better. Thus, longer queries tend to increase performance. Note that longer queries could result in more false positives which would lower both MAP and BEP scores and thus this trend is not trivial. The **third** trend is that web definition expansion helps. It is especially effective when the base query is short. It also never hurts. In addition, using all the web definitions tended to be superior to using only the first. The success of the web expansion for the highlighting task is in contrast with its performance on a traditional document retrieval task. We used the same experimental setup with respect to Zettair and the expansion methods but ran the system on the 2004 TREC Genomics Track Task 1 ad hoc retrieval task (Hersh



Annotator D

Annotator T

Figure 1: Performance measures for the following configurations in order: **keyword, 1stHigh, keyword1stHigh, keywordExp1, 1stHighExp1, keyword1stHighExp1, keywordExpAll, 1stHighExpAll, keyword1stHighExpAll**. The substring "keyword" indicates that the keyword query alone was used, "1stHigh" that the first highlighted region was used, and "keyword1stHigh" that both were used. Runs of the same base query are of the same shade. The substring "Exp1" indicates that query expansion using only the first Google definition was used and "ExpAll" that all definitions were used.

and Bhupatiraju, 2004). We tried a variety of web-definition-based expansion schemes but the results always performed lower than our baseline run.

6 Conclusion

We have presented initial work on the task of automatic highlighting of bioscience literature. We have discussed a small set of highlighted documents and their queries both of which were acquired from biology researchers. We have also build and evaluated an automated system and experimented with different kinds of base queries and methods for expanding these queries. Although the data set is small, our expansion method shows promise.

References

Breck Baldwin and Bob Carpenter. 2003. Alias-i lingpipe software. http://www.alias-i.com/lingpipe.

M. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, March.

W. R. Hersh and R. T. Bhupatiraju. 2004. Trec 2004 genomics track overview. In *The Thirteenth Text Retrieval Conference (TREC 2004)*.

William Hersh, Susan Price, and Larry Donohoe. 2000. Assessing thesaurus-based query expansion using the umls metathesaurus. In *Proc. of the AMIA Conference*.

J. Kupiec, J. Pedersen, and F. Chen. 1995. A trainable document summarizer. In *Proc. of the ACM-SIGIR Conference*.

John O'Connor. 1975. Retrieval of answer-sentences and answer-figures from papers by text searching. *Information Processing and Management*, 11:155–164.

John O'Connor. 1980. Answer-passage retrieval by text searching. *JASIS*, pages 227–239, July.

SE Roberson and S Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proc. of the 17th Annual International ACM Special Interest Group in Information Retrieval*, pages 232–241. Springer Verlag.

J. J. Rocchio, 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter Relevance Feedback in Information Retrieval, pages 313–323. Prentice Hall.

Gerard Salton, James Allan, and Chris Buckley. 1993. Approaches to passage retrieval in full text information systems. In *Proc. of the ACM-SIGIR Conference*.

E. M. Voorhees. 1999. The trec-8 question answering track report. In *Proc. of TREC-8*.

Justin Zobel, Hugh Williams, Falk Scholer, John Yiannis, and Steffen Heinz. 2004. The zettair search engine. http://www.seg.rmit.edu.au/zettair.