TREC 2006 Genomics Track Overview

William Hersh[1], Aaron M. Cohen[1], Phoebe Roberts[2], Hari Krishna Rekapalli[1]

[1]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University,
Portland, OR, USA, [hersh, cohenaa, rekapall]@ohsu.edu
[2]Biogen Idec Corp., Boston, MA, USA, Phoebe.Roberts@biogenidec.com

The TREC Genomics Track implemented a new task in 2006 that focused on passage retrieval for question answering using full-text documents from the biomedical literature. A test collection of 162,259 full-text documents and 28 topics expressed as questions was assembled. Systems were required to return passages that contained answers to the questions. Expert judges determined the relevance of passages and grouped them into aspects identified by one or more Medical Subject Headings (MeSH) terms. Document relevance was defined by the presence of one or more relevant aspects. The performance of submitted runs was scored using mean average precision (MAP) at the passage, aspect, and document level. In general, passage MAP was low, while aspect and document MAP were somewhat higher.

1. Introduction

The goal of most information retrieval (IR) systems is to retrieve documents that a user might find relevant to his or her information need. In contrast, the goal of most information extraction (IE) or text mining (TM) systems is to process document text to provide the user with one or more "answers" to a question or information need (Cohen and Hersh, 2005; Roberts, 2006). We propose that what many information seekers, especially users of the biomedical literature, really desire is something in the middle, i.e., a system that attempts to provide short, specific answers to questions and put them in context by providing supporting information and linking to original sources (Hersh, 2005). This motivated us to go beyond the ad hoc retrieval task from previous years of the TREC Genomics Track (Hersh, Cohen et al., 2005; Hersh, Bhupatiraju et al., 2006).

For the TREC 2006 Genomics Track, we developed a new task that focused on retrieval of short passages (from phrase to sentence to paragraph in length) that specifically addressed an information need, along with linkage to the location in the original source document. Topics were expressed as questions and systems were measured on how well they retrieved relevant information at the passage, aspect, and document levels. Systems were required to return passages linked to source documents, while relevance judges not only rated the passages, but also grouped them by aspect. For this task, aspect was defined similar to its definition in the TREC Interactive Track aspectual recall task (Hersh, 2001), representing answers that covered a similar portion of a full answer to the topic question. We also drew upon experience in passage retrieval from the previous TREC High Accuracy Retrieval from Documents (HARD) Track (Allan, 2003; Allan, 2004).

2. Document collection

The documents for this year's task came from a new full-text biomedical corpus. We obtained permission from a number of publishers who use Highwire Press (www.highwire.org) for electronic distribution of their journals. They agreed to allow us to include their full text in HTML format, which preserved formatting, structure, table and figure legends, etc. The document collection was derived from 49 journals and were obtained by a Web crawl of the Highwire site, with post-processing to eliminate as much non-article material as we could. The full collection contained 162,259 documents. The collection was about 12.3 GB when uncompressed. Appendix 1 lists the journals and number of documents from each.

Several notable issues were uncovered when the collection was compiled:

- The collection was not complete from the standpoint of each entire journal. That is, there were some articles that appeared in the journal but did not make it into our collection. This was acceptable to us, since we viewed the collection as a closed and fixed collection.
- Some of the PMIDs were incorrect, emanating from errors in the URLs linking to Pubmed in the source data from Highwire Press.
- Some of the HTML files were empty or nearly empty (i.e., only contained a small amount of meaningless text). Some of this was due to errors in our processing, but most was related to the incorrect or ambiguous links on the Highwire site and in the HTML documents themselves. We kept these files in the collection since they were small and unlikely to have any relevant passages.

We also created a text file, metadata.txt (Windows ASCII format, 11.9 MB), which listed the original URL of the article, the file name in our collection, and the file size in kilobytes. The name of each document file was its PMID plus the extension ".html", which facilitated accessing the associated MEDLINE record.

In addition to the full-text data, the National Library of Medicine (NLM) provided us with both ASCII and XML formatted collections of all the MEDLINE records for the full-text documents in our Highwire collection. We identified 1,767 instances (about 1% of the 162K documents) where the Highwire file PMID was invalid. We investigated the problem and found that for all of instances we checked, the problem was in the original Highwire HTML file having an incorrect PMID in the link to the PubMed record. In other words, the error was inherent in the Highwire data, and not introduced as a result of our processing.

Another file made available to participants was legalspans.txt. This file contained all "legal spans" for all documents in the collection. Legal spans were defined as any contiguous text >0 characters in length not including any HTML paragraph tags, defined as any tag that started with <P or </P (case insensitive). There were a total of 12,641,127 legal spans in the collection. We used these spans to define allowed passages in the pooling and evaluation process, and to limit the size of the passages that needed reviewing by the expert judges

Retrieved passages could contain any span of text that did not include any part of an HTML paragraph tag (i.e., one starting with <P or </P). Because there was some confusion about the different types of passages, we defined the following terms:

- Nominated passage - This was the passage that systems nominated in their runs and were scored in the passage retrieval evaluation. To be legal, these passages had to be a subset of a maximum-length legal span.
- Maximum-length legal span - These were all the passages obtained by delimited the text of each document by the HTML paragraph tags. As noted below, nominated passages could not cross an HTML paragraph boundary. So these spans represented the longest possible passage that could be designated as relevant. As also noted below, we built pools of these spans for the relevance judges. The judges were given the plain text from the entire maximum-length legal span, even if no system nominated the entire span. However, the judges did not need to designate the entire span as relevant, and were able to select just a part of the span as the relevant passage. Each maximum length span was identifier by a triple value of (PMID, offset, length).
- Relevant passage - These were the spans that the judges designated as definitely or possibly relevant. These were portions of the original HTML files, represented by the value triple: PMID, offset, and length. These spans may or may not include HTML markup tags, depending upon whether these tags were inside the relevant answer passages designated by the experts.

The following should also be noted about the maximum-length legal spans:

- The first and last spans were delimited at the beginning and end of the file respectively.
- Other HTML tags (e.g., <B>) could occur within the spans.
- “Empty” (zero character) spans were not included.

3. Topics

The topics for the 2006 track were expressed as questions. They were derived from the set of biologically relevant questions based on the Generic Topic Types (GTTs) developed last year for the 2005 track. These questions each had one or more aspects that were contained in the literature corpus (i.e., one or more answers to each question). A few things should be noted about the topics for 2006:

- Even though the questions were derived from the 2005 track topics, many of them changed, some substantially.
- Groups were instructed that if their systems made use of knowledge about the 2005 topics, then they needed to classify their 2006 runs as interactive, even if they only used automated methods in 2006.
- The official topics were the text of the questions in the text file that was provided. We also provided an Excel spreadsheet, and corresponding PDF, which showed the 2005 topics from which the 2006 topics were derived. However, the information from the 2005 questions was for reference only, and was not to be considered part of the 2006 data.

The questions (and GTTs) all had the general format of containing one or more biological objects and processes and some explicit relationship between them:

Biological object (1..many) ← relationship → Biological process (1..many)

The biological objects might be genes, proteins, gene mutations, etc. The biological process could be physiological processes or diseases. The relationships could be anything, but were typically verbs such as *causes*, *contributes to, affects, associated with*, or *regulates*. We determined that four out of the five GTTs from 2005 could be reformulated into the above structure, with the exception of the first GTT that asked about procedures or methods. The patterns for doing this from the GTTs were based on the examples in Table 1. The topics for the 2006 track are listed in Table 2.

4. Submissions

Submitted runs could contain up to 1000 passages per topic that were predicted to be relevant to answering the topic question. Passages had to be identified by the PMID, the start offset into the text file in characters, and the length of the passage in characters. The first character of each file was defined to be at offset zero.

Passages were required to be contiguous and not longer than one paragraph. As described above, this was operationalized by prohibiting any passage from containing HTML markup tags, i.e., those starting with <P or </P. Any passages containing these tags were ignored in the judgment pooling process but not omitted from the scoring process. (In other words, not counted as potentially relevant for pooling but counted as retrieved for scoring.) Each participating group was allowed to submit up to three official runs, all of which were used for building pools. Each passage was required to be assigned a corresponding rank number and value. The rank number, starting at one and ascending, was used to order nominated passages for rank-based performance computations.

Table 1 - Generic topic types used in the TREC 2006 Genomics Track.

GTT	Question Pattern	Example
Find articles describing the role of a gene involved in a given disease.	What is the role of gene in disease?	What is the role of DRD4 in alcoholism?
Find articles describing the role of a gene in a specific biological process.	What effect does gene have on biological process?	What effect does the insulin receptor gene have on tumorigenesis?
Find articles describing interactions (e.g., promote, suppress, inhibit, etc.) between two or more genes in the function of an organ or in a disease.	How do genes interact in organ function?	How do HMG and HMGB1 interact in hepatitis?
Find articles describing one or more mutations of a given gene and its biological impact.	How does a mutation in gene influence biological process?	How does a mutation in Ret influence thyroid function?

Table 2 - Topics for TREC 2006 Genomics Track.

<160>What is the role of PrnP in mad cow disease?
<161>What is the role of IDE in Alzheimer's disease
<162>What is the role of MMS2 in cancer?
<163>What is the role of APC (adenomatous polyposis coli) in colon cancer?
<164>What is the role of Nurr-77 in Parkinson's disease?
<165>How do Cathepsin D (CTSD) and apolipoprotein E (ApoE) interactions contribute to Alzheimer's disease?
<166>What is the role of Transforming growth factor-beta1 (TGF-beta1) in cerebral amyloid angiopathy (CAA)?
<167>How does nucleoside diphosphate kinase (NM23) contribute to tumor progression?
<168>How does BARD1 regulate BRCA1 activity?
<169>How does APC (adenomatous polyposis coli) protein affect actin assembly
<170>How does COP2 contribute to CFTR export from the endoplasmic reticulum?
<171>How does Nurr-77 delete T cells before they migrate to the spleen or lymph nodes and how does this impact autoimmunity?
<172>How does p53 affect apoptosis?
<173>How do alpha7 nicotinic receptor subunits affect ethanol metabolism?
<174>How does BRCA1 ubiquitinating activity contribute to cancer?
<175>How does L2 interact with L1 to form HPV11 viral capsids?
<176>How does Sec61-mediated CFTR degradation contribute to cystic fibrosis?
<177>How do Bop-Pes interactions affect cell growth?
<178>How do interactions between insulin-like GFs and the insulin receptor affect skin biology?
<179>How do interactions between HNF4 and COUP-TF1 suppress liver function?
<180>How do Ret-GDNF interactions affect liver development?
<181>How do mutations in the Huntingtin gene affect Huntington's disease?
<182>How do mutations in Sonic Hedgehog genes affect developmental disorders?
<183>How do mutations in the NM23 gene affect tracheal development?
<184>How do mutations in the Pes gene affect cell growth?
<185>How do mutations in the hypocretin receptor 2 gene affect narcolepsy?
<186>How do mutations in the Presenilin-1 gene affect Alzheimer's disease?
<187>How do mutations in familial hemiplegic migraine type 1 (FHM1) gene affect calcium ion influx in hippocampal neurons?

Each submitted run was submitted in a separate file, with each line defining one nominated passage using the following format based loosely on trec_eval. Each line in the file had to contain the following data elements, separated by white space:

- Topic ID - from 160 to 187.
- Doc ID - name of the HTML file minus the .html extension. This was the PMID that was designated by Highwire, even though this may not have been the true PMID assigned by the NLM (i.e., used in MEDLINE). But this was the official identifier for the document within the corpus.
- Rank number - rank of the passage for the topic, starting with 1 for the top-ranked passage and preceding down to as high as 1000.
- Rank value - system-assigned score for the rank of the passage, an internal number that should descend in value from passages ranked higher.
- Passage start - the character offset in the Doc ID file where the passage begins, where the first character of the file is offset 0.
- Passage length - the length of the passage in 8-bit ASCII characters.
- Run tag - a tag assigned by the submitting group that should be distinct from all the group's other runs.

Because of the complex nature of this year's task, and most groups' not having a system in place before release of the topics, the classification of runs was complicated. "Usual" TREC rules (detailed at http://trec.nist.gov/act_part/guidelines/trec8_guides.html) would ordinarily categorize runs as follows:

- Automatic - no human modification of topics into queries for a system whatsoever.
- Manual - human modification of queries entered into a system but no modification based on results obtained (i.e., you cannot look at the output from your runs to modify the queries).
- Interactive - human interaction with the system, including modification of the queries after viewing the output (i.e., you look at the output from the topics and corpus and adjust your system to produce different output).

However, because we reused topics (with modification, sometimes substantial) from 2005, and because people were building systems up to the last minute, we adopted the following rules to be applied to classification of runs:

- If a group made any tuning or optimization of their retrieval system based on last year's topics, then their run should have been categorized as interactive this year, even if they did everything else in an automated fashion.
- If a group made any human generated modifications to the topic statements or their system for queries entered into their system, then the run should have been classified as manual.
- If groups made any modifications to the topic statements or their system for the queries entered into it based on looking at the output of passages and/or documents, then their run should have been classified as interactive.

As with many TREC tasks, groups were allowed to manually modify topics to create their queries to their systems. In addition, they were allowed to consult outside resources on the Web (e.g., gene databases), but only in a fully automated fashion. In other words, the original queries could be manually modified, but interaction with external resources could only be done in an automated fashion. For example, if a system pulled information from SOURCE, GenBank, or any other resource, the query to those sources and the information obtained from them had to be done in an automated way, i.e., without manual intervention.

5. Relevance assessments

a. Pooling

There were 92 submitted runs, with each nominating up to 1000 passages over 28 topics. Given our resources, this was far too much data to perform an exhaustive expert evaluation. Instead, we used a pooling method, similar to that used by other document retrieval tasks in TREC.

For each topic a separate pool of passages was created for expert judging. Each ranked and submitted passage consisted of a (PMID, offset, length) triple, which was mapped to its corresponding maximum-length span, also identified as a (PMID, offset, length) triple. These spans were distributed in the legalspans.txt made available before submissions were due. Then, for each topic, pooling was done by taking the top ranked maximal legal span from each submitted run in a round-robin fashion, until the topic pool contained 1000 unique spans. In other words, the top ranked passage was taken from each submitted run, and then the second ranked passage if not yet included, and so on, until the submissions were exhausted or a pool contained 1000 spans.

To consistently subdivide source documents into shorter passages, the HTML <P> tag was used to approximate splitting up the document into paragraphs; as noted above we called these maximum-length legal spans. Likewise, legal submitted passages were limited to not include any HTML <P> tags. By definition, maximal legal spans did not overlap. Therefore, any legally submitted passage would have to be either a maximal legal span or a subset of exactly one maximal legal span.

In addition to HTML <P> tags, additional markup characters were embedded in the text, hampering the readability (thought they generally rendered well in a browser). Maximal legal spans generated in the previous step were converted to plain text by removing the HTML markup. This allowed the judges to concentrate on the content of the passages instead of having to deal with erratic formatting issues. Despite the attempt to remove HTML formatting, plain text was not fully restored to publication quality. Common modifications included loss of inline images that represented characters such as Greek symbols, and lack of conversion of HTML entity codes to more easily readable plain text punctuation characters such as ampersands At times, for some judges, these changes proved to be a distraction.

The plain text content from the pooled spans was then imported into an Excel spreadsheet. Columns were added to allow easy relevance judging. A drop down menu was provided to set the relevance of each passage, and cells were provided for the judges to cut-and-paste relevant plain text from the maximal legal span text field into an “answer text” field. Another column was provided for judges to cut-and-paste MeSH terms corresponding to relevant passage aspects. To make the Excel forms more user friendly for the judges, hyperlinks were added to the PubMed record for the PMID for the journal article for each passage, and also to enable quick access to the PubMed MeSH browser.

b. Judging

Relevance judges were provided with guidelines and a one-hour training course to improve the judging process. As this year’s track was developed by the steering committee, the question and answer nature of the task raised discussion about what constituted a complete answer, prompting development of guidelines for dealing with anaphora and abbreviations to benefit participants and judges alike. In addition, the guidelines offered a brief introduction to MeSH, and tips for taking advantage of Excel features to monitor consistency and completion. Nine judges participated, and they were provided with an email list to discuss issues as they came up.

To assess relevance, judges were instructed to break down the question into required elements (e.g., the biological entities and processes that make up the GTT) and isolate the minimum contiguous substring that answered the question. In general, a passage was definitely relevant if it contained all required elements of the question and it answered the question. A passage was possibly relevant if it contained the majority of required elements, missing elements were within the realm of possibility (i.e. more general terms are mentioned that probably include the missing elements), and it possibly answered the question.

It was possible for a judge to designate any number of relevant passages from an individual article. It was also possible for a judge to designate multiple non-overlapping relevant passages from an individual pooled span. The judges evaluated the text of the maximum-length legal span for relevance, and identified the portion of this text that contained an answer, hereafter called the gold standard passage. This could be all of the text of the maximum legal span, or any contiguous substring. It was possible that one maximum legal span could contain two or more separate gold standard passages. Judges were instructed to duplicate rows with more than one gold standard passage, and process each row independently. Judges were not shown how many systems had retrieved each maximum-length span. Appendix 2 shows the number of maximum-length legal spans where part or all of the span was judged as definitely or possibly relevant; the remainder were counted as not relevant.

Relevance judges next determined the "best" answer passages and grouped them into related concepts. The judges then assigned one or more Medical Subject Headings (MeSH) terms (possibly with subheadings) to capture similarities and differences among retrieved passage aspects. We originally considered using Gene Ontology (GO) terms for this purpose, but an early analysis by our genomics domain expert determined that GO lacked sufficient coverage in many areas needed for this task and MeSH terms alone would provide sufficient coverage.

Judges assigned MeSH term-based aspects to each gold standard passage. They were instructed to use the most specific MeSH term, with the option of adding subheadings, similar to the NLM literature indexing process. If one term was insufficient to denote all aspects of the gold standard passage, judges assigned additional MeSH terms. All passages judged as definitely or possibly relevant were required to have a gold standard passage and at least one MeSH term.

A total of six topics were selected randomly for judgment in duplicate: 160, 165, 176, 179, 181, and 185. (We hoped to have more topics judged in duplicate but were unable to recruit judges for additional work.) Table 3 shows the agreement of the original and duplicate judges, where agreement indicates that any part of a maximum-length legal span was judged as relevant or not. The kappa statistic was calculated to assess chance-corrected inter-rater agreement. For five of the topics, the kappa statistic indicated "good" inter-rater agreement, with a value of 0.60. For topic 181, however, the kappa statistic was poor, with a value of 0.028. This outlier brought the overall kappa value for the six topics down to 0.032. What happened for topic 181 was that one judge interpreted relevance to the question very broadly and the other very narrowly. Table 4 shows the agreement of original and duplicate judges for MeSH terms assigned for aspects, which shows an even poorer rate of agreement. (Kappa could not be calculated due to the inability to calculate the number of MeSH terms not assigned.)

c. Processing

The final result of the judging process was a set of filled-out forms in Excel spreadsheet format. Each spreadsheet corresponded to the judged passages for one topic, one row per passage. If a passage was marked "Not" relevant, no further processing needed to be done, as this passage was not included in the gold standard. Passages marked "Definitely" or "Possibly" relevant were treated as relevant for purposes of the gold standard. The "Definitely" and "Possibly" relevant passages also had two additional associated data items: the relevant answer text cut and pasted from the maximal legal span, and a list of pipe character-separated MeSH terms.

The text and MeSH data associated with the relevant passages was processed to create a set of gold standard passages for each topic. Each gold standard passage consisted of the PMID of the document that the passage was from, the starting character offset, the length of the gold standard passage, and a list of pipe character-separated MeSH terms corresponding to the aspects for that passage.

Table 3 - Agreement of original and duplicate judges for relevant passages, where agreement indicates that any part of a maximum-length legal span was judged as relevant or not.

	Five topics (not including 181)		**Six topics (including 181)**	
	Duplicate judge relevant	Duplicate judge not relevant	Duplicate judge relevant	Duplicate judge not relevant
Original judge relevant	234	228	253	789
Original judge not relevant	53	4485	53	4905

Table 4 - Agreement of original and duplicate judges for MeSH terms assigned. (The cell where neither assigned in undefined.)

	Five topics (not including 181)		**Six topics (including 181)**	
	Duplicate judge assigned	Duplicate judge did not assign	Duplicate judge assigned	Duplicate judge did not assign
Original judge assigned	83	730	90	2407
Original judge did not assign	632	N/A	652	N/A

The starting character offset and length of the gold standard passage in the HTML journal article file was determined by an automated process. Using a dynamic programming algorithm similar to the third stage alignment step in BLAST (McGinnis and Madden, 2004), the relevant answer text selected by the expert judge was aligned with the text of the corresponding maximum length span in the HTML file in order to determine the best overlap. This step had the effect of finding the plain answer text in the HTML file, accounting and skipping over any intervening HTML markup. The starting offset into the HTML file, along with the length in characters in the HTML file matching up with the answer text was taken to be the gold standard passage for that judgment.

As noted above, judges assigned MeSH terms to designate the aspects of a complete topic answer that were addressed by each relevant gold standard passage. This allowed grouping of answer passages and estimatation of the performance of systems in providing a complete answer. Ideally, the MeSH terms provided by the expert judges would have been copied from the MeSH browser without error. However, an additional processing step was necessary because of several types of variation. First, sometimes judges typed in MeSH terms instead of cut and pasting them. Spelling errors were introduced, and these needed to be corrected. A second type of error was created by judges using a MeSH entry term instead of the official MeSH descriptor. These entry terms needed to be mapped to the official term. A few errors were introduced by the judges when non-MeSH terms were used, these needed to be mapped to the closest official MeSH term.

Except for the spelling variations, judges were consistent within a topic, and so the MeSH term variations did not have any effect on the final results. However the MeSH assignments were normalized by mapping to the official MeSH descriptor in order to improve the overall quality and reusability of the test collection. A table driven program was created to fix these errors and map all aspects to official MeSH terms. The table was reviewed by our lead biological expert (P.R.) before finalizing the gold standard aspects. All MeSH terms were also normalized to upper case. Subheadings were preserved if used by the relevance judges.

After mapping the answer text to the HTML source documents and correcting variation in the

MeSH terms, the gold standard passages for each topic were combined into a single file. This file contained 3451 gold standard passages, giving the topic identifier, the source document PMID, the starting offset and length of the relevant passage, and a list of pipe character separated normalized MeSH terms.

Appendix 3 lists the number, average length, and standard deviation of passages per topic as well as the number of aspects per topic. Table 5 shows the minimum, mean, median, and maximum for the topics of these values. It is clear that there is significant variation among the topics as far as number of relevant passages in the literature corpus, the length of those passages, and the number of aspects per topic that were found by the judges. Note that two topics, 173 and 180, had no relevant passages.

6. Performance Measures

For this year's track, there were three levels of retrieval performance that we measured: passage retrieval, aspect retrieval, and document retrieval. Each of these provided insight into the overall performance for a user trying to answer the given topic questions. Each was measured by some variant of mean average precision (MAP). Because this was a new task, and uncharted research territory, we decided to measure the three types of performance separately. We did not propose any summary metric to grade overall performance, but instead wished to examine each aspect of performance in a way that was both as meaningful and straightforward as we could at our current level of experience with this task.

a. Passage-level MAP

This measure used a variation of MAP, computing individual precision scores for passages based on character-level precision, using a variant of a similar approach used for the TREC 2004 HARD Track (Allan, 2004). For each nominated passage, the number of characters that overlapped with those deemed relevant by the judges in the gold standard was determined. For each relevant retrieved passage, precision was computed as the fraction of characters overlapping with the gold standard passages divided by the total number of characters included in all nominated passages from this system for the topic up until that point. Similar to regular MAP, remaining relevant passages that were not retrieved (no overlap with any nominated passages) were added into the calculation as well, with precision set to 0 for these relevant non-retrieved gold standard passages. Then the mean of these average precisions over all topics was calculated to compute the MAP for passages. Note that this measure is essentially the fraction of retrieved characters that are part of an answer to the topic question.

b. Aspect-level MAP

Aspect retrieval was measured using the average precision for the aspects of a topic, averaged across all topics. To compute this, for each submitted run, the ranked passages were transformed to two types of values, either the aspect(s) of the gold standard passage that the submitted passage overlapped with or the value "not relevant". This resulted was a ranked list, for each run and each topic, of lists of aspects per passage, Non-relevant passages had empty lists of aspects. Because we were uncertain of the utility for a user of a repeated aspect (e.g., same aspect occurring again further down the list), we discarded these from the output to be analyzed. For the remaining aspects of a topic, we calculated MAP similar to how it is calculated for documents, with the additional wrinkle that a single passage may have associated with it multiple aspects. Therefore the precision for the retrieval of each aspect was computed as the fraction of relevant passages for the retrieved passages up to the current passage under consideration. These fractions at each point of first aspect retrieval were then averaged together to compute the average aspect precision. Relevant passages that did not contribute any new aspects to the aspects retrieved by higher ranked passages were removed from the ranking. Taking the mean over all topics produced the final aspect-based MAP.

Table 5 - Range and central tendency of relevant passages, their length, and distinct aspects per topic.

Measure	**Relevant passages per topic**	**Mean relevant passage length**	**Distinct aspects per topic**
Minimum	3	27	7
Mean	35	400	22
Median	133	229	30
Maximum	593	6928	96

c. Document-level MAP

For the purposes of this measure, any PMID that had a passage associated with a topic ID in the set of gold standard passages was considered a relevant document for that topic. All other documents were considered not relevant for that topic. System run outputs were collapsed by PMID document identifier, with the documents appearing in the same order as the first time the corresponding PMID appeared in the nominated passages for that topic. For a given system run, average precision was measured at each point of correct (relevant) recall for a topic. The MAP was the mean of the average precisions across topics.

Two topics, 173 and 180, had no relevant passages. These were not included in the scoring for any of the three measures.

7. Results

Information about each run is listed in Appendix 4, including a brief system description provided by the group. The results from all submissions are provided in Appendix 5. A summary of the medium and maximum run results by run type is shown in Figure 1. The best results per topic are seen in Figure 2. In general, document MAP scores are highest, followed by aspect, and then passage, although these scores are not directly comparable since they measure precision at recall of different things. There was a general, though far from perfect, correlation between passage, aspect, and document MAP, as shown in Figure 3. As seen in many TREC-style evaluations and demonstrated in Figure 4, statistical significance, based on pair-wise comparison with the top-ranking score in an ANOVA model, was not achieved for any measure until well down the ranked list of runs.

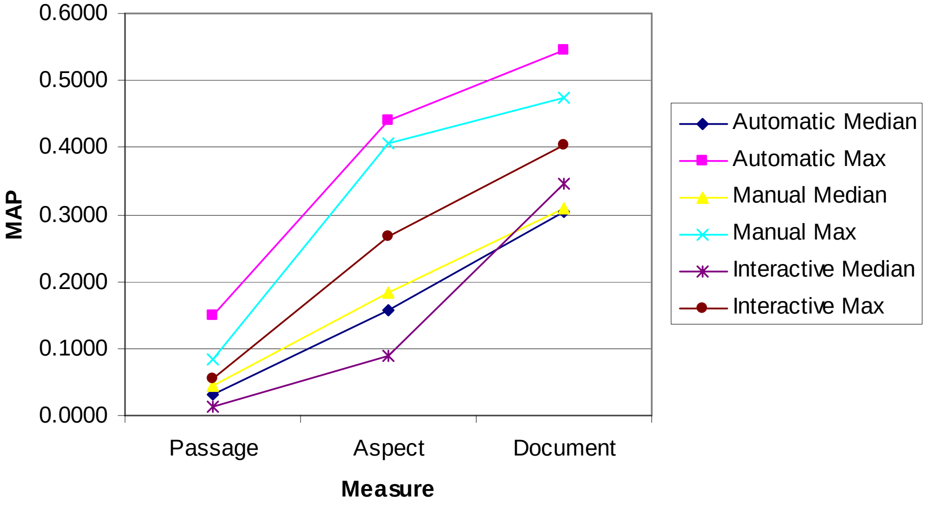

Figure 1 - MAP for all runs and those categorized as automatic, manual, and interactive.

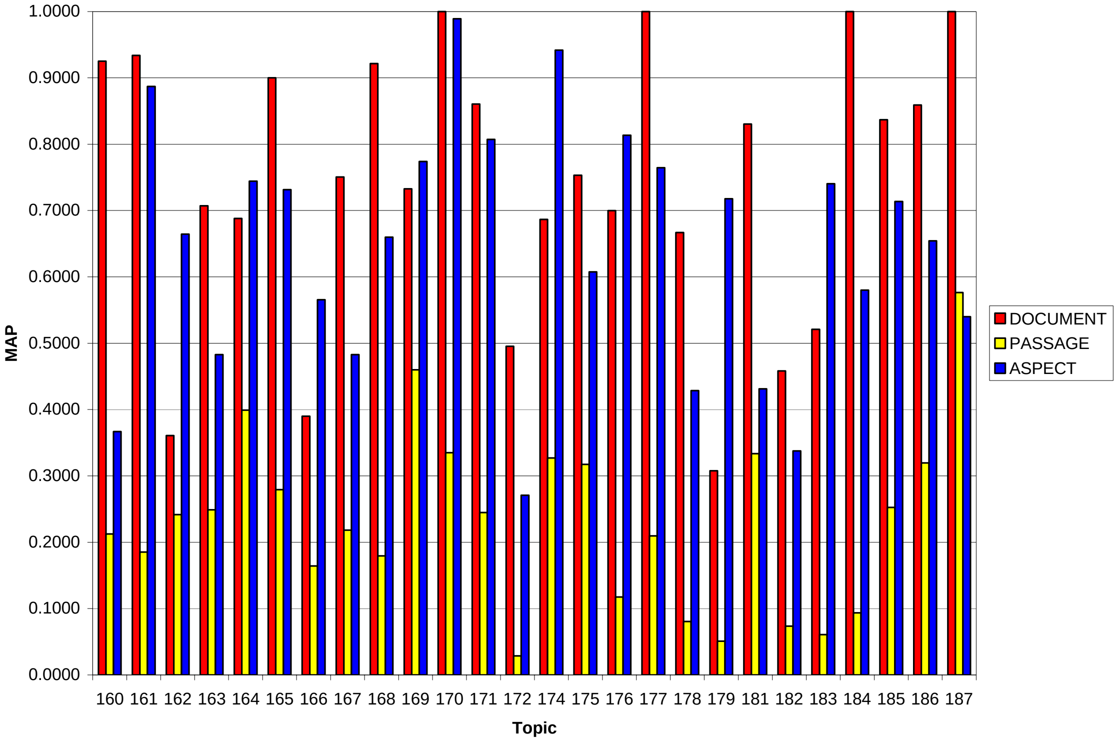

Figure 2 - Best results per topic

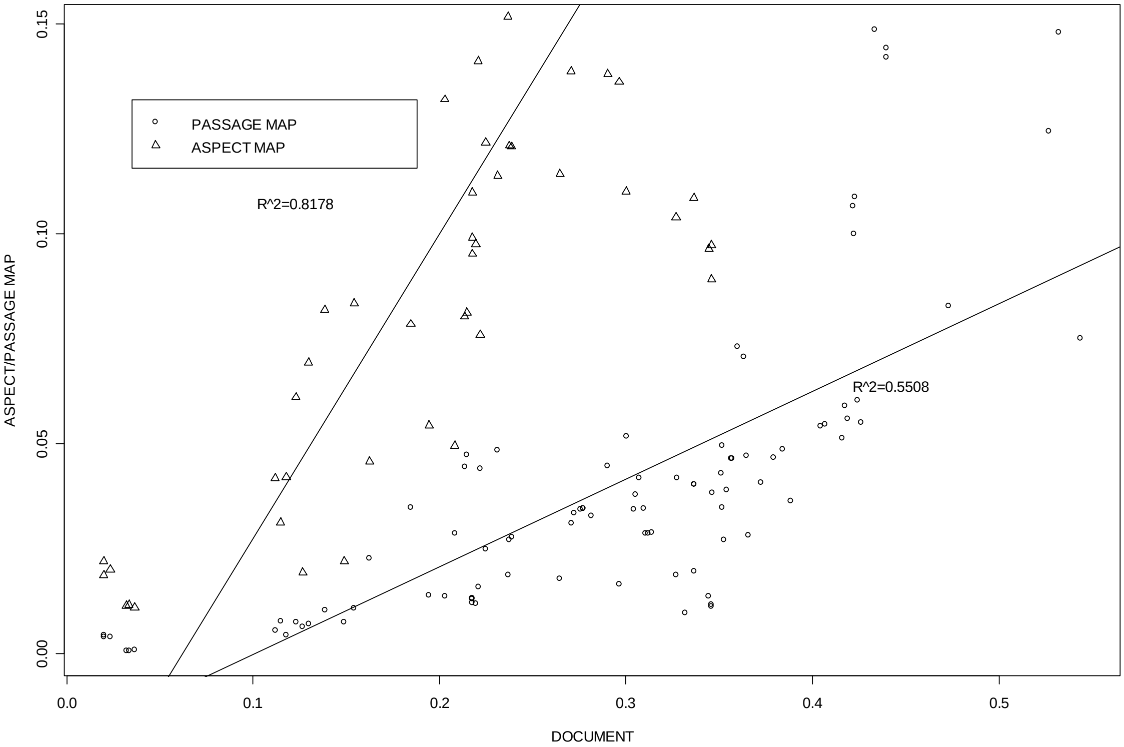

Figure 3 - Plot of passage and aspect MAP versus document MAP for all submitted entries.

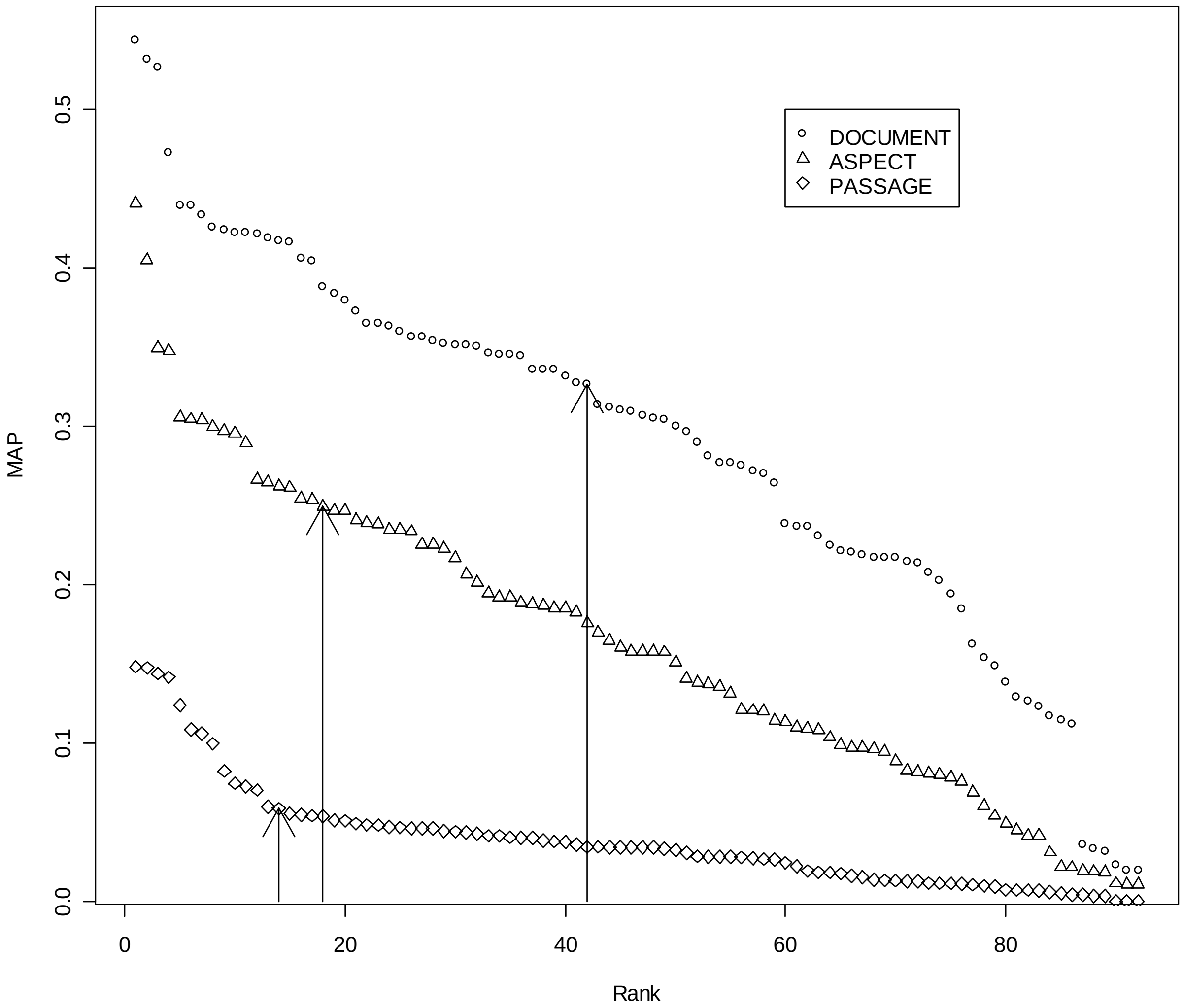

Figure 4 - Ranked MAP showing the first run that is statistically significantly different from the best run.

8. Analysis

Overall there was a wide variation in system performance across submissions for each of the three measures. In general, scores grouped into three sets. A few groups dominated the high scores of each measure, followed by a large group with scores around the mean, and then another large group of relatively low scores. Submissions that scored well on document retrieval tended to score well on both passage and aspect. While a correlation between passage and document retrieval might have been expected, the correlation between document and aspect retrieval is more surprising since aspect retrieval places an emphasis on novelty and document retrieval does not.

Certainly the task and the three measures provided a significant challenge to the participants. The best scores for document retrieval were moderate, and the highest scores on the passage and aspect measures were moderate and fairly low, respectively. No MAP for any system or measure was much greater than 0.50.

For all three measures, the best automatic approaches were as good or better than manual or interactive systems. Manual and interactive approaches did not appear to provide an advantage over automatic methods. However, because the definitions of automatic, manual, and interactive were not as solid as in previous years because systems had the topic questions available during development, inference should be limited from these observations.

Although a comprehensive analysis was not performed, it was clear from the results and techniques of the top-performing groups in passage retrieval that certain approaches were quite effective. In particular, "trimming" passages to shorten them was done in all the runs with the highest passage MAP. Indeed, some groups noted that non-content manipulations of passages had substantial effects on passage MAP, with one group claiming that breaking passages in half with no other changes doubled their (otherwise low) score. To this end, we defined an alternative passage MAP (PASSAGE2) that calculated MAP as if each character in each passage were a ranked document. In essence, the output of passages was concatenated, with each character being from a relevant passage or not. The complete results are shown in Appendix 6, and summarized in Figure 5, where it can be seen that some re-ranking of runs occurred.

9. Conclusions and Future Directions

This novel approach to the TREC 2006 Genomics Track was carried out successfully, leading to the development of a new test collection with new documents and tasks, as well as a new evaluation method and the software to administer and score it. While further analysis of results is required for more definitive conclusions, it can be noted that passage retrieval in this context is quite difficult, with results quite low. Fortunately, retrieval at the aspect and document levels is much better, indicating users still might be able to find answers to their questions in the biomedical literature. Duplicate relevance assessments showed relatively good levels of reproducibility, with one exceptional outlier.

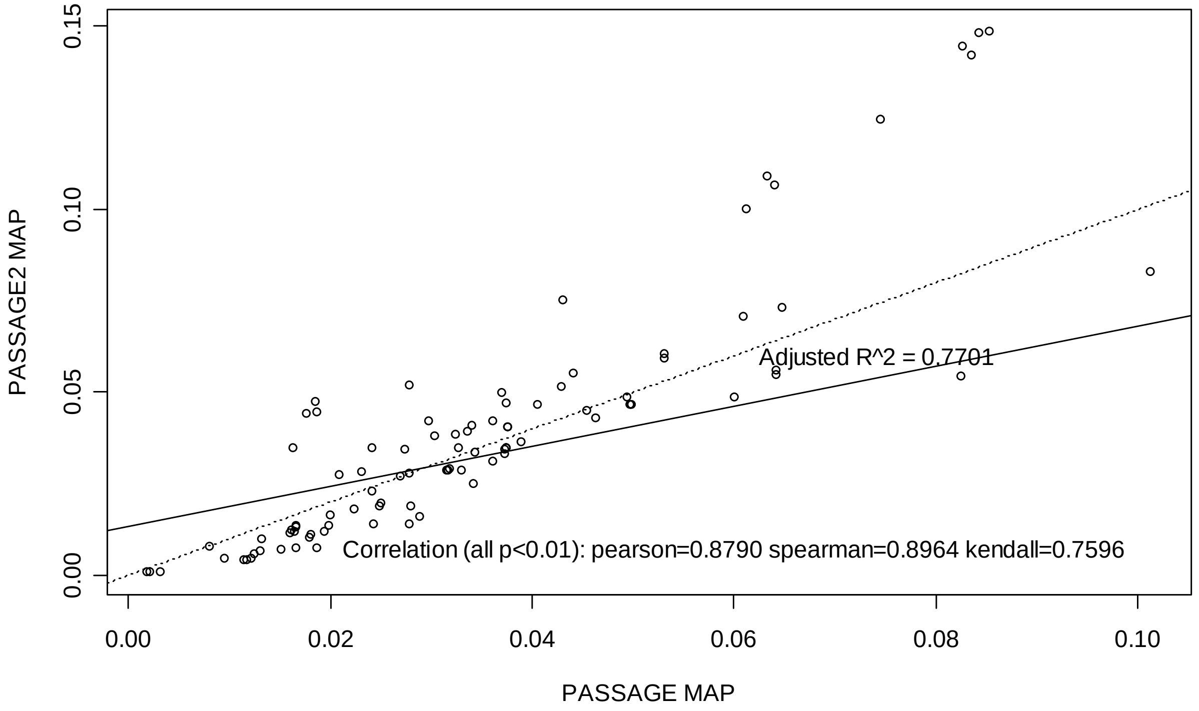

Figure 5 - Comparison of runs using original passage MAP and revised measure (PASSAGE2).

We plan to continue the TREC 2007 Genomics Track in the same direction, using the existing document collection and task structure but adding completely new topics and potentially new topic types. The 2007 track will be the last running of the Genomics Track within TREC, although future options to continue biomedical IR challenge evaluations are being explored.

Acknowledgements

The TREC Genomics Track is funded by grant ITR-0325160 from the U.S. National Science Foundation. The track also thanks Ellen Voorhees, Ian Soboroff, and Lori Buckland of NIST for their help in various ways. We also thank John Sack and Highwire Press (www.highwire.org) for facilitating the use of documents from the respective publishers.

References

Allan, J. (2003). HARD Track overview in TREC 2003 - high accuracy retrieval from documents. *The Twelfth Text REtrieval Conference - TREC 2003*, Gaithersburg, MD. Naitonal Institute of Standards and Technology. 24-37.

Allan, J. (2004). HARD Track overview in TREC 2004 - high accuracy retrieval from documents. *The Thirteenth Text Retrieval Conference (TREC 2004)*, Gaithersburg, MD. National Institute of Standards and Technology.

Cohen, A. and Hersh, W. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics,* 6: 57-71.

Hersh, W. (2001). Interactivity at the Text Retrieval Conference (TREC). *Information Processing and Management,* 37: 365-366.

Hersh, W. (2005). Evaluation of biomedical text mining systems: lessons learned from information retrieval. *Briefings in Bioinformatics,* 6: 344-356.

Hersh, W., Bhupatiraju, R., et al. (2006). Enhancing access to the bibliome: the TREC 2004 Genomics Track. *Journal of Biomedical Discovery and Collaboration,* 1: 3.

Hersh, W., Cohen, A., et al. (2005). TREC 2005 Genomics Track overview. *The Fourteenth Text Retrieval Conference - TREC 2005*, Gaithersburg, MD. National Institute for Standards & Technology.

McGinnis, S. and Madden, T. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research,* 32: W20-W25.

Roberts, P. (2006). Mining literature for systems biology. *Briefings in Bioinformatics,* 7: 399-406.

Appendices

Appendix 1 - List of journals in TREC 2006 Genomics Track full-text collection.

Journal Name	File Size (MB)	Number of Docs	Journal URL
American Journal of Epidemiology	24	1777	aje.oxfordjournals.org
American Journal of Physiology - Cell Physiology	62	2906	ajpcell.physiology.org
American Journal of Physiology - Endocrinology And Metabolism	48	2462	ajpendo.physiology.org
American Journal of Physiology - Gastrointestinal and Liver Physiology	48	2472	ajpgi.physiology.org
American Journal of Physiology - Heart and Circulatory Physiology	99	5170	ajpheart.physiology.org
American Journal of Physiology - Lung Cellular and Molecular Physiology	48	2426	ajplung.physiology.org
American Journal of Physiology - Renal Physiology	39	1897	ajprenal.physiology.org
Alcohol and Alcoholism	9.7	657	alcalc.oxfordjournals.org
Journal of Andrology	7.1	482	www.andrologyjournal.org
Annals of Oncology	16	1273	annonc.oxfordjournals.org
British Journal of Anaesthesia	21	1843	bja.oxfordjournals.org
The British Journal of Psychiatry	17	1531	bjp.rcpsych.org
Blood	209	11291	www.bloodjournal.org
Carcinogenesis	36	2022	carcin.oxfordjournals.org
Cerebral Cortex	22	917	cercor.oxfordjournals.org
Development	62	2402	dev.biologists.org
Diabetes	37	2156	diabetes.diabetesjournals.org
Endocrinology	104	5517	endo.endojournals.org
European Heart Journal	15	1160	eurheartj.oxfordjournals.org
Glycobiology	15	719	glycob.oxfordjournals.org
Human Reproduction	50	3784	humrep.oxfordjournals.org
Human Molecular Genetics	58	3105	hmg.oxfordjournals.org
International Journal of Epidemiology	13	1203	ije.oxfordjournals.org
International Immunology	23	1175	intimm.oxfordjournals.org
Journal of Antimicrobial Chemotherapy	29	2720	jac.oxfordjournals.org
Journal of Applied Physiology	105	5751	jap.physiology.org
Journal of Biological Chemistry	74	4368	www.jbc.org
Journal of Biological Chemistry	33	4733	www.jbc.org
Journal of Biological Chemistry	60	3098	www.jbc.org
Journal of Biological Chemistry	59	2918	www.jbc.org
Journal of Biological Chemistry	49	2432	www.jbc.org
Journal of Biological Chemistry	111	5361	www.jbc.org
Journal of Biological Chemistry	69	3262	www.jbc.org
Journal of Biological Chemistry	119	5539	www.jbc.org
Journal of Biological Chemistry	76	3510	www.jbc.org
Journal of Biological Chemistry	132	6214	www.jbc.org
Journal of Biological Chemistry	109	4886	www.jbc.org
The Journal of Cell Biology	93	3996	www.jcb.org

Journal of Clinical Endocrinology & Metabolism	6.9	758	jcem.endojournals.org
Journal of Cell Science	54	2417	jcs.biologists.org
Journal of Experimental Biology	41	1911	jeb.biologists.org
Journal of Experimental Medicine	70	3492	www.jem.org
The Journal of General Physiology	25	1014	www.jgp.org
Journal of General Virology	40	2375	vir.sgmjournals.org
Journal of Histochemistry and Cytochemistry	24	1592	www.jhc.org
Journal of the National Cancer Institute	34	3214	jncicancerspectrum.oxfordjournals.org
Journal of Neurophysiology	68	2874	jn.physiology.org
Molecular & Cellular Proteomics	9.5	426	www.mcponline.org
Microbiology	46	2400	mic.sgmjournals.org
Molecular Biology and Evolution	25	1303	mbe.oxfordjournals.org
Molecular Endocrinology	36	1610	mend.endojournals.org
Molecular Human Reproduction	14	817	molehr.oxfordjournals.org
Nucleic Acids Research	126	7606	nar.oxfordjournals.org
Nephrology Dialysis Transplantation	38	3629	ndt.oxfordjournals.org
Protein Engineering Design and Selection	15	834	peds.oxfordjournals.org
Physiological Genomics	13	656	physiolgenomics.physiology.org
Rheumatology	21	1985	rheumatology.oxfordjournals.org
RNA	11	544	www.rnajournal.org
Toxicological Sciences	33	1667	toxsci.oxfordjournals.org

Appendix 2 - Relevance judgments per part or all of each maximum-length legal span sent to the judge. If any part of the span was judged relevant, it was counted as definitely or possibly relevant in this table; otherwise it was counted as not relevant.

Topic	**Definitely Relevant**	**Possibly Relevant**	**Not Relevant**	**Total**
160	214	179	607	1000
161	40	28	932	1000
162	1	17	982	1000
163	99	163	738	1000
164	4	3	993	1000
165	7	10	983	1000
166	2	32	966	1000
167	158	50	792	1000
168	56	187	757	1000
169	54	48	898	1000
170	28	8	964	1000
171	14	36	950	1000
172	305	46	648	999
173	0	0	1000	1000
174	18	18	964	1000
175	0	33	967	1000
176	4	10	986	1000
177	6	3	991	1000
178	3	4	993	1000
179	1	12	987	1000
180	0	0	1000	1000
181	418	162	420	1000
182	94	50	856	1000
183	0	19	981	1000
184	3	2	995	1000
185	17	8	975	1000
186	281	107	612	1000
187	1	2	997	1000

Appendix 3 - Number, average length, and standard deviation of relevant passages and number of aspects per topic.

Topic	**Number of Relevant Passages**	**Mean Passage Length**	**Standard Deviation of Passage Length**	**Number of Distinct MeSH Aspects**
160	527	307	234	32
161	68	390	449	94
162	18	350	334	20
163	262	289	171	35
164	7	405	210	14
165	17	251	125	11
166	34	485	553	19
167	208	605	612	35
168	243	251	186	35
169	103	1012	1077	32
170	36	234	168	23
171	50	306	134	13
172	593	171	232	78
173	0	0	0	0
174	36	461	285	12
175	33	416	554	27
176	14	412	281	9
177	9	366	240	12
178	7	410	155	21
179	13	360	283	7
180	0	0	0	0
181	589	775	691	96
182	144	293	239	35
183	19	188	116	11
184	5	318	103	10
185	25	209	183	55
186	388	286	291	32
187	3	1107	954	13
Total	3451			781

Appendix 4 - List of groups, runs, run type, and a brief description (provided by the group) for the TREC 2006 Genomics Track

Group	Run	Type	Brief Description
arizona-stateu.gonzalez	asubaral	automatic	First complete run after question variants in.
arizona-stateu.gonzalez	asubaral2	automatic	Using subject-verb-object as part of ranking together with keyword frequency, distance between keywords.
arizona-stateu.gonzalez	asubaral3	automatic	Similar to first run, but less restrictive in filtering. Only require the subject to be in the passage.
concordiau.bergler	BioKI1	interactive	Weighted keyphrases interactively optimized over 2005 data for each query. Output limited to sentence boundaries.
concordiau.bergler	BioKI2	interactive	Weighted keyphrases interactively optimized over 2005 data for each query. Output limited to paragraph boundaries.
concordiau.bergler	BioKI3	interactive	Weighted keyphrases (weight fixed at 25) interactively optimized over 2005 data for all queries. Output limited to paragraph boundaries.
dalianu.yang	DUTgen1	interactive	Rocchio feedback based on 2005's gold standard, Two levels of indexes, BM25, Paragraph-first reranking
dalianu.yang	DUTgen2	interactive	Rocchio feedback based on 2005's gold standard, Two levels of indexes, BM25, Combining reranking
dalianu.yang	DUTgen3	interactive	Rocchio feedback and SVM based on 2005's gold standard, Two levels of indexes, BM25, Paragraph-first reranking
erasmus.schuemie	EMCUT1	automatic	Document retrieval is performed using a language-modelling approach. Passage selection is based on identification of concepts from the UMLS metathesaurus and a gene thesaurus in both the query and the documents.
erasmus.schuemie	EMCUT2	manual	Document retrieval is performed using a language-modelling approach. Passage selection is based on identification of concepts from the UMLS metathesaurus and a gene thesaurus in both the query and the documents. The concepts identified in the query were manually checked and corrected.
fudanu.niu	fdugen1	manual	passage retrieval, svm classification.
fudanu.niu	fdugen2	manual	passage retireval , svm classification, less positive files
fudanu.niu	fdugen3	manual	sentence retrieval, pattern matching.
iit.urbain	iitx1	automatic	sentMatchRatioNormSC + passMatchRatioNormSC
iit.urbain	iitx2	automatic	sentmatchrationormsc+sentnormsc+passmatchrationormsc+passnormsc)/4
iit.urbain	iitx3	automatic	(1*sentmatchrationormsc+0.1*passmatchrationormsc+0.01*sentnormsc+0.001 *passnormsc)
inst-infocomm-res.yu	i2rg061	automatic	document retrieval
inst-infocomm-res.yu	i2rg062	automatic	document reranking
inst-infocomm-res.yu	i2rg063	automatic	Passage Retrieval
kyotou.wan	kyoto1	automatic	Paragraph-level IR with impact-based retrieval and a probabilistic model for term co-occurrence with their scores merged. Queries expanded automatically with synonyms.
kyotou.wan	kyoto2	automatic	a combination of IR impact-based retrieval at document level with a probabilistic model of term coocurance at paragraph level; for the first phase, queries are automatically expanded using synonyms.
kyotou.wan	kyoto20	automatic	a combination of IR impact-based retrieval at document level with a probabilistic model of term coocurance at paragraph level; for the first phase, original queries are employed.
nlm.aronson	NLMfusion	automatic	This run is the equally-weighted fusion of the results of four automatic methods (1) Essie, a search engine developed specifically for biomedical text supporting flexible query expansion; (2) NCBI, a method that performs selective query expansion based on theme analysis; (3) UniGe, a method based on the EasyIR search engine using term and document weightings as well as pivoted normalization; and (4) Smart, a method based on the Smart search engine. Automatic query expansion based on MetaMap and Theme was available to each of the basic methods. Each method produced paragraphs which were then merged into a final list.
nlm.aronson	NLMinter	interactive	This run consists of manually constructed queries generally consisting of a conjunction of topic terms each of which is a disjunction of synonyms. The synonyms were obtained both by introspection and by consulting databases such as Entrez Gene, GeneCards and MeSH. Query development sometimes also involved examination of PubMed and Essie results of preliminary query

			formulations. The queries were processed by Essie, and the results were automatically trimmed of text unrelated to the topics.
nlm.aronson	NLMmanual	manual	This is similar to the automatic Essie method which is part of our automatic fusion run but with some manually modified queries and with results automatically trimmed of text unrelated to the topics.
ntu.chen	NTUadh1	automatic	The underlying retrieval model is KL-divergence. Synonyms for query expansion are selected by checking that the synonyms co-occur with the original query terms in Pubmed's Medline abstracts.
ntu.chen	NTUadh2	automatic	A baseline run using KL-divergence retrieval model.
ntu.chen	NTUadh3	manual	Same as NTUadh1, except that Nur-77 is manually added to queries containing Nurr-77.
ohsu.hersh	OHSUBigclu	automatic	Same as cluster run. Reranking by clustering of similar returns. Parameters for clustering were modified so that cluster were looser.
ohsu.hersh	OHSUCluster	automatic	Same as noclu. The returned passages were further processed by clustering with CLUTO. Features for clustering are text words from the passage with stopwords filtered out and stemming.
ohsu.hersh	OHSUNoclu	automatic	Automatically generated queries with concept expansion. Documents indexed at legal span granularity with Lucene. Retrieved passages scored by tfidf.
purdueu.si	PCPsgAspect	automatic	Combine multiple types of resources for constructing queries; Hierarchical language model smoothing; Post result filter; Aspect retrieval based on vector representation of MMR
purdueu.si	PCPsgClean	automatic	Combine multiple types of resources for constructing queries; Hierarchical language model smoothing; Post result filter
purdueu.si	PCPsgRescore	automatic	Combine multiple types of resources for constructing queries; Hierarchical language model smoothing; Post result filter; Combine multiple types of evidence
queenslandu.geva	Baseline1M	automatic	Baseline run, Identify paragraphs
queenslandu.geva	Z1KL5KX	automatic	Legal span
queenslandu.geva	Z1KL5KY	automatic	Max 5K span
queenslandu.geva	zoom0p5K1M	automatic	Identify complete paragraphs
queenslandu.geva	zoom1K1M	automatic	zoom on passage (500 chars either size)
suny-buffalo.ruiz	UBexp1	automatic	This run uses a pre-retrieval query expansion method that adds gene names and synonyms. Retrieval is performed using SMART Lnu.ltu and returning full paragraphs.
suny-buffalo.ruiz	UBexp1M	automatic	The run has been generated with SMART using pivoted normalization.
suny-buffalo.ruiz	UBexp2	automatic	This run uses automatic pre-retrieval query that adds gene names and synonyms. Retrieval is performed using SMART with atn ann weighting scheme. Retrieval step returns full paragraphs.
suny-buffalo.ruiz	UBexp2M	automatic	The run has been generated with SMART using pivoted normalization (2nd run from Miguel Ruiz).
technion.gabrilovich	LARAg06pe0	automatic	In the preprocessing phase, documents are indexed with BOW and with an additional set of knowledge-rich features based on Wikipedia concepts. First, a simple BOW query is generated from the topic (no expansion or other enhancements). Then, the top 10 returned documents are mapped into most relevant Wikipedia concepts. The resulting concepts are used to query the second index of documents. No explicit domain-specific knowledge is used. Due to lack of time, retrieval is of entire paragraphs, not passages.
technion.gabrilovich	LARAg06pe5	automatic	Note this run is identical to LARAg06pe0 except the use of query expansion. In the preprocessing phase, documents are indexed with BOW and with an additional set of knowledge-rich features based on Wikipedia concepts. First, a simple BOW query is generated from the topic, with blind feedback query expansion. Then, the top 10 returned documents are mapped into most relevant Wikipedia concepts. The resulting concepts are used to query the second index of documents. No explicit domain-specific knowledge is used. Due to lack of time, retrieval is of entire paragraphs, not passages.
technion.gabrilovich	LARAg06t	automatic	Document and query are represented using features generated by an auxiliary classifier that was built using world knowledge extracted from Wikipedia. No other domain-specific or general information is used. Due to lack of time, retrieval is of entire paragraphs, not passages.
tsinghuau.zhang	THU1	automatic	Our best result.
tsinghuau.zhang	THU2	automatic	Shorter Passages to return.
tsinghuau.zhang	THU3	automatic	Longer Passages to return.

uamsterdam.meij	UAmsBaseLine	automatic	Baseline. Just some naive index-specific acronym expansion on identified (and extracted) NP's
uamsterdam.meij	UAmsExp	automatic	Massive query expansion, using online resources and iteratively gathered acronyms
uamsterdam.meij	UAmsExpSel	automatic	Automatically identified obligatory terms (and expansions)
ucal-berkeley.larso	biotext1	automatic	Basic run. Returns complete legal spans. Ranking based on Lucene score.
ucal-berkeley.larso	biotext3	automatic	Ranked
ucal-berkeley.larso	biotextweb	automatic	Reranking of the first submission run, using n-grams from the Web.
ucolorado.cohen	uchsc1	interactive	Expanded queries are sent to the search engine Lemur. Results undergo zone filtering, and top remaining Lemur results are sent to a singular value decomposition algorithm to expand the results pool by selecting similar paragraphs based on a latent semantic Dirichlet similarity score. Results of the SVD are filtered using Naive Bayes with lexical and conceptual features with training data dervied from manual evaluation of Lemer output.
ucolorado.cohen	uchsc2	interactive	Expanded queries are sent to the search engine Lemur. Results undergo zone filtering. A second, less strict, set of queries is sent to the Lemur search engine and results are filtered using zone filtering and Naive Bayes with lexical and conceptual features with training data dervied from manual evaluation of Lemer output.
ucolorado.cohen	uchsc3	manual	Expanded queries are sent to the search engine Lemur. Results undergo zone filtering.
uguelph.song	UofG0	automatic	Retrieval based on the language modeling approach.
uguelph.song	UofG1	automatic	Retrieval based on the language modeling approach. The results are further filtered based on document coverage.
uguelph.song	UofG2	automatic	Retrieval based on language modeling approach. The results are further filtered based on document and aspect coverage.
uhosp-geneva.ruch	UniGe	automatic	Use the easyIR engine a vector-space with tf.idf weightings and a modified version of pivoted normalization. Basic run.
uhosp-geneva.ruch	UnigeGO	automatic	Use the easyIR engine a vector-space with tf.idf weightings and a modified version of pivoted normalization. GO specific reranking.
uhosp-geneva.ruch	UnigeMesh	automatic	Use the easyIR engine a vector-space with tf.idf weightings and a modified version of pivoted normalization. Template-specific semantic filtering and expansion.
uillinois.chicago.zhou	UICGenRun1	automatic	two-dimensional ranking
uillinois.chicago.zhou	UICGenRun2	automatic	two-dimensional ranking query expansion
uillinois.chicago.zhou	UICGenRun3	automatic	2-dimensional ranking; query expansion; passage retrieval
uiowa.eichmann	UIowa06Geno1	automatic	NLP processing of question, entire paragraph returned as result
uiowa.eichmann	UIowa06Geno2	automatic	NLP processing of question, paragraphs contracted to only those sentences mentioning query terms.
uiowa.eichmann	UIowa06Geno3	automatic	NLP processing of question, entire paragraphs returned, but only those at least 300 characters long (as an ad hoc citation exclusion mechanism).
uiuc.zhai	UIUCauto	automatic	Automatic run.
uiuc.zhai	UIUCinter	interactive	Interactive run.
uiuc.zhai	UIUCinter2	interactive	Interactive run 2.
umass.allan	UMassCIIR1	interactive	Query-biased pseudo relevance feedback. 250 word passages with overlap removed.
umass.allan	UMassCIIR1L	interactive	Query-biased pseudo relevance feedback. The UMassCIIR1 run was "legalized" to only be spans from the legalspans file. Legal spans less than 750 chars were excluded.
umass.allan	UMassCIIR2	interactive	Query-biased pseudo relevance feedback. 500 word passages with overlap removed.
uneuchatel.savoy	UniNE1	automatic	Data fusion of two IR systems (based on normalized RSV values using Z-score) IR system 1 Divergence from randomness, word-based indexing, spelling correction & word variant generation IR system 2 Divergence from randomness, 5-gram indexing
uneuchatel.savoy	UniNE2	automatic	Data fusion of two IR systems (based on normalized RSV values (max)) IR system 1 Divergence from randomness, word-based indexing, spelling correction & word variant generation the document title is included to all passages generated from the article IR system 2 Divergence from randomness, 5-gram indexing
uneuchatel.savoy	UniNE3	automatic	Data fusion of two IR systems (based on normalized RSV values (Z-score), baserun for comparisons) IR system 1 Divergence from randomness, word-

			based indexing IR system 2 Divergence from randomness, 5-gram indexing
utokyo.ishii	Tlab6rGT1	automatic	Automatically calculating abstract level of biomedical concepts and disambiguation of them.
utokyo.ishii	Tlab6rGT2	automatic	Automatically calculating abstract level of biomedical concepts and disambiguation of them. Another condition.
utokyo.ishii	Tlab6rGT3	automatic	Automatically calculating abstract level of biomedical concepts and disambiguation of them. Yet another condition.
uwisconsin.madison	WiscRun1	automatic	Performs POS chunking on topic questions to identify significant noun phrases - Automatically generates expansion term lists for each NP using the MeSH database - Uses Lemur/Indri toolkit to execute queries that require one item in each term list to be found in a paragraph - Ranks results using likelihood of paragraphs given all the expansion term lists concatenated together - Adjusts passage boundaries to include only sentences between the first and last occurrence of key terms
uwisconsin. madison	WiscRun2	automatic	Begins with the same baseline results as our WiscRun1 run - Re-Ranks these results by performing hierarchical clustering on passage bag-of-words vectors - Interleaves results from clusters to promote aspect diversity (Note that clusters are repeatedly considered in order of their average initial rank)
uwisconsin. madison	WiscRun3	automatic	same baseline results as our WiscRun1 run - Re-Ranks using GRASSHOPPER, a graph theoretical algorithm that Performs random walk with absorbing states on the results, to Automatically balance the representativeness and diversity of the final rank
weill-med-cornellu	icb1	interactive	Run 1 was performed with queries at the full article level only. Slider position 200. In this run, we used the MG4J Vigna scorer as baseline. The Vigna scorer favors matches where search terms appear in short text intervals. All runs are performed with the Twease slider at position 200. At this position, the slider expands the query with all the morphological word variants, abbreviations, and MeSH synonyms that match the query words. Morphological word variants are discovered at runtime, with a statistical model trained on Medline 2006 (Campagne, F. unpublished, 2006). Passages are assigned as the minimal intervals where the query match the documents.
weill-med-cornellu	icb2	interactive	Run 2 was performed with parts of the queries at the sentence-level, when appropriate, other terms matching the rest of the article, and ranking by context. Slider at position 200. Context ranking is a new ranking strategy implemented in our textractor framework for the 2006 TREC genomics track. Context queries are expressed as (query)/(context). Briefly, context ranking allows to rank documents matching query by a context, specified as a query expression (e.g., "colon cancer" as a phrase or keywords with boolean clauses). The words in the context do not necessarily occur in the document being ranked. The documents matching the context part of the query are used to infer words that are associated with the context in the corpus. These words are then used to rank the specific set of documents. All runs are performed with the Twease slider at position 200. At this position, the slider expands the query with all the morphological word variants, abbreviations, and MeSH synonyms that match the query words. Morphological word variants are discovered at runtime, with a statistical model trained on Medline 2006 (Campagne, F. unpublished, 2006). Passages are assigned as the minimal intervals where the query match the documents.
weill-med-cornellu	icb3	interactive	Run 3 was performed with queries at the full article level, ranked by context as in Run 2. The context of queries in Run 2 were added to queries from Run 1 to form queries for this run. Slider at position 200. For each topic, queries have the form (query run 1) / (context run 2). All runs are performed with the Twease slider at position 200. At this position, the slider expands the query with all the morphological word variants, abbreviations, and MeSH synonyms that match the query words. Morphological word variants are discovered at runtime, with a statistical model trained on Medline 2006 (Campagne, F. unpublished, 2006). Passages are assigned as the minimal intervals where the query match the documents.
yorku.huang	york06ga1	automatic	1. Use Okapi BM25 for concept-based structured query 2. Use the blind feedback with term selection technique 3. Use a dual index model for passage retrieval 4. No aspect-level retrieval
yorku.huang	york06ga3	automatic	Split the top 500 retrieved passages into 5 groups with 100 passages in each

			group and then use the EM clustering algorithm to re-rank the 100 passages in each group for aspect-level retrieval
yorku.huang	york06ga4	automatic	This run is for document-level retrieval. That is documents will appear in the front of list for only once and those retrieved by different passage previously will be put at the end of list. No aspect-level retrieval.

Appendix 5 - Results of runs sorted by passage, aspect, and document MAP.

Run	Passage	Run	Aspect	Run	Document
	MAP		MAP		MAP
THU2	0.1486	UICGenRun1	0.4411	UICGenRun1	0.5439
UICGenRun3	0.1479	NLMinter	0.4051	UICGenRun3	0.532
THU1	0.1442	UICGenRun3	0.3492	UICGenRun2	0.5269
THU3	0.1419	UICGenRun2	0.3479	NLMinter	0.473
UICGenRun2	0.1244	THU1	0.3058	THU1	0.4395
PCPsgRescore	0.1088	THU3	0.3047	THU3	0.4395
PCPsgAspect	0.1065	THU2	0.304	THU2	0.4335
PCPsgClean	0.0999	PCPsgAspect	0.2997	iitx1	0.4261
NLMinter	0.0827	UIUCinter	0.2976	UIUCinter2	0.4243
UICGenRun1	0.075	PCPsgRescore	0.2958	PCPsgRescore	0.4228
DUTgen2	0.073	UIUCinter2	0.29	PCPsgClean	0.4223
DUTgen1	0.0707	NLMmanual	0.2664	PCPsgAspect	0.4217
UIUCinter2	0.0604	PCPsgClean	0.2652	uchsc2	0.4189
UIUCinter	0.0591	iitx1	0.2624	UIUCinter	0.4176
uchsc2	0.056	NLMfusion	0.2617	iitx3	0.4161
iitx1	0.0549	iitx3	0.2546	uchsc1	0.4066
uchsc1	0.0546	BioKI2	0.2537	uchsc3	0.4042
uchsc3	0.0542	uchsc1	0.2496	iitx2	0.3885
icb1	0.0517	uchsc2	0.2472	UIUCauto	0.3842
iitx3	0.0513	uchsc3	0.2467	NLMfusion	0.3793
UofG0	0.0496	UIUCauto	0.2407	UniNE3	0.3725
UIUCauto	0.0486	biotext1	0.2397	UofG1	0.3655
UAmsExpSel	0.0484	Tlab6r2GT3	0.2386	NLMmanual	0.3648
i2rg061	0.0473	Tlab6r2GT2	0.2351	DUTgen1	0.3634
NLMmanual	0.047	NTUadh2	0.2349	DUTgen2	0.3601
NLMfusion	0.0466	Tlab6rGT1	0.2338	NTUadh3	0.3571
NTUadh1	0.0465	UniNE3	0.2259	NTUadh1	0.3563
NTUadh3	0.0464	NTUadh1	0.2256	UniNE1	0.3539
DUTgen3	0.0447	NTUadh3	0.2232	UofG2	0.3526
i2rg063	0.0445	BioKI1	0.2171	biotext1	0.3517
i2rg062	0.0441	UniNE1	0.207	UofG0	0.3517
NTUadh2	0.0429	UniNE2	0.2018	NTUadh2	0.351
BioKI1	0.0419	OHSUNoclu	0.1946	UniNE2	0.346
OHSUNoclu	0.0419	UBexp2	0.1922	EMCUT1	0.3459
UniNE3	0.0407	UBexp2M	0.1922	EMCUT2	0.3459
UBexp2	0.0403	OHSUBigclu	0.1892	york06ga4	0.3444
UBexp2M	0.0403	OHSUCluster	0.188	york06ga1	0.3365
UniNE1	0.039	iitx2	0.1869	UBexp2	0.3364
UniNE2	0.0384	DUTgen1	0.1857	UBexp2M	0.3364
OHSUBigclu	0.0379	UofG0	0.1856	UMassCIIR2	0.3317
iitx2	0.0363	BioKI3	0.1828	OHSUNoclu	0.3274
biotext1	0.0348	UMassCIIR2	0.1761	york06ga3	0.3269
icb2	0.0348	UniGe	0.1702	Tlab6r2GT2	0.3139
BioKI2	0.0346	DUTgen2	0.1648	Tlab6r2GT3	0.3121
UBexp1	0.0346	UofG1	0.1608	Tlab6rGT1	0.3105
UBexp1M	0.0346	UofG2	0.1583	BioKI2	0.3093
OHSUCluster	0.0344	UBexp1	0.1578	BioKI1	0.3072

UniGe	0.0343	UBexp1M	0.1578	OHSUBigclu	0.3051
BioKI3	0.0335	UnigeMesh	0.1577	OHSUCluster	0.3042
UnigeMesh	0.0328	WiscRun1	0.1516	icb1	0.3003
UnigeGO	0.0309	WiscRun3	0.1411	UMassCIIR1	0.2964
Tlab6r2GT2	0.0288	UnigeGO	0.1386	DUTgen3	0.2902
Tlab6r2GT3	0.0287	DUTgen3	0.1379	UnigeMesh	0.2814
Tlab6rGT1	0.0286	UMassCIIR1	0.1361	UBexp1	0.277
UAmsExp	0.0286	WiscRun2	0.1319	UBexp1M	0.277
UofG1	0.0282	kyoto1	0.1217	UniGe	0.2755
Z1KL5KY	0.0277	Z1KL5KX	0.1209	BioKI3	0.2724
UofG2	0.0271	Z1KL5KY	0.1207	UnigeGO	0.2706
Z1KL5KX	0.027	UMassCIIR1L	0.1143	UMassCIIR1L	0.2647
kyoto1	0.0248	UAmsExpSel	0.1137	Z1KL5KY	0.2386
UAmsBaseLine	0.0226	icb1	0.11	Z1KL5KX	0.2375
york06ga1	0.0197	Baseline1M	0.1097	WiscRun1	0.2368
WiscRun1	0.0188	york06ga1	0.1084	UAmsExpSel	0.2312
york06ga3	0.0187	york06ga3	0.1039	kyoto1	0.2248
UMassCIIR1L	0.0179	zoom1K1M	0.099	i2rg062	0.2219
UMassCIIR1	0.0164	biotextweb	0.0974	WiscRun3	0.2208
WiscRun3	0.0159	EMCUT1	0.0972	biotextweb	0.2195
fdugen3	0.0138	york06ga4	0.0964	Baseline1M	0.2176
WiscRun2	0.0137	zoom0p5K1M	0.0952	zoom0p5K1M	0.2176
york06ga4	0.0135	EMCUT2	0.0891	zoom1K1M	0.2176
zoom1K1M	0.0132	LARAg06pe0	0.0833	i2rg061	0.2148
zoom0p5K1M	0.0131	LARAg06pe5	0.0818	i2rg063	0.2135
Baseline1M	0.0121	i2rg061	0.0812	UAmsExp	0.2081
biotextweb	0.0118	i2rg063	0.0802	WiscRun2	0.203
EMCUT1	0.0117	icb2	0.0784	fdugen3	0.1943
EMCUT2	0.0113	i2rg062	0.0758	icb2	0.1846
LARAg06pe0	0.0109	kyoto2	0.0692	UAmsBaseLine	0.1624
LARAg06pe5	0.0103	kyoto20	0.061	LARAg06pe0	0.1542
UMassCIIR2	0.0097	fdugen3	0.0544	fdugen1	0.1488
icb3	0.0076	UAmsExp	0.0495	LARAg06pe5	0.1385
fdugen1	0.0075	UAmsBaseLine	0.0457	kyoto2	0.1297
kyoto20	0.0075	biotext3	0.0419	fdugen2	0.1267
kyoto2	0.0071	LARAg06t	0.0418	kyoto20	0.1231
fdugen2	0.0065	icb3	0.0313	biotext3	0.1178
LARAg06t	0.0056	fdugen1	0.022	icb3	0.1147
biotext3	0.0044	UIowa06Geno3	0.0219	LARAg06t	0.1119
UIowa06Geno2	0.0044	UIowa06Geno1	0.0199	asubaral3	0.0365
UIowa06Geno1	0.0039	fdugen2	0.0193	asubaral	0.0334
UIowa06Geno3	0.0039	UIowa06Geno2	0.0187	asubaral2	0.0319
asubaral3	0.0008	asubaral	0.0116	UIowa06Geno1	0.0234
asubaral	0.0007	asubaral2	0.0114	UIowa06Geno2	0.02
asubaral2	0.0007	asubaral3	0.011	UIowa06Geno3	0.0198
Mean	0.0392	Mean	0.1643	Mean	0.2887
Median	0.0345	Median	0.1581	Median	0.3083
Min	0.0007	Min	0.011	Min	0.0198
Max	0.1486	Max	0.4411	Max	0.5439

Appendix 6 - Comparison of results and ranks of original (PASSAGE) and modified (PASSAGE2) passage MAP.

Run	PASSAGE MAP	PASSAGE2 MAP	PASSAGE MAP Rank	PASSAGE2 MAP Rank
THU2	0.148593	0.085316	1	2
UICGenRun3	0.147916	0.084342	2	3
THU1	0.144239	0.082738	3	5
THU3	0.141929	0.083562	4	4
UICGenRun2	0.124390	0.074536	5	7
PCPsgRescore	0.108766	0.063310	6	12
PCPsgAspect	0.106500	0.064048	7	11
PCPsgClean	0.099922	0.061270	8	13
NLMinter	0.082714	0.101262	9	1
UICGenRun1	0.075050	0.043047	10	24
DUTgen2	0.073024	0.064767	11	8
DUTgen1	0.070666	0.061039	12	14
UIUCinter2	0.060380	0.053200	13	16
UIUCinter	0.059062	0.053124	14	17
uchsc2	0.055976	0.064229	15	10
iitx1	0.054941	0.044172	16	23
uchsc1	0.054570	0.064268	17	9
uchsc3	0.054223	0.082599	18	6
icb1	0.051705	0.027911	19	52
iitx3	0.051309	0.042971	20	25
UofG0	0.049608	0.037067	21	35
UIUCauto	0.048644	0.049393	22	20
UAmsExpSel	0.048445	0.060108	23	15
i2rg061	0.047251	0.018594	24	70
NLMmanual	0.047048	0.037467	25	30
NLMfusion	0.046584	0.040631	26	26
NTUadh1	0.046493	0.049792	27	19
NTUadh3	0.046379	0.049894	28	18
DUTgen3	0.044680	0.045511	29	22
i2rg063	0.044458	0.018773	30	68
i2rg062	0.044096	0.017759	31	73
NTUadh2	0.042941	0.046341	32	21
BioKI1	0.041915	0.036084	33	37
OHSUNoclu	0.041866	0.029858	34	49
UniNE3	0.040747	0.034017	35	40
UBexp2	0.040306	0.037583	36	28
UBexp2M	0.040306	0.037583	37	29
UniNE1	0.038983	0.033616	38	41
UniNE2	0.038359	0.032431	39	44
OHSUBigclu	0.037946	0.030458	40	48
iitx2	0.036266	0.039009	41	27
icb2	0.034804	0.016423	42	78
biotext1	0.034778	0.024210	43	61
UBexp1	0.034650	0.037421	44	31
UBexp1M	0.034650	0.037421	45	32
BioKI2	0.034603	0.032756	46	43

OHSUCluster	0.034366	0.027431	47	55
UniGe	0.034294	0.037394	48	33
BioKI3	0.033540	0.034368	49	38
UnigeMesh	0.032847	0.037338	50	34
UnigeGO	0.030936	0.036192	51	36
Tlab6r2GT2	0.028798	0.031839	52	45
Tlab6r2GT3	0.028680	0.031514	53	47
Tlab6rGT1	0.028639	0.031713	54	46
UAmsExp	0.028589	0.033008	55	42
UofG1	0.028192	0.023125	56	62
Z1KL5KY	0.027745	0.027867	57	53
UofG2	0.027139	0.020943	58	64
Z1KL5KX	0.026958	0.026984	59	56
kyoto1	0.024776	0.034174	60	39
UAmsBaseLine	0.022634	0.024233	61	60
york06ga1	0.019689	0.025089	62	57
WiscRun1	0.018783	0.027995	63	51
york06ga3	0.018684	0.024995	64	58
UMassCIIR1L	0.017901	0.022477	65	63
UMassCIIR1	0.016448	0.020021	66	65
WiscRun3	0.015890	0.028893	67	50
fdugen3	0.013789	0.027836	68	54
WiscRun2	0.013735	0.024343	69	59
york06ga4	0.013542	0.019954	70	66
zoom1K1M	0.013236	0.016688	71	75
zoom0p5K1M	0.013069	0.016699	72	74
Baseline1M	0.012056	0.016280	73	79
biotextweb	0.011773	0.019502	74	67
EMCUT1	0.011705	0.016463	75	77
EMCUT2	0.011284	0.016136	76	80
LARAg06pe0	0.010871	0.018109	77	71
LARAg06pe5	0.010268	0.017938	78	72
UMassCIIR2	0.009669	0.013213	79	82
icb3	0.007629	0.008082	80	89
kyoto20	0.007493	0.016605	81	76
fdugen1	0.007480	0.018689	82	69
kyoto2	0.007093	0.015123	83	81
fdugen2	0.006471	0.013105	84	83
LARAg06t	0.005562	0.012587	85	84
biotext3	0.004443	0.009515	86	88
UIowa06Geno2	0.004425	0.012204	87	85
UIowa06Geno1	0.003856	0.011418	88	87
UIowa06Geno3	0.003851	0.011783	89	86
asubaral3	0.000759	0.003154	90	90
asubaral2	0.000690	0.001915	91	92
asubaral	0.000684	0.002142	92	91