TREC 2007 Genomics Track Overview

William Hersh[1], Aaron Cohen[1], Lynn Ruslen[1], Phoebe Roberts[2]

[1]Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University, Portland, OR, USA
[2]Pfizer Corp., Cambridge, MA, USA

The TREC 2007 Genomics Track employed an entity-based question-answering task. Runs were required to nominate passages of text from a collection of full-text biomedical journal articles to answer the topic questions. Systems were assessed not only for the relevance of passages retrieved, but also how many aspects (entities) of the topic were covered and how many relevant documents were retrieved. We also classified the features of runs to explore which ones were associated with better performance, although the diversity of approaches and the quality of their reporting prevented definitive conclusions from being drawn.

For the TREC 2007 Genomics Track, we undertook a modification of the question answering extraction task used in the 2006 track [1]. We continued to task systems with extracting out relevant passages of text that answer topic questions. However for this year, instead of categorizing questions by generic topic type (GTT), we derived questions based on biologists' information needs where the answers were, in part, lists of named entities of a given type. Systems were required to return a passage of text, which provided one or more relevant list items within the context of supporting text.

Similar to 2006, systems were tasked to return passages of text. Relevance judges with expertise in biological research assigned the relevant passage "answers," or items belonging to a single named entity class, analogous to the assignment of MeSH aspects in 2006. After pooling the top nominated passages as in past years, judges selected relevant passages and then assigned one or more answer entities to each relevant passage. Passages had to contain one or more named entities of the given type with supporting text that answered the given question in order to be marked relevant. Judges created their own entity list for each topic, based on the passages they judged as relevant. Passages were given credit for each relevant and supported answer. This was required because it was assumed that the passage would not answer the list entity question unless it contains an entity of the type for which the judges were looking. The experts were instructed to perform their relevance judgments in this manner.

The evaluation measures for 2007 were a refinement of the measures used in 2006. We added a new character-based mean average precision (MAP) measure (called Passage2 MAP) to compare the accuracy of the extracted answers, modified from the original measure in 2006 (called Passage MAP). Passage2 MAP treated each individually retrieved character in published order as relevant or not, in a sort of "every character is a mini relevance-judged document" approach. This was done to increase the stability of the Passage MAP measure against arbitrary passage splitting techniques. We included the 2006 passage retrieval measure as well. The Aspect MAP measure remained the same, except that instead of using assigned MeSH aspects we used the answer entities assigned by the relevance judges. We continued to use Document MAP as is, i.e., a document that contained a passage judged relevant was deemed relevant.

Documents

We used the same full-text document corpus that we assembled for the TREC 2006 Genomics Track. The documents in this corpus came from the Highwire Press (www.highwire.org) electronic distribution of journals and were in HTML format. There were about 160,000 documents in the corpus from about 49 genomics-related journals. Highwire Press agreed to allow us to include their full text in HTML format, which preserved formatting, structure, table and figure legends, etc.. In 2006, we found some known issues with the document collection:

- The collection was not complete from the standpoint of each journal. That is, there were many journals where some articles appeared in the journal but did not make it into our collection. (Neither the article nor the MEDLINE record.) This was not an issue to us, since we viewed the corpus as a closed and fixed collection.
- Some of the PMIDs in the source data from Highwire Press were inconsistent with PubMed PMIDs (see next paragraph for an explanation).
- 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 some was also related to the incorrect PMID problem of Highwire. We froze the corpus for the test collection and, since these files were small, they were unlikely to have any relevant passages or even be retrieved by most systems.

Also discovered in 2006 were some errors between the PMIDs designated by Highwire and the actual PMIDs from NLM in MEDLINE. We identified 1,767 instances (about 1% of the 162K documents) where the Highwire file PMID was invalid, in the sense that it returned zero hits when searching for it on PubMed. Some invalid PMIDs are due to the fact that the corresponding documents represented errata and author responses to comments (e.g., author replies to letters). These were assigned PMIDs in publisher-supplied data, but NLM generally does not cite them separately in PubMed, and therefore deleted the PMIDs, although they remained in publisher data. There were documents already assigned a PMID submitted by Highwire that NLM, by policy, decided not to index at all, in which case, again, NLM deleted the PMID, but it was retained in Highwire data. We also found instances of invalid PMIDs in Highwire data for documents that were cited in PubMed but with a different PMID which is absent from Highwire data; such instances could be characterized as errors. In any case, we investigated the problem of invalid PMIDs and found that for all instances we checked, the problem was the original Highwire file having an invalid PMID. In other words, invalid PMIDs were in the Highwire data, not a result of our processing. For this reason, we decided not to delete these files from the collection. They represented, in our view, normal dirty data, whether due to errors or policy differences between NLM and publishers, and should be part of what real-world systems need to be able to handle.

Since the goal of the task was passage retrieval, we developed some additional data sources that aided researchers in managing and evaluating runs. As noted below, 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). We also used these delimiters to extract text that was assessed by the relevance judges. Because there was much confusion in the discussion about the different types of passages, we defined the following terms:

- Nominated passage - This is the passage that systems nominated in their runs and was scored in the passage retrieval evaluation.
- Maximum-length legal span - These were all the passages obtained by the delimited 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 entire span if any system nominated any part of the 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 could select just a part of the span to be relevant.
- Relevant passage - These were the spans that the judges designated as definitely or possibly relevant, had to contain at least one answering entity of the given type, and had entities assign to them by the expert judges. A relevant passage must consist of all or part of a maximum-length legal span.

We note some other things 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.

In order to facilitate our management of the data, and perhaps be of use to participants, we created a 215-megabyte file, legalspans.txt, which included all of the maximum-length legal spans for the collection. The first span for each document included all of the HTML prior to the first <p>, which contained the HTML header information and usually was not part of any relevant passage. This file identified all of the maximum-length legal spans in all of the documents, which consisted of all spans >0 bytes delimited by HTML paragraph tags. These spans were identified by the byte character offset and length in the HTML file. The index number of the first character of the file was 0.

These span definitions can be illustrated with the example in Table 1. The last line of the following data is sample text from an HTML file hypothetically named 12345.html (i.e., having PMID 12345). The numbers above the text represent the tens (top line) and ones (middle) digits for the file position in bytes.

The maximum-length legal spans in this example are from bytes 0-4, 8-29, and 39-50. Our legalspans.txt file would include the following data in PMID, offset, and length order:

```
12345 0  5
12345 8  22
12345 39 12
```

Let us consider the span 8-29 further. This is a maximum-length legal span because there is an HTML paragraph tag on either side of it. If a system nominates a passage that exceeds these boundaries, it will be disqualified for further analysis or judgment. But anything within the maximum-length legal span, e.g. 8-19, 18-19, or 18-28, could be nominated or relevant passages.

Table 1 - Example text for span definitions.

```
000000000011111111112222222222333333333344444444445
012345678901234567890123456789012345678901234567890
Aaa. <p> Bbbbb <b>cc</b> ddd. <p><p><p> Eee ff ggg.
```

We note that it would be possible for there to be more than one relevant passage in a maximum-length legal span. While this will be unlikely, our character-based scoring approach (see below) would handle it fine. However, this was a problem for the judges as the judging interface did not support an easy way to split a judged maximum-length span into multiple relevant passages. In this case judges were instructed to include all of the relevant text within a span in the relevant passage, even if that required the inclusion of some text that the judge thought not relevant. This was most likely to be an issue in spans originating in the references section of the original documents, where two references with informative titles are separated by one or more non-relevant references.

Topics

There were 36 official topics for the track in 2007, which were in the form of questions asking for lists of specific entities. The definitions for these entity types were based on controlled terminologies from different sources, with the source of the terms depending on the entity type. We gathered new information needs from working biologists. This was done by modifying the questionnaire used in 2004 to survey biologists about recent information needs. In addition to asking about information needs, biologists were asked if their desired answer was a list of a certain type of entity, such as genes, proteins, diseases, mutations, etc., and if so, to designate that entity type. Fifty information needs statements were selected after screening them against the corpus to ensure that relevant paragraphs with named entities were present, of which 36 were used as official topics and 14 used as sample topics. Table 2 lists the 36 topics and Table 3 shows the entities and the number of topics in which they occurred.

An example of our topic development process is as follows. Suppose that the information need was:
What is the genetic component of alcoholism?
This is transformed into a list question of the form:
What [GENES] are genetically linked to alcoholism?
Answers to this question are passages that relate one or more entities of type GENE to alcoholism. For example, a valid and relevant answer to the above question would be: *The DRD4 VNTR polymorphism moderates craving after alcohol consumption.* (from PMID 11950104 for those who want to know) And the GENE entity supported by this statement would be DRD4.

Table 2 - TREC 2007 Genomics Track official topics.

<200>What serum [PROTEINS] change expression in association with high disease activity in lupus?
<201>What [MUTATIONS] in the Raf gene are associated with cancer?
<202>What [DRUGS] are associated with lysosomal abnormalities in the nervous system?
<203>What [CELL OR TISSUE TYPES] express receptor binding sites for vasoactive intestinal peptide (VIP) on their cell surface?
<204>What nervous system [CELL OR TISSUE TYPES] synthesize neurosteroids in the brain?
<205>What [SIGNS OR SYMPTOMS] of anxiety disorder are related to coronary artery disease?
<206>What [TOXICITIES] are associated with zoledronic acid?
<207>What [TOXICITIES] are associated with etidronate?
<208>What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to zoledronic acid?
<209>What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to etidronate?
<210>What [MOLECULAR FUNCTIONS] are attributed to glycan modification?
<211>What [ANTIBODIES] have been used to detect protein PSD-95?
<212>What [GENES] are involved in insect segmentation?
<213>What [GENES] are involved in Drosophila neuroblast development?
<214>What [GENES] are involved axon guidance in C.elegans?
<215>What [PROTEINS] are involved in actin polymerization in smooth muscle?
<216>What [GENES] regulate puberty in humans?
<217>What [PROTEINS] in rats perform functions different from those of their human homologs?
<218>What [GENES] are implicated in regulating alcohol preference?
<219>In what [DISEASES] of brain development do centrosomal genes play a role?
<220>What [PROTEINS] are involved in the activation or recognition mechanism for PmrD?
<221>Which [PATHWAYS] are mediated by CD44?
<222>What [MOLECULAR FUNCTIONS] is LITAF involved in?
<223>Which anaerobic bacterial [STRAINS] are resistant to Vancomycin?
<224>What [GENES] are involved in the melanogenesis of human lung cancers?
<225>What [BIOLOGICAL SUBSTANCES] induce clpQ expression?
<226>What [PROTEINS] make up the murine signal recognition particle?
<227>What [GENES] are induced by LPS in diabetic mice?
<228>What [GENES] when altered in the host genome improve solubility of heterologously expressed proteins?
<229>What [SIGNS OR SYMPTOMS] are caused by human parvovirus infection?
<230>What [PATHWAYS] are involved in Ewing's sarcoma?
<231>What [TUMOR TYPES] are found in zebrafish?
<232>What [DRUGS] inhibit HIV type 1 infection?
<233>What viral [GENES] affect membrane fusion during HIV infection?
<234>What [GENES] make up the NFkappaB signaling pathway?
<235>Which [GENES] involved in NFkappaB signaling regulate iNOS?

Table 3 - TREC 2007 Genomics Track entities, definitions, sources of term, and topics with each entity.

Entity Type	Definition	Potential Source of Terms	Topics With Entity Type
ANTIBODIES	Immunoglobulin molecules having a specific amino acid sequence by virtue of which they interact only with the antigen (or a very similar shape) that induced their synthesis in cells of the lymphoid series (especially plasma cells).	MeSH	1
BIOLOGICAL SUBSTANCES	Chemical compounds that are produced by a living organism.	MeSH	3
CELL OR TISSUE TYPES	A distinct morphological or functional form of cell, or the name of a collection of interconnected cells that perform a similar function within an organism.	MeSH	2
DISEASES	A definite pathologic process with a characteristic set of signs and symptoms. It may affect the whole body or any of its parts, and its etiology, pathology, and prognosis may be known or unknown.	MeSH	1
DRUGS	A pharmaceutical preparation intended for human or veterinary use.	MEDLINEplus	2
GENES	Specific sequences of nucleotides along a molecule of DNA (or, in the case of some viruses, RNA) which represent functional units of heredity.	iHoP, Harvester	11
MOLECULAR FUNCTIONS	Elemental activities, such as catalysis or binding, describing the actions of a gene product or bioactive substance at the molecular level.	GO	2
MUTATIONS	Any detectable and heritable change in the genetic material that causes a change in the genotype and which is transmitted to daughter cells and to succeeding generations	MeSH	1
PATHWAYS	A series of biochemical reactions occurring within a cell to modify a chemical substance or transduce an extracellular signal.	BioCarta, KEGG	2
PROTEINS	Linear polypeptides that are synthesized on ribosomes and may be further modified, crosslinked, cleaved, or assembled into complex proteins with several subunits.	MeSH	5
STRAINS	A genetic subtype or variant of a virus or bacterium.	Ad hoc	2
SIGNS OR SYMPTOMS	A sensation or subjective change in health function experienced by a patient, or an objective indication of some medical fact or quality that is detected by a physician during a physical examination of a patient.	MeSH	1
TOXICITIES	A measure of the degree and the manner in which which something is toxic or poisonous to a living organism.	MeSH	2
TUMOR TYPES	An abnormal growth of tissue, originating from a specific tissue of origin or cell type, and having defined characteristic properties, such as a recognized histology.	MeSH	1

Submissions

Submitted runs could contain up to 1000 passages per topic in ranked order 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.

Passages were required to be contiguous and not longer than one paragraph. This was operationalized by prohibiting any passage from containing HTML markup tags, i.e., those starting with <P or </P. Any passage that included those tags was ignored in the relevance judgment process but not omitted from the scoring process. (In other words, they were not including in the pooling and judgment for creating the gold standard, but they could be scored and may include some relevant characters.) Each participating group was be allowed to submit up to three official runs, each of which was used for building the judgement pools. Each passage also needed to be assigned a corresponding rank number and value, which was used to order nominated passages for rank-based performance computations. Rank values could be integers or floating point numbers, such as confidence values.

Each submitted run had to be 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 (spaces or a tab characters):

- Topic ID - from 200 to 235.
- Doc ID - name of the HTML file minus the .html extension. This is the PMID that has been designated by Highwire, even though we now know that this may not be the true PMID assigned by the NLM (i.e., used in MEDLINE). But this is the official identifier for the document.
- 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 byte 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 bytes, in 8-bit ASCII, not Unicode.
- Run tag - a tag assigned by the submitting group that should be distinct from all the group's other runs (and ideally any other group's runs, so it should probably have the group name, e.g., OHSUbaseline).

Here is an example of the submission file format:

```
200 12474524 1 1.0   1572 27  tag1
200 12513833 2 0.373 1698 54  tag1
200 12517948 3 0.222 99   159 tag1
201 12531694 1 0.907 232  38  tag1
201 12545156 2 0.456 789  201 tag1
```

A Perl script that checked runs to insure that the submission file was in the proper format was available (check_genomics.pl). Runs also needed to include a “dummy” passage for any topic for which no passages were retrieved. It was recommended that the dummy passage use “0” as a docid, “0” as the passage start, and “1” as the passage length. This worked for the Perl script and did not correspond to a document in the collection.

Runs were also classified based on amount of intervention in converting topics to queries. We adopted the “usual” TREC rules (detailed at http://trec.nist.gov/act_part/guidelines/trec8_guides.html) for categorizing runs:

- Automatic - no human modification of topics into queries for your system whatsoever
- Manual - human modification of queries entered into your system (or any other 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 or the system after viewing the output from your system or any other system (i.e., you look at the output from the topics and corpus and adjust your system to produce different output)

Relevance Judgments

The expert judging for this evaluation used the pooling method, with passages corresponding to the same topic ID pooled together. The judges were presented with the text of the maximum-length legal span containing each pooled passage, with pool composed of the top ranked 1000 passages for each topic. They then evaluated the text of the maximum-length legal span for relevance, and identified the portion of this text that contains an answer. This could be all of the text of the maximum legal span, or any contiguous substring. If a maximum legal span contained more than one relevant passage, judges were instructed to select the minimum contiguous passage that contained all relevant passages, even if the passages were separated by irrelevant text. Maximum legal spans comprised of the journal article bibliography frequently generated multiple relevant sub-passages that needed to all be included in the singe designated passage.

Judges were recruited from the institutions of track participants and other academic or research centers. They were required to have significant domain knowledge, typically in the form of a PhD in a life science. They were trained using a 12-page manual and a one-hour videoconference, with the option of testing out of the videoconference by successfully judging a mini-topic based on a practice topic from 2006 made up of an equal mix of definitely, possibly, and not relevant maximum-length legal spans. The self-training option had the unexpected benefit of highlighting and correcting potential problems with the judging tool or ambiguous guidelines before judging began in earnest. The training manual is on the track Web site at: http://ir.ohsu.edu/genomics/2007judgeguidelines.pdf

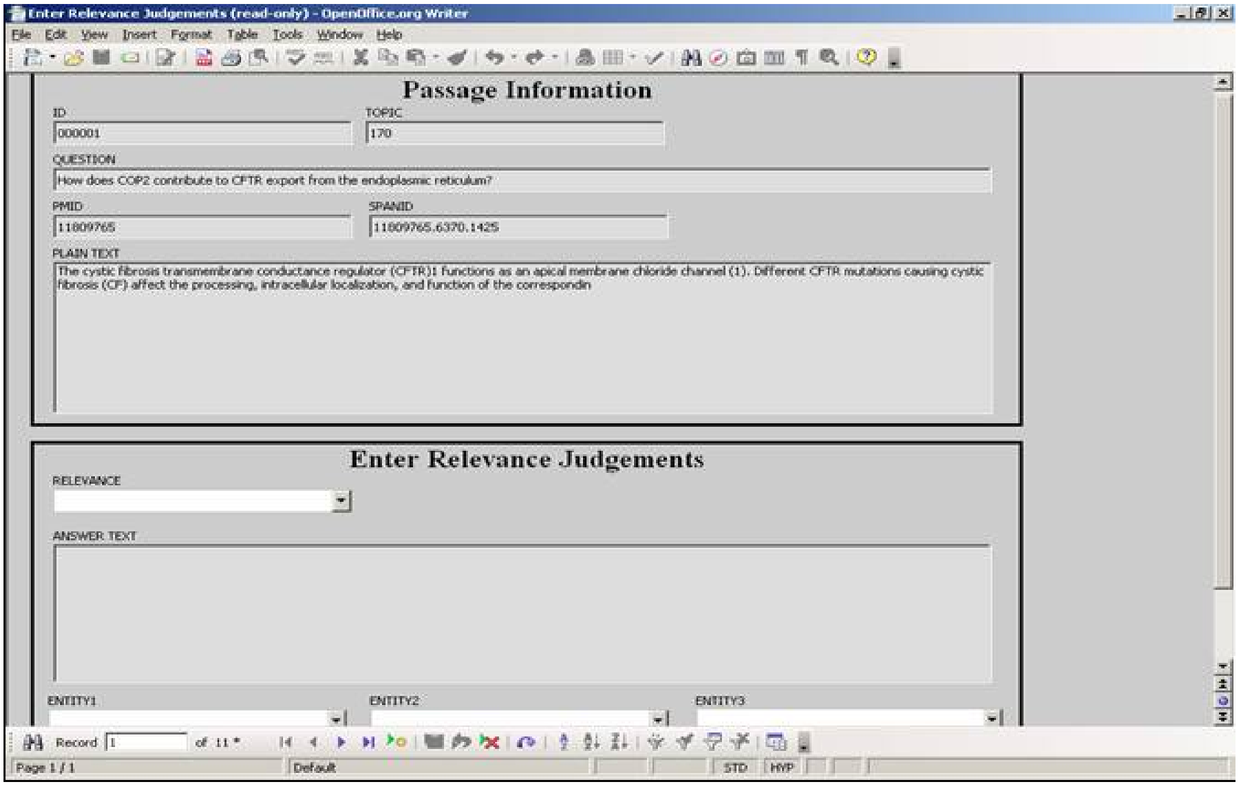
In summary, judges were given the following instructions:

1. Review the topic question and identify key concepts.
2. Identify relevant paragraphs and select minimum complete and correct excerpts.
3. Develop controlled vocabulary for entities based on the relevant passages and code entities for each relevant passage based on this vocabulary.

Judgments were made using database files created and accessed via the OpenOffice Base application. As shown in Figure 1, judges were presented passages as a form view of individual records in the database with the topic, question, and text of the full-text legal passage. If part or all of the passage was relevant, the judges then:

1. Selected the level of relevance ("Definitely Relevant" or "Possibly Relevant").
2. Copied the relevant portion of the passage from the passage plain text field into the answer text box.
3. Selected entities (ENTITY1, ENTITY2, etc.) they had added using the Add Entities form (not shown).

A gold standard was created by extracting out the relevance passages and entities from the database file for each topic. Selected relevant text was transformed into file character offset and length using a text alignment algorithm. A summary of the gold standard developed from the results of the judging process is shown in Table 4. Topics ranged from a low of 1 relevant passage to a high of 377. Individual topics had a range of 1 to 300 relevant entities, with an average ranging between 1.0 to 3.5 entities assigned per relevant passage.


Enter Relevance Judgements (read-only) - OpenOffice.org Writer
File Edit View Insert Format Table Tools Window Help
Record 1 of 11 *
Page 1 / 1
Default
STD
HYP

Passage Information

ID: 000001

TOPIC: 170

QUESTION: How does COP2 contribute to CFTR export from the endoplasmic reticulum?

PMID: 11809765

SPANID: 11809765.6370.1425

PLAIN TEXT: The cystic fibrosis transmembrane conductance regulator (CFTR)1 functions as an apical membrane chloride channel (1). Different CFTR mutations causing cystic fibrosis (CF) affect the processing, intracellular localization, and function of the correspondin

Enter Relevance Judgements

RELEVANCE

ANSWER TEXT

ENTITY1

ENTITY2

ENTITY3

Figure 1 - Passage judgment form.

Table 4 - Relevant passages, relevant documents, mean and standard deviation (SD) of relevant passage length, number of aspects, and mean number of aspects per relevant passage.

Topic	**Relevant Passages**	**Relevant Documents**	**Mean Relevant Passage Length**	**SD of Relevant Passage Length**	**Aspects**	**Mean Aspects Per Relevant Passage**
200	320	193	2380.58	5387.02	300	2.15
201	37	12	1701.86	2894.64	7	1.16
202	53	43	522.77	293.60	28	1.45
203	321	147	2163.60	4237.72	245	1.91
204	164	74	1989.90	4670.61	36	1.79
205	93	65	788.67	1277.35	17	1.23
206	38	19	363.79	362.85	24	1.87
207	15	12	357.60	671.28	8	1.07
208	22	16	615.36	317.50	13	1.23
209	78	11	1239.63	720.81	15	1.50
210	71	57	669.79	623.70	21	1.10
211	57	42	191.68	217.10	29	1.14
212	358	133	1165.97	969.94	142	2.16
213	377	185	456.94	594.39	165	1.88
214	209	98	414.91	1095.21	54	1.42
215	137	73	750.96	580.54	80	1.66
216	42	34	1058.12	3141.51	13	1.12
217	38	34	1491.18	1019.48	34	1.03
218	163	74	632.23	635.55	80	1.28
219	22	16	623.64	503.66	43	3.41
220	16	6	425.75	218.10	6	1.75
221	183	87	1373.32	1705.58	108	1.44
222	57	42	1249.51	914.23	72	2.18
223	18	8	269.72	138.24	12	1.17
224	3	3	1009.33	666.59	1	1.00
225	1	1	745.00	0.00	1	1.00
226	152	57	753.82	1648.91	18	2.25
227	281	172	1307.02	863.14	183	2.25
228	15	14	632.20	413.79	13	1.87
229	150	57	528.81	978.41	34	1.79
230	82	29	1186.65	933.99	25	1.30
231	16	13	472.00	406.56	7	1.06
232	93	57	388.57	907.63	49	1.12
233	19	16	1186.68	1070.54	1	1.00
234	609	483	1777.02	3124.85	577	3.24
235	182	107	1963.25	1737.40	141	2.54
Mean	124.8	69.2	968.0	1276.2	72.3	1.63

Evaluation Measures

For this year’s track, there were three levels of retrieval performance measured: passage retrieval, aspect retrieval, and document retrieval. Each of these provides insight into the overall performance for a user trying to answer the given topic questions. Each was measured by some variant of MAP. We again measured the three types of performance separately. There was not any summary metric to grade overall performance. A Python program to calculate these measures (http://ir.ohsu.edu/genomics/trecgen2007_score.py) with the appropriate gold standard data files is available.

Passage-level retrieval performance - character-based MAP

The original passage retrieval measure for the 2006 track was found to be problematic in 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 measure (Passage2 MAP) 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. We used Passage2 MAP as the primary passage retrieval evaluation measure in 2007.

The original Passage MAP measure was also calculated. This measure computed individual precision scores for passages based on character-level precision, using a variant of a similar approach used for the TREC 2004 HARD Track [2]. For each nominated passage, a fraction of characters overlaps with those deemed relevant by the judges in the gold standard. At 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 at all were added into the calculation as well, with precision set to 0 for relevant passages not retrieved. Then the mean of these average precisions over all topics was calculated to compute the mean average passage precision.

Aspect-level retrieval performance - aspect-based MAP

Aspect retrieval was measured using the average precision for the aspects of a topic, averaged across all topics. For 2007, the aspects were the different named entities of the given type for each question. To compute this, for each submitted run, the ranked passages were transformed to two types of values, either:

- the aspects of the gold standard passage that the submitted passage overlaps with, or
- not relevant

This resulted in an ordered list, for each run and each topic, of aspects and not-relevant. 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 them from the output to be analyzed and only kept the first appearance of an aspect. For these remaining aspects of a topic, we calculated Aspect MAP similar to how it was calculated for documents.

Document-level retrieval performance - document-based 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 nonrelevant for that topic. System run outputs were similarly collapsed, with the documents appearing in the same order as the first time the corresponding PMID appears in a nominated passage for that topic. For a given system run, average precision was measured at each point of correct (relevant) recall for a topic, with Document MAP being the mean of the average precision values across topics.

Results

A total of 66 runs were submitted by 27 groups. Of the submitted runs, 49 were classified as automatic, 8 as manual, and 9 as interactive. Appendix 1 lists the type and description of each submitted run. Table 5 lists the performance statistics for all of the runs and for the runs subdivided by categories. Appendix 2 shows the overall scores for each run, sorted by each measure.

We also measured correlation of the four measures (Passage2 MAP, Passage MAP, Aspect MAP, and Document MAP) for each run. As is seen in Table 6, the new Passage2 MAP measure was highly correlated with Aspect MAP and Document MAP ($R^2 > 0.8$), with the older Passage MAP measure less correlated.

Table 5 - Descriptive statistics for all runs and subdivided by categories.

All	**Passage2 MAP**	**Passage MAP**	**Aspect MAP**	**Document MAP**
Min	0.0008	0.0029	0.0197	0.0329
Median	0.0377	0.0565	0.1311	0.1897
Mean	0.0398	0.0560	0.1326	0.1862
Max	0.1148	0.0976	0.2631	0.3286
Automatic				
Min	0.0008	0.0029	0.0197	0.0329
Median	0.0391	0.0587	0.1272	0.1954
Mean	0.0421	0.0582	0.1286	0.1891
Max	0.1097	0.0976	0.2494	0.3105
Manual				
Min	0.0032	0.0177	0.0204	0.0541
Median	0.0149	0.0276	0.1136	0.1696
Mean	0.0169	0.0328	0.0964	0.1526
Max	0.0458	0.0654	0.1503	0.2309
Interactive				
Min	0.0268	0.0394	0.1411	0.0892
Median	0.0384	0.0620	0.1865	0.1940
Mean	0.0475	0.0648	0.1868	0.2007
Max	0.1148	0.0968	0.2631	0.3286

Table 6 - MAP measure correlation matrix using Pearson correlation coefficient (all values significantly different from 0 with a significance level p < .05).

MAP	**Passage2**	**Passage**	**Aspect**	**Document**
Passage2	1	0.656	0.845	0.812
Passage	0.656	1	0.591	0.830
Aspect	0.845	0.591	1	0.775
Document	0.812	0.830	0.775	1

We attempted to analyze the automatic runs to discern whether there was any association between individual methods used (as reported in conference notebook papers and not final proceedings papers) and overall performance as measured by Passage2 MAP. The task was challenging since groups approached entity-based question answering with a myriad of methods. Submissions employed multiple approaches for query expansion, various levels of passage retrieval granularity, varying IR models with many different scoring schemes, and several methods of post-processing. In all, these runs exercised over 70 different features, any of which could have impacted Passage2 MAP separately or in combination. With so many features and a limited number of runs (43) having a corresponding notebook paper describing methods, data sparseness was an issue. We therefore distilled the features into high-level categories, or meta-features shown in Table 7.

If retrieval was done in two steps, e.g., to pare down results for secondary concept-based retrieval, and each step uses a different level of granularity for passage retrieval, we chose the granularity level of the second one in order to focus on features of the core strategy rather than a filtering step designed to reduce computer processing burdens. This only affected runs from ASU and Tsinghua. Each run was represented as a vector of meta-features deemed either present (1) or absent (0). The decision was binary since there is no uniform way to say something was partially done, such as in the case of fusion runs, or to weigh the impact of a paring step for concept-based retrieval. If fusion was done, the union of features used by the individual component runs was chosen since they presumably all contributed to the ultimate result. All meta-features were given the same weight. A hierarchical clustering algorithm using a centroid similarity metric grouped runs based on their meta-features, as shown in Figure 2. Runs were clustered as a “group” when their correlation was > 70%. Clustering using Dice’s coefficient similarity measure produced similar results.

Originally, we had also clustered by statistical rank group. This simply revealed that many different paths lead to roughly the same performance, and was less informative as far as whether individual meta-features had an overall positive or negative impact. Although not used for clustering, the rank group is included in the heat map to indicate how a run performed. Given that the MAP measures were highly correlated (see Table 6), only Passage2 MAP rank is shown for clarity.

Table 7 - Meta-features of runs.

Meta-Feature Name	**Description**
SynExp	query expansion with synonyms
OrthExp	query expansion with orthographic variants using any source or method
ParGranularity	passage retrieval by paragraph
SentGranularity	passage retrieval by sentence
BlckGranularity	passage retrieval by block, including blocks of words or sentences greater than a single sentence yet smaller than a paragraph
ConcptIR	concept-based retrieval - a general retrieval strategy attempting to align concepts and, for some runs, relationships between a topic and a passage; uses external knowledge sources such as UMLS as a source of "concepts"; and finds concepts in the results as an inherent part of the retrieval process rather than a post-processing step to "trim" a passage
TermIR	term-based retrieval – a general retrieval strategy focusing on terms rather than concepts
FusionIR	fusion - combining results from 2 or more systems regardless of fusion operator used
TfIdfIR	passage retrieval using a vector space model with any variant of TF-IDF
OkapiIR	passage retrieval using a vector space model with any variant of Okapi
DfrIR	passage retrieval using a vector space model with any variant of divergence from randomness (DFR)
LatentSemIR	passage retrieval using a vector space model with any variant of latent semantic analysis
LmIR	passage retrieval using any language model
Feedback	feedback using pseudo-relevance feedback or a custom method
FilterPostProc	filter post-processing - removing passages for any reason
TrimPostProc	passage trimming - post-processing of passages by removing sentences from the ends regardless of method

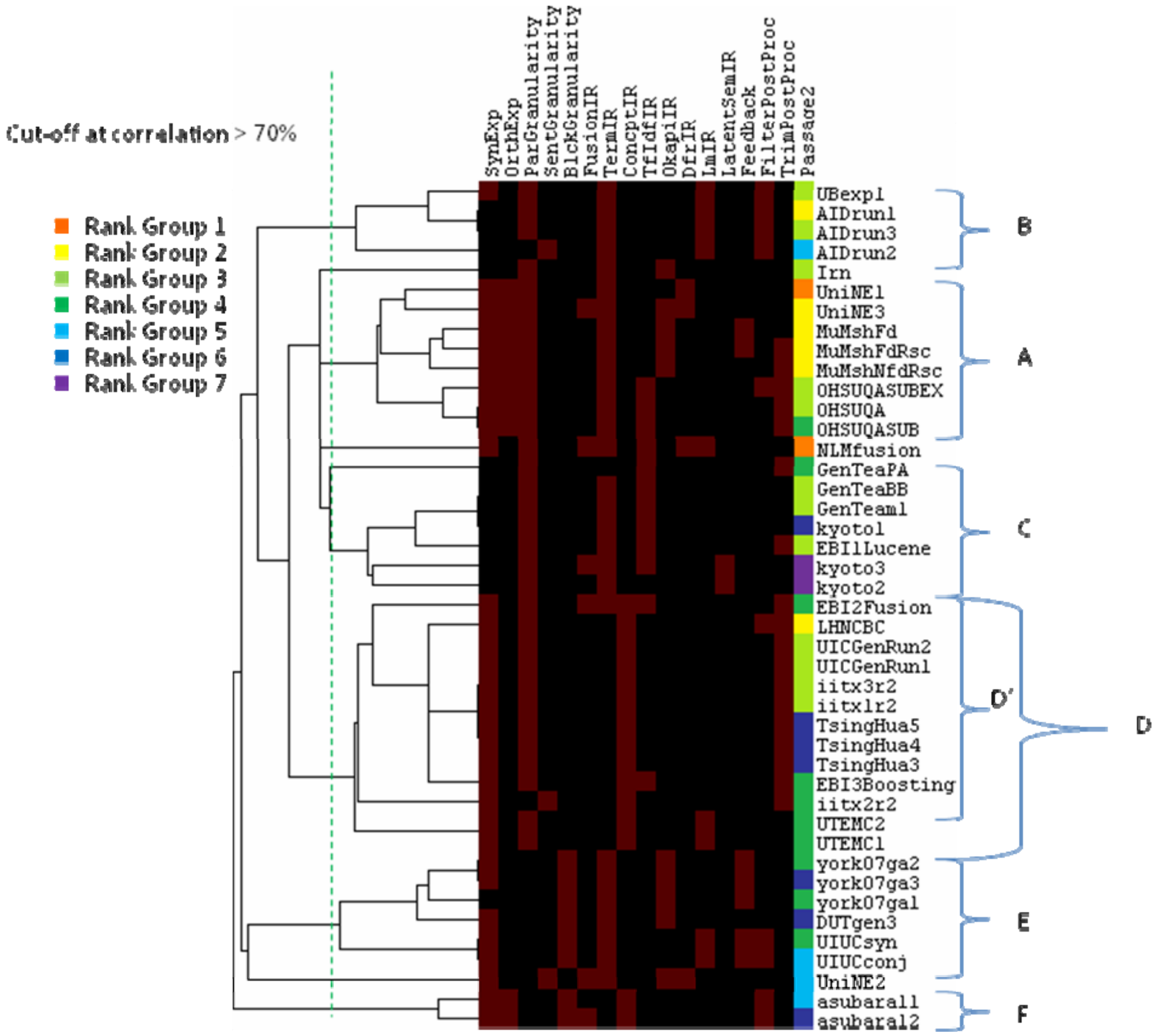


Figure 2 - Heat map for meta-features, their use in runs, and rank group clustering.

Focusing on groups which “on average” used similar methods allowed us to make generalizations about some of the strategies used. Inevitably, abstracting out features in this manner does not precisely identify sources of changes in performance. Furthermore, important details such as corpus pre-processing was not included since papers often lacked details on how this was done. In spite of these limitations, however, we could make some general observations and identify potential causes for performance differences.

Group A runs expanded queries with synonyms and orthographic variants, defined passages by paragraph, and used vector space models for retrieval. Feedback, trimming, and passage removal all changed Passage2 MAP, but by no more than 10%. The greatest performance decline occurred between MuMshNfdRsc and OHSUQA. The most prominent difference between the runs was the use of Okapi (MuMshNfdRsc) versus TF-IDF (OHSUQA), though OHSU added some words related to the entity types to their queries. Furthermore, it was not clear that some aspect of corpus preprocessing contributed significantly to the decline.

Group B runs used various language models for retrieval and filtering passages. Group C runs used no query expansion, defined passages by paragraph, and retrieved them using a TF-IDF based vector space model. In spite of the differences in the approaches used in groups B and C, they performed similarly with the exception of AIDRun2, which defined passages by sentence, and the Kyoto runs. Kyoto1, the only run in group B belonging to a low rank group, used a different scoring scheme than the pivoted-document normalization used by the others. Unlike the runs in group A, not all runs in groups B and C performed above the mean and median on Passage2.

Group E included those runs defining passages by block. With one exception, all runs performed below the mean on Passage2 even though each used methods employed by higher scoring runs. Additionally, experiments conducted by Neuchatel, IIT, and Amsterdam suggest that defining passages by units other than paragraphs hurt performance.

Groups D and F represented the concept-based retrieval runs. The former used methods such as synonym query expansion, defined passages by paragraph and, for subgroup D', trimming of passages to ensure high concept density. Group F differed from Group D primarily in that passages were defined by block rather than paragraph. If submissions defining passages at the paragraph level (since any other seems to degrade performance, see Group E), are compared by those examining concepts (Group D) and those simply using terms (every other group besides Groups D and F), the mean and median dropped on all metrics though most significantly for Passage 2 (31% decline for the mean, 15% for the median). At best, the extra processing required for concept-based retrieval did not seem to help as a general approach. Only LHNCBC and the two UIC runs performed above both the TREC mean and median on all metrics. The exact impact of concept retrieval was difficult to ascertain as most runs did not compare it against a baseline; only NLM and EBI attempted to do so (with no difference and a decline, respectively). SUNY Buffalo, although not submitting an automatic run using concept retrieval, submitted a manual one representing each passage as a list of concepts to be compared to those of the topic. All three metrics dropped significantly.

According to NLM, the effect of trimming was neutral. The other runs in Group D' did not examine the effects of removing it, but there were runs outside this group that did so. Like them, OHSU and Geneva used external knowledge to identify the part of a passage with the highest density of "concepts" matching the topic. However, the ordering of passages returned from the retrieval step was unchanged. OHSU reported a small improvement (6%) on Passage2 MAP, but Geneva's dropped 41% for the same metric. This was a surprising result in that both methods employed NER, albeit differently. Other runs that trim using only word matches had results more in line with NLM's and OHSU's. Melbourne improved slightly by 4% and EBI improved by 7%, both for Passage2 MAP.

Across groups, synonym expansion was a popular method. Ostensibly, submissions using it scored about 20% higher on Passage2 MAP and Aspect MAP with no significant difference on Document MAP. But those groups conducting runs with and without synonym expansion differed in their results. Some, like OHSU, Melbourne, and Neuchatel, improved on all metrics (up to 40% for Passage2 MAP, 51% for Aspect MAP, 44% for Document MAP). However, some like EBI and York did worse (up to 39% decrease for Passage2 MAP, 40% for Aspect MAP, 19% for Document MAP). Yet others like UIUC only improved marginally on Document MAP (10%). Such an equivocal outcome may have been due to the fact that groups used different knowledge sources for synonyms and/or processed those knowledge sources in different ways that resulted in different precision/recall tradeoffs for synonym expansion.

The performance of NLMFusion, the top scoring automatic run for all three metrics, suggested that combining results from different IR models may improve score. But other runs using fusion (UniNE3, EBI2Fusion, and kyoto3) showed slight declines in performance from their baseline non-fusion runs. Each used a different method, however, for fusing the individual runs, and this may have contributed to the differences in performance. Divergence from randomness (DFR) was another approach used in the NLMFusion run by its highest scoring subcomponent run. Neuchatel also reported success in using it. However, with only two groups using it in any form, it is hard to say in general that it is a superior method to other lexical-statistical methods.

Discussion

Although our analysis is incomplete and difficult to interpret due to incomplete experimentation and reporting, we can draw some conclusions from the results. In terms of the overall results, the level of performance of the top systems was somewhat lower than the TREC 2006 Genomics Track. This may imply that the list-entity type question was more difficult than the GTT question. This would not be unexpected since list entity questions are more open-ended, involve more different entity types, and are closer to natural language than the GTT question used last year. The top systems did consistently well on all measures. The measures were highly correlated.

We can also conclude that, unlike last year, Aspect MAP was a meaningful measure of system topic coverage in the 2007 track. While the range of the average number of aspects per relevant passages was low (1-3), the number of aspects per topic was relatively high (could be over 300). Therefore for a system to do well on the Aspect MAP measure, a number of passages with complementary aspect information had to be retrieved and ranked highly, since for most topics, almost no single passages would cover all of the required entities.

We are able to draw some conclusions from our extraction of meta-features and their comparison with results of runs as reported in conference notebook papers. First, we conclude that no single strategy or combination of strategies was clearly superior, as indicated by both the diversity of methods used by runs clustering in the same rank group and the diversity of scores within the same methods cluster group. Second, concept-based retrieval using external knowledge sources, as used by the runs in the competition, at the very least did not help results in spite of the extra processing. Third, results with synonym query expansion, once again with external knowledge sources, were mixed but tended to improve results. Finally, passage retrieval by sentence or block-level was detrimental to performance compared to paragraph-level. Clearly, further experimentation as well as descriptions of runs must be provided by participating groups to reach conclusions about performance of features with more confidence.

Future Directions

The 2007 track is the last year of the TREC Genomics Track. We are exploring future challenge evaluations in biomedicine, probably in concert with the ImageCLEF medical image retrieval task [3]. We hope that the test collections created over the years of the track will be used for further research in biomedical information retrieval and related areas. We will continue to maintain the track Web site for the foreseeable future, with the resources posted there as well as instructions for accessing them.

Acknowledgements

The TREC Genomics Track was funded by grant ITR-0325160 from the U.S. National Science Foundation. The track thanks Ellen Voorhees, Ian Soboroff, and Lori Buckland of NIST for their help in various ways. We also thank John Sack and Highwire Press for facilitating the use of documents from the respective publishers as well as the U.S. National Library of Medicine for their use of MEDLINE data.

References

1. Hersh W, et al. *TREC 2006 Genomics Track overview*. *The Fifteenth Text Retrieval Conference (TREC 2006)*. 2006. Gaithersburg, MD: National Institute for Standards & Technology. 52-78. http://trec.nist.gov/pubs/trec15/papers/GEO.OVERVIEW.pdf.
2. Allan J. *HARD Track overview in TREC 2004 - high accuracy retrieval from documents*. *The Thirteenth Text Retrieval Conference (TREC 2004)*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/HARD.OVERVIEW.pdf.
3. Hersh WR, et al., *Advancing biomedical image retrieval: development and analysis of a test collection.* Journal of the American Medical Informatics Association, 2006. 13: 488-496.

Appendix 1 - Type and description of submitted runs.

Run	Type	Description
AIDrun1	A	Baseline
AIDrun2	A	Same as baseline (AIDrun1), but with more elaborate passage identification.
AIDrun3	A	baseline, re-ranked according to a language model of the entity-type assiciated with the topic.
asubaral1	A	Finding relatedness between words in the passages and keywords in the corresponding question - keyword expansion by utilizing the terms appearing in the definitions of the keywords of the questions
asubaral2	A	Using both Lucene and Indri indexing systems for retrieval - passage length is as minimal as possible - finding relatedness between words in the passages and keywords in the corresponding question - keyword expansion by utilizing the terms appearing in the definitions of the keywords of the questions
asubaral3	I	Similar to our first run, except we interactively modified queries to improve the answers
biokiP	I	Interactive selection of keyphrases and weights. Result span expanded to paragraph.
biokiS	I	Interactive selection of keyphrases and weights. Result span expanded to sentence boundaries.
biokiST	I	Interactive selection of keyphrases and weights. Entity-type matching on some queries. Result span expanded to sentence boundaries.
DUTgen1	I	Indri; named entity recognition; sentence-leval overlapped window; query expansion based on MeSH;post-processing using templates
DUTgen2	I	Indri; named entity recognition; sentence-leval overlapped window; query expansion based on MeSH;result combination
DUTgen3	A	BM25; named entity recognition; NP extraction for topics; sentence-leval overlapped window; query expansion based on MeSH; result combination;
EBI1Lucene	A	System based on a Lucene index that considers the spans as documents. The Lucene scoring function has been modified to better deal with large and small documents based on the article by Singhal about pivoted cosine normalization. A postprocessing of the spans has been done removing HTML without any content and by finding the zone with relevant information based on the similarity of the query and the sentences in the spans.
EBI2Fusion	A	This run is the fusion of two configurations of our system. Our system is based on a Lucene index that considers the spans as documents (configuration 1); in addition, query expansion and boosting of some spans based on the entities matched between the query and the spans can be done (configuration 2). The Lucene scoring function has been modified to better deal with large and small documents based on the article by Singhal about pivoted cosine normalization. In addition, the spans have been processed by removing HTML without any content and by finding the zone with relevant information based on the similarity of the query and the sentences in the spans.
EBI3Boosting	A	System based on a Lucene index that considers the spans as documents. The Lucene scoring function has been modified to better deal with large and small documents based on the article by Singhal about pivoted cosine normalization. A postprocessing of the spans has been done analyzing the entities in the query and in the span and boosting the spans based on the entities that are matched. In addition, the spans have been processed by removing HTML without any content and by finding the zone with relevant information based on the similarity of the query and the sentences in the spans.
fdgerun1	A	Automatically extract the relevant concepts of each topic, retrieval sentences according to those concepts.
fdgerun2	A	Automatically extract relevant concepts of each topic, combine the result of sentence retrieval and the one of context retrieval.
fdgerun3	M	Score each sentence according to the concurrency of concept terms from different groups.
GenTeaBB	A	Same as GenTeam1 but with Boolean boosting.
GenTeam1	A	Basic run using easyIR as IR engine (dtu.dtn, Porter).
GenTeaPA	A	Same run as GenTeaBB, but with passage selection based on assesing the density of semantic targets.
HFmanual	M	Hongfang's run
hltcairo1	A	First 25 results from the search engine
hltcairo2	A	First 50 results from the search engine

icbdoc	M	Rank fusion of seven different search techniques implemented in Twease (includes both automatic and manual runs). Optimized for document MAP.
icbpassage	M	Rank fusion of seven different search techniques implemented in Twease (includes both automatic and manual runs). Passages are marked as a post-processing step after document retrieval and fusion. At most, ten passages are included per document. Optimized for passage MAP.
icbtwease	M	Manual run with minimal interval semantics performed with Twease using a slider value of 80. This run was performed with the same Twease software version deployed at Twease.org as of July 2007.
iitx1r2	A	MST passage extraction by concept sc = (1.0*result.getPassConceptSCNorm() + 0.1*result.getSentConceptSCNorm() + 1.0*result.getPassConceptIdfSumNorm() + 0.1*result.getSentConceptIdfSumNorm())
iitx2r2	A	MST passage extraction by concept sc = (1.0*result.getPassConceptSCNorm() + 0.1*result.getSentConceptSCNorm() + 1.0*result.getPassConceptIdfSumNorm() + 0.1*result.getSentConceptIdfSumNorm()) With sentence boosting dependency grammar, sumidf, nconcepts
iitx3r2	A	MST passage extraction by concept sc =((passConceptSCNorm+sentConceptSCNorm+passConceptIdfSumNorm+sentConceptIdfSumNorm)/4)
IRn	A	This run has been performed by applying the Information Retrieval technique based on passages. The passages are composed of four sentences. The indexing of the document collection applies the Okapi measure.
kyoto1	A	Paragraph-level impact-based retrieval combined with a probabilistic model for term co-occurrence. Passages scored using a variant of TF/IDF, but results are ranked using only the IR system's scores.
kyoto2	A	Paragraph-level impact-based retrieval combined with a probabilistic model for term co-occurrence. Passages scored using a variant of TF/IDF, but more results were used and then ranked using only the PM's scores.
kyoto3	A	Paragraph-level impact-based retrieval combined with a probabilistic model for term co-occurrence. Passages scored using only the probabilistic model and final ranking determined using equal weight on both systems.
LHNCBC	A	An automatic run based on LHC's Essie search engine and for which results are reranked based on relationships extracted from Essie results using MetaMap and SemRep.
MuMshFd	A	Automatic query expansion with entities and ontological terms, but without passage reduction and reranking.
MuMshFdRsc	A	Automatic query expansion with entities and ontological terms, followed by passage reduction and reranking.
MuMshNfdRsc	A	Automatic query expansion with ontological terms only, followed by passage reduction and reranking.
ncbi2007a	A	Reranked Essie hits from NCBI
ncbi2007b	A	generated by Larry
NLMfusion	A	An automatic run obtained by applying fusion to the LHNCBC run, a Terrier run, an NCBI Themes run, an INDRI run and an easyIR run.
NLMinter	I	An interactive run based on an interactively created filter applied to the NLMfusion run.
OHSUQA	A	Two stage query generation with MESH and gene synonym expansion, and entity-specific keywords. Lucene maximal passage index, TF*IDF. MMTX based sentence entity count passage trimming.
OHSUQASUB	A	Two stage query generation with MESH and substances expansion, and entity-specific keywords. Lucene maximal passage index, TF*IDF. MMTX based sentence entity count passage trimming.
OHSUQASUBEX	A	Two stage query generation with MESH and substances expansion, and entity-specific keywords. Lucene maximal passage index, TF*IDF. MMTX based sentence entity count passage trimming.
TsingHua3	A	(run3)Machine learning and dictionary based NE recognition, BM2500, Treble passage retrieval.
TsingHua4	A	(run4) Machine learning and dictionary based NE, sigma local df for weighting, Treble passage retrieval
TsingHua5	A	(run5_new) strictly generated dictionary for NE, max local df for weighting, Treble passage retrieval, reduction

UBexp1	A	Automatic runs generated with Indri. Queries were build automatically by expanding them with MetaMap and discarding common terms. Gene and proteins names were expanded using Gene Ontology. Query formulated using synonyms to represent the expanded terms and multiword phrases when appropriate. Reference sections were discarded by restricting results to those passages that did not have the word "Medline".
UBHFmanual	M	Queries were expanded using publicly available resources. This list was manually filtered to discard ambiguous names of gene and proteins.
uchsc1	M	Queries were manually expanded and individual terms were assigned weights. Lists of terms matching keyword classes were included in the queries; those terms recieved equal weights. The queries were submitted to the Indri search engine of the Lemur toolkit. Post-processing included filtering out passages that did not contain genes, mutations or biological substances, according to query type.
uchsc2	I	Queries were manually expanded and individual terms were assigned weights based on MeSH distance. Additionally, salient biomedical predicates were also expanded for 5 of the queries. Lists of terms matching keyword classes were included in the queries; those terms recieved equal weights. The queries were submitted to the Indri search engine of the Lemur toolkit. Post-processing included filtering out passages that did not contain genes, mutations or biological substances, according to query type.
UICGenRun1	A	Utilize UMLS to get some of the entities.
UICGenRun2	A	Do not differentiate the importance of entities in passages as long as some entity presents in passages.
UIowa07Gen01	M	title of reference identified from logical document structure
UIUCconj	A	automatic run
UIUCrelfb	I	Interactive run with relevance feedback
UIUCsyn	A	automatic run with synonym expansion
UniNE1	A	Retrieval based on Divergence from randomness. Query expansion using forms generated from query words. The length of a passage is delimited by the <p> tag.
UniNE2	A	Data fusion of three IR systems. 1 Retrieval based on Okapi model with query expansion using forms generated from query words. 2 Retrieval based on Okapi model, using only the original query words. Re-ranking based on distance between query words and entity in the query. 3 Retrieval based on Divergence from randomness. Query expansion using forms generated from query words. Each passage is a sentence.
UniNE3	A	Data fusion of three IR systems 1- Retrieval based on Divergence from randomness. Query expansion using forms generated from query words and word variant generation for entity and query terms. 2 Retrieval based on Okapi model with query expansion using forms generated from query words. 3 Retrieval based on Divergence from randomness. Query expansion using forms generated from query words. Re-ranking based on distance between query words and entity in the query. Each passage is delimited by the <p> tag.
UTEMC1	A	UMLS-based thesaurus in combination with language-modeling. Run optimized for aspect-retrieval.
UTEMC2	A	UMLS-based thesaurus in combination with language-modeling. Run optimized for precision.
york07ga1	A	No query expansion. Use only terms extracted from the raw topics for retrieval. Use BM25 for term weighting in structured queries. Use Okapi to build word-based index.
york07ga2	A	Expand query terms for 11 gene-related topics by using Entrez Gene. Use BM25 for term weighting in structured queries. Use Okapi to build sentence-based index.
york07ga3	A	Expand query terms for all the topics by using UMLS. Use BM25 for term weighting in structured queries. Use Okapi to build word-based index.

Appendix 2 - Overall MAP for each run, sorted by each measure.

Run	Passage2	Run	Passage	Run	Aspect	Run	Document
NLMinter	0.1148	UICGenRun2	0.0976	NLMinter	0.2631	NLMinter	0.3286
NLMfusion	0.1097	NLMinter	0.0968	NLMfusion	0.2494	NLMfusion	0.3105
UniNE1	0.0988	york07ga1	0.0947	biokiP	0.2254	MuMshFd	0.2906
UniNE3	0.0970	iitx3r2	0.0940	UniNE1	0.2189	MuMshFdRsc	0.2880
MuMshFd	0.0895	iitx2r2	0.0926	MuMshNfdRsc	0.2079	UniNE1	0.2777
MuMshFdRsc	0.0893	NLMfusion	0.0921	MuMshFd	0.2068	UniNE3	0.2710
MuMshNfdRsc	0.0809	UniNE3	0.0914	UniNE3	0.2043	MuMshNfdRsc	0.2682
UBexp1	0.0698	MuMshFdRsc	0.0880	LHNCBC	0.2030	iitx2r2	0.2462
AIDrun1	0.0694	york07ga2	0.0859	ncbi2007a	0.2022	iitx1r2	0.2454
LHNCBC	0.0680	iitx1r2	0.0852	MuMshFdRsc	0.2016	iitx3r2	0.2414
GenTeaBB	0.0665	MuMshFd	0.0840	IRn	0.1976	AIDrun1	0.2412
GenTeam1	0.0647	UICGenRun1	0.0834	biokiS	0.1968	UTEMC2	0.2398
ncbi2007a	0.0612	UIUCrelfb	0.0811	biokiST	0.1923	UICGenRun2	0.2393
IRn	0.0606	biokiST	0.0803	DUTgen1	0.1865	IRn	0.2351
UICGenRun2	0.0511	UniNE1	0.0802	UICGenRun2	0.1807	UTEMC1	0.2335
biokiST	0.0472	MuMshNfdRsc	0.0794	GenTeaBB	0.1795	icbdoc	0.2309
york07ga2	0.0472	UniNE2	0.0787	UBexp1	0.1790	LHNCBC	0.2266
biokiS	0.0462	biokiS	0.0768	asubaral3	0.1782	ncbi2007a	0.2222
uchsc2	0.0458	UTEMC2	0.0738	GenTeam1	0.1749	biokiP	0.2222
uchsc1	0.0458	AIDrun2	0.0708	UIUCsyn	0.1629	biokiS	0.2222
UICGenRun1	0.0445	UTEMC1	0.0687	AIDrun1	0.1561	biokiST	0.2216
iitx3r2	0.0442	icbpassage	0.0654	UTEMC1	0.1535	UBexp1	0.2209
OHSUQA	0.0440	fdgerun2	0.0650	uchsc2	0.1525	york07ga1	0.2153
OHSUQASUBEX	0.0439	AIDrun1	0.0636	EBI1Lucene	0.1513	york07ga2	0.2150
OHSUQASUB	0.0434	UIUCsyn	0.0633	uchsc1	0.1503	UICGenRun1	0.2092
EBI1Lucene	0.0404	DUTgen1	0.0620	UTEMC2	0.1500	GenTeam1	0.1991
EBI2Fusion	0.0401	GenTeam1	0.0615	EBI2Fusion	0.1470	UIUCsyn	0.1962
AIDrun3	0.0399	GenTeaBB	0.0611	UIUCrelfb	0.1452	GenTeaBB	0.1962
GenTeaPA	0.0392	LHNCBC	0.0609	UICGenRun1	0.1451	GenTeaPA	0.1962
UIUCsyn	0.0391	york07ga3	0.0595	GenTeaPA	0.1415	AIDrun2	0.1954
iitx1r2	0.0388	DUTgen2	0.0595	DUTgen2	0.1411	UIUCrelfb	0.1940
DUTgen1	0.0384	DUTgen3	0.0587	icbdoc	0.1390	york07ga3	0.1917
biokiP	0.0378	UBexp1	0.0576	icbtwease	0.1317	UniNE2	0.1903
UTEMC2	0.0376	TsingHua4	0.0555	york07ga2	0.1306	kyoto1	0.1892
york07ga1	0.0373	GenTeaPA	0.0551	asubaral1	0.1303	DUTgen2	0.1851
UTEMC1	0.0367	ncbi2007a	0.0549	UIUCconj	0.1302	icbpassage	0.1833
UIUCrelfb	0.0364	fdgerun1	0.0535	iitx1r2	0.1272	DUTgen1	0.1818
DUTgen2	0.0339	UIUCconj	0.0497	iitx3r2	0.1253	EBI1Lucene	0.1799
EBI3Boosting	0.0339	TsingHua5	0.0490	EBI3Boosting	0.1247	UBHFmanual	0.1799
iitx2r2	0.0335	IRn	0.0486	kyoto1	0.1208	HFmanual	0.1773
DUTgen3	0.0314	kyoto1	0.0474	iitx2r2	0.1166	EBI2Fusion	0.1768
UIUCconj	0.0296	TsingHua3	0.0463	UBHFmanual	0.1138	fdgerun2	0.1759
UniNE2	0.0278	uchsc1	0.0460	HFmanual	0.1134	OHSUQA	0.1719
asubaral3	0.0268	uchsc2	0.0459	OHSUQASUBEX	0.1104	DUTgen3	0.1705
AIDrun2	0.0248	icbdoc	0.0420	asubaral2	0.1102	OHSUQASUBEX	0.1695
york07ga3	0.0227	EBI1Lucene	0.0420	UniNE2	0.1102	OHSUQASUB	0.1684
fdgerun2	0.0216	asubaral3	0.0416	OHSUQASUB	0.1080	icbtwease	0.1619
kyoto1	0.0209	EBI2Fusion	0.0404	OHSUQA	0.1075	uchsc2	0.1614
UBHFmanual	0.0189	OHSUQA	0.0403	york07ga1	0.1017	uchsc1	0.1610
HFmanual	0.0188	biokiP	0.0394	fdgerun2	0.0894	TsingHua4	0.1603
TsingHua4	0.0182	OHSUQASUBEX	0.0392	DUTgen3	0.0883	AIDrun3	0.1536
fdgerun1	0.0178	AIDrun3	0.0390	AIDrun2	0.0882	fdgerun1	0.1522
TsingHua5	0.0168	OHSUQASUB	0.0388	AIDrun3	0.0848	EBI3Boosting	0.1522
asubaral1	0.0157	icbtwease	0.0354	fdgerun1	0.0769	UIUCconj	0.1495
icbtwease	0.0156	asubaral2	0.0351	icbpassage	0.0691	TsingHua5	0.1413
TsingHua3	0.0145	EBI3Boosting	0.0346	TsingHua5	0.0670	TsingHua3	0.1331
icbdoc	0.0141	asubaral1	0.0287	TsingHua4	0.0642	kyoto2	0.1191
asubaral2	0.0140	kyoto2	0.0235	york07ga3	0.0611	kyoto3	0.1022

icbpassage	0.0123	kyoto3	0.0204	TsingHua3	0.0560	asubaral2	0.0932
ncbi2007b	0.0111	fdgerun3	0.0199	ncbi2007b	0.0552	asubaral3	0.0892
fdgerun3	0.0068	UBHFmanual	0.0179	fdgerun3	0.0333	asubaral1	0.0737
kyoto3	0.0065	UIowa07Gen01	0.0178	kyoto3	0.0312	fdgerun3	0.0725
kyoto2	0.0054	HFmanual	0.0177	kyoto2	0.0302	ncbi2007b	0.0568
UIowa07Gen01	0.0032	ncbi2007b	0.0095	UIowa07Gen01	0.0204	UIowa07Gen01	0.0541
hltcairo2	0.0013	hltcairo2	0.0042	hltcairo2	0.0203	hltcairo2	0.0396
hltcairo1	0.0008	hltcairo1	0.0029	hltcairo1	0.0197	hltcairo1	0.0329