Combining resources to find answers to biomedical questions

Dina Demner-Fushman,[a] Susanne M. Humphrey,[a] Nicholas C. Ide,[a] Russell F. Loane,[a] James G. Mork,[a] Patrick Ruch,[b] Miguel E. Ruiz,[a] Lawrence H. Smith,[a] W. John Wilbur,[a] Alan R. Aronson[a]

[a]National Library of Medicine, Bethesda, Maryland
{ddemner, shumphrey, nide, rloane, mork, ruizmi, larsmith, wilbur, alaronson} @mail.nih.gov

[b]University Hospital of Geneva, Geneva, Switzerland
patrick.ruch@sim.hcuge.ch

Abstract
One of the NLM experimental approaches to the 2007 Genomics track question answering task followed the track evaluation design: we attempted identifying exact answers in the form of semantic relations between biomedical entities named in questions and the potential answer types and then marked the passages containing the relations as containing the answers. The goal of this knowledge-based approach was to improve the answer precision. To boost recall, evidence obtained through relation extraction was combined with passage scores obtained by semantic filtering and passage retrieval.

Our second approach, the fusion of retrieval results of several search engines established to be reliably successful in the past, was used as the baseline, which ultimately was not improved upon by the knowledge-based approach.

The impact of the relevance of whole documents on finding passages containing answers was tested in the third approach, an interactive retrieval experiment, in which the relevance of a document was determined by virtue of its retrieval in an expert PubMed® search and an occasional examination of its abstract. This relatively moderately labor-intensive approach significantly improved the fusion retrieval results.

Keywords: Genomics; MEDLINE/PubMed; MeSH; Statistical Natural Language Processing; Machine Learning; Thematic Analysis.

1. Introduction

The 2007 TREC Genomics track focused on answering questions gathered from working biologists. Rather than finding an exact answer, the task was to extract passages containing answers from about 160,000 full-text scientific articles published in 49 genomics-related journals. As the task required finding passages potentially containing lists of named entities of a given answer type, and sample questions for each answer type were provided, we attempted to find answers in the form of relations characteristic of a given answer type. The relations were extracted using SemRep (Rindflesch et al, 2003), a natural language processing system that relies on semantics and domain knowledge encoded in the Unified Medical Language System® (UMLS®) (Lindberg et al, 1993) to determine relations between entities found in the text by MetaMap (Aronson, 2001). As there were no restrictions on question type and form, we did not expect to see all question types in the training set, and therefore we did not attempt a deeper question understanding. Because of this fairly general approach and because SemRep was used "as is" without any adjustments for the task, we anticipated missing quite a few answers and the relevant passages containing these answers. Therefore we decided to combine the knowledge-based approach with corpus-based and statistical methods. The latter approaches rely upon Essie, the LHNCBC experimental search engine for biomedical text in structured XML (Ide et al, 2007). Because SemRep processing of the whole collection is computationally intensive, only the 1,000 top documents retrieved by Essie were submitted to SemRep.

In the 2005 and 2006 tasks, our information retrieval approach, in which the spans retrieved by the base systems were merged using the sum fusion method (Fox and Shaw, 1994), achieved good performance. This year, we used this approach as the baseline, combining retrieval results of Essie, Theme (Wilbur, 2002), Indri (Metzler and Croft, 2004), EasyIR (Ruch et al, 2006), and Terrier (Ounis I et al, 2006).

Building upon the approach to document processing developed and tested in the 2006 evaluation, we treated each passage of text delimited by the HTML paragraph tags (the maximum-length legal spans) as an individual document. We also tested if passage retrieval could be improved by adding information about relevance of the whole full-text document containing the passage (see section 3.3). Section 2 describes preparation of the documents. Section 3

provides a detailed description of our methods. Section 4 presents some preliminary results.

2. Document Preparation

We pulled all 12,641,127 spans identified in the official legalspans.txt file from their respective articles. To make processing easier for our various tools, we created a "cleaned" version by removing all of the HTML tags, converting HTML codes into their respective ASCII characters, and replacing all of the UTF-8 characters with their ASCII equivalents (see Figure 1 for an example). The "cleaned" spans were then indexed using EasyIR, Essie, Indri, and Terrier. The top 1,000 spans retrieved by Essie for each of the 36 official topics resulted in 29,746 unique spans to be further processed using SemRep. To facilitate processing, we split the spans into smaller sub-spans, where possible. We developed two different methods of identifying and splitting out the specific components from the spans. These two methods worked on reference-based spans since these tended to be longer and more in need of breaking up.

Original Span	34. Dunst J, Jurgens H, Sauer R et al. Radiation therapy in Ewing’s sarcoma: an update of the CESS 86 trial. <I>Int J Radiat Oncol Biol Phys</I> 1995; <B>32</B>: 919–930.<!-- HIGHWIRE ID="13:1:23:34" --><A HREF="/cgi/external_ref?access_num=A1995RL55500003&link_type=ISI" >[ISI]</A><A HREF="/cgi/external_ref?access_num=7607966&link_type=MED" >[Medline]</A><!-- /HIGHWIRE --><P><A NAME="MDF047C35"><!-- null --></A>
"Cleaned" Span	34. Dunst J, Jurgens H, Sauer R et al. Radiation therapy in Ewing's sarcoma: an update of the CESS 86 trial. Int J Radiat Oncol Biol Phys 1995; 32: 919-930. [ISI][Medline]
Parsed Sub-span	Radiation therapy in Ewing's sarcoma: an update of the CESS trial.
Span ID	11863105_122_68154_428 [ID: 11863105, Unique Span #within document: 122, starting byte position: 68154, and length of span: 428]

Figure 1: Span cleaning and parsing example

Sub-spanning Method 1:

The first method involved identifying spans that started with the sequence of "<number> <period> <space>". These reference spans were very well behaved and easy to parse, for example, see Figure 1 where we identified the span "34. Dunst …" as meeting the conditions for this method, and then we were able to correctly parse out most of the article's title from the reference. If this span had contained multiple references, all of them meeting the conditions for this method would have been parsed and assigned a unique sub-span number as well.

Sub-spanning Method 2:

The second method used a set of 24 trigger phrases (see Figure 2 for the list) to identify the ending of each specific article reference in a large reference section span.

[Abstract]	[Editorial]	[Letter]	[Swedish]
[Abstr.]	[Engl. Ed]	[Med]	[abstract
[Abstract	[Extract]	[Medline]	[article
[Clinical conference]	[Free Full Text]	[Online]	[comment]
[CrossRef]	[Full Text]	[Review]	[editorial]
[Dissertation]	[In	[Russian]	[letter

Figure 2: List of Reference Section Trigger Phrases

The list of trigger phrases was identified by a manual review of some "cleaned" spans. We used these trigger phrase positions at the end of each reference to parse the larger span into sub-spans broken on the trigger phrases. This method was not able to parse out the specific title for each article reference as was Method I, but by using the smaller sub-spans we were better able to narrow down the part of the span that was triggering our results. Figure 3 has an example showing the trigger phrases in bold.

Identifying Reference Spans:

We wanted to be able to identify spans that were reference sections so we could attempt to parse the data into smaller chunks to better refine our results. We came up with two metrics for determining whether a span was a reference or not. The first method used the list of trigger phrases described above as *Sub-spanning Method 2*, except for reference span determination where we focused only on spans that were greater then 3,000 characters in length. If the span contained one of the triggers, it was determined to be a reference span. The second metric used punctuation to decide whether a span was a reference. We simply counted the punctuation characters in a span and if the percentage of punctuation to total length was greater then 3.19%, the span was considered to be a reference span. We

used a manual review of some "cleaned" spans to determine the 3.19% cut-off, where we focused on only three punctuation symbols (period, comma, and colon) since counting all of the punctuation gave us worse results. For this second metric, we also exempted any spans that started with "Received" or "Accepted" since they contained a lot of punctuation, but were definitely not references. Figure 4 details examples of where the metrics worked well and where they didn't. We identified 11,126 of the 29,746 unique spans (59.75%) in our final results as being reference spans. Given more time, we would have done more refinement of the reference span identification and also improved the parsing that was done. The list of trigger phrases needs to be cleaned up to better refine what is being selected as a reference section. The initial parsing on the well-formed references provided much better text to process and it would have been nice to expand that processing into the type II references

"Cleaned" Span	American College of Rheumatology Task Force on Osteoporosis Guidelines: Recommendations for the prevention and treatment of glucocorticoid-induced osteoporosis. Arthritis Rheum1996;39: 1791-1801[ISI]**[Medline]** . . . Weinstein RS, Jilka RL, Parfitt AM, Manolagas SC. Inhibition of osteoblastogenesis and promotion of apoptosis of osteoblasts and osteocytes by glucocorticoids. Potential mechanisms of their deleterious effects on bone. J Clin Invest1998;102: 274-282**[Abstract/Free Full Text]** . . .
Example Sub-span	American College of Rheumatology Task Force on Osteoporosis Guidelines: Recommendations for the prevention and treatment of glucocorticoid-induced osteoporosis. Arthritis Rheum1996;39: 1791-1801[ISI][Medline]
Span ID	10809790_40_18217_6898 [ID: 10809790, Unique Span #within document: 40, starting byte position: 18217, and length of span: 6898]

Figure 3: Span splitting example

Example of Good Reference Designation Span (Has trigger phrase as well as 9.89% punctuation)	16176946_74_79835_470|Ackley, B. D., Crew, J. R., Elamaa, H., Pihlajaniemi, T., Kuo, C. J. and Kramer, J. M. (2001). The NC1/endostatin domain of Caenorhabditis elegans type XVIII collagen affects cell migration and axon guidance. J. Cell Biol. 152,1219 -1232.[Abstract/Free Full Text]
Incorrect Reference Designation Span (10.17% well above our cut-off)	15539493_6_3243_66|Key words: C. elegans, UNC-14, UNC-51, VAB-8, Axon guidance
Exempted Span (has 3.89% punctuation, but contains "Received" at beginning)	10024662_3_1082_99|Received on May 11, 1998; revised on July 21, 1998; accepted on July 21, 1998

Figure 4: Reference Span Designation Examples

3. Automatic answer extraction and passage retrieval

We explored two approaches to automatic answer extraction: 1) a method that involved describing the meaning of a passage of text in the form of extracted relations and entities; and 2) a pure information retrieval approach in which each passage was treated as a document and the retrieval results obtained using five search engines were merged using a method that consistently improved our results in the previous evaluations.

3.1 Knowledge-based approach

Our knowledge-based approach to question answering followed the Genomics track evaluation design: we identified exact answers in the form of relations, and then marked the passage as containing an answer.

With enough computing resources and a better understanding of question and answer types, this process could have been applied to the whole collection. Lacking both, we used a two-stage question answering design: we first identified "promising" passages using our in-house search engine Essie, and then submitted the passages to SemRep. SemRep results (consisting of the lists of identified entities and relations) were post-processed and combined with evidence from entity identification and Essie ranking to generate the final ranking of extracted passages. The final submitted passages consisted of the full-length legal spans trimmed to sub-spans containing relevant entities and relations.

3.1.1 Essie query generation

Essie queries consisted of three parts: 1) search terms taken from the topics, 2) expansions for search terms, where possible, and 3) broad weightings in favor of the topic answer types (domain query expansion).

Search Term Extraction: Terms were extracted from the topics using MetaMap and a list of model organisms obtained from the NCBI taxonomy of the organisms commonly used in molecular research (The NCBI Taxonomy).

Restrictions on semantic type, part of speech, and exception lists were used to rate the terms with respect to their usefulness for search as shown in Figure 5. The useful semantic types were determined based on the 2006 question templates and the 2007 training topics.

The part-of-speech inclusion/exclusion rules were generated based on the training topics. The exclusion lists contained the stop lists and the lists of terms too general to be useful in biomedical question answering, such as disease, or protein. The lists were compiled in the process of developing the CQA-1.0 clinical question answering system (Demner-Fushman and Lin, 2007).

Topic 200: What serum [PROTEINS] change expression in association with high disease activity in lupus?

Terms extracted from topic 200 and their attributes:						Domain query expansion for answer type [PROTEINS]		
Term	**SemTypes**	**STStrength**	**Rules**	**POS**	**Weight**	**Term**	**Quality**	**Weight**
high	qnco	weak	stop	n,j,r	zero	protein	A	0.5
lupus	dsyn	good		n	high	synthesis	B	0.3
change	ftcn	weak		n,v	low	biosynthesis	B	0.3
expression	genf	good	general	n	low	domain	C	0.2
serum	bdsu	good		n,j	high	amino	C	0.2
association	socb	weak	stop	n	zero	peptide	C	0.2
activity	dora	very low	general	n	low	c terminal	C	0.2
disease	dsyn	good	stop	n	zero	…	…	…

Essie query for topic 200:

```
( CONST[0.3] OR DOMAIN[PROTEINS] )
        AND ( CONST[0.05] OR lupus )
        AND ( CONST[0.05] OR serum )
        AND ( CONST[0.5]  OR change )
        AND ( CONST[0.5]  OR expression )
        AND ( CONST[0.5]  OR activity )
```

Figure 5: Essie query details for topic 200

Term Expansion: Where possible, search terms were expanded with synonyms. Essie synonymy expansion is limited to synonyms from the UMLS. For non-UMLS terms, such as model organisms, the expansion terms were obtained from the taxonomy of model organisms. Gene names identified in SemRep via MetaMap and AbGene (Tanabe & Wilbur, 2002) were expanded using the NCBI Gene database and EUtilities (Benson et al, 2000).

Answer Type Expansion: The Essie search engine is optimized to search for multi-word terms. Query expansion with inflectional variants and synonyms allows one to search for concepts. Further expansion with broader, narrower, and closely related terms produces what we call a domain search. For example, a domain search for proteins consists of a list of related terms such as {*protein*, *synthesis*, *biosynthesis*, *domain*, *amino*, *peptide*, etc}. Each term is given a small weight and their contributions are ORed together. A typical document about proteins will have many occurrences of these related terms, and the accumulation of many small weights will approach the maximum score of 1.0. Domain searches were used to represent the answer types, such as [PROTEINS], in the TREC 2007 topics. The only difficulty was in identifying broader, narrower, and closely related terms.

A future version of Essie will include a sentence parser and recognize all terms in the corpus. Until then, we use n-grams. As Essie indexes a corpus, it accumulates n-grams for all ***n*** up to 8. Many of these are too rare or otherwise uninteresting (e.g., starts or ends with a stop word, has unbalanced parentheses, includes a line break, etc.). The threshold for keeping an n-gram depends on available memory and the corpus size. For the TREC 2007 collection, the threshold was 64 occurrences of an n-gram. The remaining n-grams are collapsed with term

normalization (remove plurals, possessives, compound words, other hyphens, spelling variants, etc.), leaving a total of 1,250,391 different n-grams. An extra index was built to store which n-grams occur in each document and how many times.

A document to n-gram index allows finding all n-grams that occur in a search result (a list of documents). If an n-gram occurs more frequently in a search result than expected by random chance, there may be a relationship between the n-gram and the search term. For simplicity, we assume terms occur independently and follow Poisson statistics. This defines 1) an expected number of occurrences of any given n-gram in any given search result, and 2) a standard deviation of the random variation in the number of occurrences. If the observed number of occurrences is more than 3 standard deviations greater than expected, the search term and n-gram are unlikely to occur together by random chance. This simple approach has been shown to work by capturing many known relationships. For example, the chance that the n-gram *CHD* (coronary heart disease) is found by random chance in a search of *Heart Attack* is out at 160 standard deviations. *HOCl-LDL* (hypochlorite-oxidized LDL) is at 127 standard deviations, and *Heart Disease* is at 124. There are typically hundreds of n-grams with significance greater than 50 for broad search terms like the answer types. One drawback is that n-grams are not terms, and often include meaningless strings like *risk of heart* (part of *risk of heart disease* and *risk of heart attack*).

Using the above approach, Essie can produce a list of related n-grams for a given search term. This capability was used to suggest broader, narrower, and closely related terms for the answer types in the TREC 2007 training topics. Human review finalized these domain query expansions (an example is shown in Figure 5). The process was tedious and error prone, but possible with less than an hour of human effort per answer type.

Essie Queries: Past experience has taught us several tricks in creating effective Essie queries. Essie scores documents with a value, P, between zero and one, which can be loosely interpreted as a probability of relevance. An AND search of two terms (A AND B) will require both terms to be present and will score the result document with the joint probability of relevance, PA*PB. A naive approach is to AND together all search terms, but this is extremely restrictive. Documents missing any one of the search terms would be excluded. A finite penalty for missing search terms can be included by ORing in constant background scores. An OR search of two terms (A OR B) will require either term to be present and will score the result document with the probability, PA+PB-PA*PB. A 10-fold reduction in score if term A is missing and a 2-fold reduction for B can be represented by ((CONST[0.1] OR A) AND (CONST[0.5] OR B)). Term expansions were also included as OR clauses as in (CONST[0.1] OR A1 OR A2 OR A3). The domain query expansions are essentially large OR clauses and were treated similarly: (CONST[0.1]OR DOMAIN[PROTEINS]). An example of a final Essie query is shown in the bottom part of Figure 5.

Our goal was to find the 1,000 most relevant legal ranges for each topic. Filtering and ranking of these results was performed in the post-processing of retrieval results. Therefore our search strategy was recall-oriented.

Relation	**found in sentence:**
ASSOCIATED_WITH(Vascular Cell Adhesion Molecule-1, Lupus Nephritis)	Plasma **sVCAM-1** concentration is significantly elevated in patients with active **lupus nephritis** of WHO classes III and IV, and is decreased during remission.
Up-regulated expression of adhesion molecules on leucocytes and vascular endothelium leads to the adherence of inflammatory cells to the blood vessel wall and their subsequent extravasation [3, 23]. Soluble adhesion molecules have been detected in plasma and thus serve as useful markers of both leucocyte and endothelial cell activation in different diseases, such as autoimmune disorders, including rheumatoid arthritis, vasculitis and SLE [24-27]. A long-term study has revealed that elevated levels of soluble VCAM-1 (sVCAM-1), but not of soluble ICAM-1 (sICAM-1) and soluble E-selectin (sE-selectin), in SLE sera correlate positively with disease activity [28]. **Plasma sVCAM-1 concentration is significantly elevated in patients with active lupus nephritis of WHO classes III and IV, and is decreased during remission [28, 29].** These results suggest that sVCAM-1 may be a promising marker for monitoring patients with lupus nephritis.	

Figure 6: Finding a passage containing a part of an exact answer to question 200

3.1.2 SemRep post-processing

The goal of this processing was to identify passages containing not only the semantic types of an expected answer and those found in a question, but also the potential exact answers.

Identification of exact answers was based on the assumption that an answer can be expressed in the form of a relation between a query term and an instantiation of an answer type, as shown in Figure 6, where the semantic type of the term *Vascular Cell Adhesion Molecule-1* maps to the answer type PROTEINS, and the term *Lupus* is found in the query.

In addition to finding a relation between an answer type and a query term, we needed to define the relation types that potentially answer the questions. As the question types were not explicitly defined in the task, we approximated the question-answering relations using all "interesting" relations for a given answer type, whether the answer type was an object or a subject in a relation. These approximations were based on extensive use of SemRep in genomics, pharmacogenomics, and knowledge discovery (Ahlers et al, 2007). See Appendix A for the full list of answer and relation types.

In SemRep post-processing, each passage was scored based on the full credit given to appropriate relations containing an answer type and a query term, and partial credit given to appropriate relations containing either an answer type or a query term.

3.1.3 Combining evidence from exact answers, semantic filtering and passage relevance

The additional sources of evidence with respect to relevance of a given passage were sought because we anticipated that many extracted relations will point to approximate answers, and many useful relations are not yet defined in SemRep. For example, it remains to be seen if the judges consider the exact answer to question 200 in Figure 6 relevant, as the question contains an underspecified term that could have several interpretations, and the relation contains a specific disease term. However, expanding the answer to the whole passage we captured another sentence containing a potentially relevant relation between the *sera selectins* and *SLE*. Unfortunately, the relation captured by SemRep in this sentence was AFFECTS(Selectins, Disease), which is too general for an exact answer.

Semantic type filter: The semantic type filter checked that the passages retrieved by Essie contain the terms and semantic types mentioned in the query and the expected answer type. Using the training questions we derived a set of semantic types likely corresponding to each answer type (see Appendix 1). For example, the answer to the question "Which MUTATIONS are associated with xxxx" was expected to belong to one of the following semantic types: gene function (genf) or gene/genome (gngm). Sentences were identified in the passages and each sentence was assigned a score based on the coverage of the terms and semantic types mentioned in the query. The coverage was measured by counting the number of unique terms that were common between the query and a sentence. We also checked whether the sentence contained a term with the expected semantic type of the answer. The semantic matching was measured using a modified Jackard similarity measure:

$$Sem\_sim = \left(\frac{N\_common\_terms}{Qsize + Ssize - N\_common\_terms} \right) \times N\_answers$$

where *N_common_terms* is the number of the query terms that are covered in the sentence, *Qsize* is the total number of terms in the query, *Ssize* is the total number of terms in the sentence, and *N_answers* is the total number of possible answers (terms that have the semantic type of the expected answer) in the sentence. A paragraph could have one or more sentences and we selected the maximum score among all sentences in the paragraph.

The final score assigned to a passage was a linear combination of the normalized scores assigned to passages during semantic filtering with those generated in SemRep post-processing and assigned by Essie. The scores were normalized as follows:

$$norm(X) = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Based on the observations of combining the scores for training topics, a slightly higher weight (0.4) was given to scores assigned by SemRep. Equal weights (0.3) were given to the scores generated by the two remaining systems.

Examining retrieval results for training topics, we determined that, in general, reference sections tend to top the list because they contain all query terms and synonyms; however very few titles were actually relevant to the questions. Given SemRep results, we can pinpoint the relevant titles. Therefore in the final post-processing step, these titles were extracted from the reference sections as described below. The reference sections containing no useful relations were demoted to the bottom of the list.

3.1.4 Sub-span computation

In order to provide a more focused answer than an entire maximal legal span, we attempted to locate the answer sentences within the span. This was made difficult because the sentences had been stripped of special characters and XML syntax that is present in the maximal legal spans. For simplicity we just did an exact substring match for each of the sentences and formed the union of the discovered spans. If the sentence boundaries were not found exactly, we used the entire maximal legal span as the answer.

3.2 Information retrieval approach

The fusion approach described in our 2005 TREC paper was applied to merge the retrieval results obtained by Essie (see section 3.1.1), Indri (Ruiz et al, 2007), Terrier, Theme, and EasyIR (Gobeill et al, 2007). The original topics prepared for Essie queries were used in Terrier retrieval without expansion or any other further processing. For Terrier, the InL2 model was used with its default parameters. The description of Theme and Easy IR follows.

3.2.1 Theme

The NCBI contributed a run that was incorporated into the fusion run and also independently submitted through our UMD colleague (Lin et al, 2007). The run was based on an updated and simplified "theme query" process that we utilized in previous years.

To process a query topic, a program scanned the topic text to extract a list of features. For each feature, a probability was assigned to each TREC document, and the product of these feature probabilities was used for the final TREC ranking.

The TREC collection was processed by indexing on all words and bigrams using an in-house C++ class library that also tokenizes and removes stop words. The indexes enabled efficient retrieval and cross referencing of documents and the words and bigrams.

The features extracted from each topic consisted of all common nouns and verbs, all bigrams of adjectives and nouns, and the entity type itself (determined by matching the given names). Parts of speech were determined using the MedPost part-of-speech tagger (Smith et al., 2004).

The document score for each feature was determined by query expansion on MEDLINE®. For terms, i.e. words and bigrams, this was accomplished by querying MEDLINE (with the same indexing software) and using the result set to compute Bayesian weights for all MEDLINE terms. Those MEDLINE terms with weights greater than 9.64 were retained and applied to the TREC collection (Kim and Wilbur, 2005). That is, the score for a given term and a given TREC document was obtained by summing the retained Bayesian weights over all of the terms appearing in the TREC document.

In a similar way, a score was assigned to each document for each of the 12 given answer types excluding GENES and PROTEINS. For each answer type, a manually-constructed query was performed on MEDLINE, often using MeSH® headings and subheadings. For example, for the answer type CELL OR TISSUE TYPES, the entire MeSH hierarchy starting at A10 (Tissues) and A11 (Cells) was queried. The resulting set of MEDLINE documents was used to compute Bayesian weights for all possible MEDLINE terms, and these were used in the same way to compute scores for all TREC documents.

At this point, the raw scores were set to 0 for those documents that we determined were likely to correspond to whole bibliographies, as these contained a large number of relevant features in unrelated contexts. This included documents with more than 100,000 characters, and documents with more than 2,000 characters also having more than 1 in 15 characters as a period, colon, or digit. The raw scores were then converted to a probability with the formula $p = 1 / (1 + \exp(a z + b))$ where z is the raw score and a and b are chosen separately for each feature so that the top 10 scores received a probability of 0.99 and all documents containing at least one weighted term received a probability of 0.5.

The answer types GENES and PROTEINS were treated differently. We utilized a system that recognizes sentences containing gene names (following Tanabe and Wilbur, 2002), trained on the GENETAG collection (Tanabe et al., 2005). The value of the highest scoring sentence in a TREC document was taken as the score of the entire document for both the GENES and PROTEINS entity types. To map these scores to a probability, whole bibliographies were removed as before and the same formula was applied taking $a = -1$ and $b = 10$ to achieve a shallow, fuzzy cutoff near a raw score of 10.

3.2.2 EasyIR

Bibliographical sections were a priori removed from the indexed collection, as well as very short text passages (< 12 characters). A subset of the original collection (about 800,000 passages) was selected based on an automatic run generated by Indri (Ruiz et al, 2007) using the original queries. Porter stemming was applied on this collection with a specific

handling of hyphens. Thus, the following 3-word expression a-b-c was expanded into the following set of words {abc, ab, bc, a, b, c}. A specific pivoted normalization (dtu.dtn) weighting formula (Gobeill et al., 2007) was used at retrieval time with slope = 13. The original DF (document frequency) of the collection was mixed with a DF list computed on the whole MEDLINE collection.

3.3 Adding document relevance

To find documents potentially containing answers to the test topics, one of the researchers built manual PubMed queries to search MEDLINE. The set of retrieved documents was used to re-order the passage ranks obtained in the fusion run, promoting passages from documents retrieved using PubMed and deemed relevant. This process is illustrated by the query in Figure 7, which retrieved two MEDLINE documents that were also in the test collection. Simply by virtue of this retrieval (without examining the full documents) scores for the passages from these two documents that were also retrieved by the Essie search were boosted. For instance, the best-ranked Essie passage 14693703_3813_2415 was boosted from a rank of 34 to a rank of 4. It appears to be relevant to the query based primarily on the sentence, "Caspase-1 is important in the regulation of IFN production induced by lipopolysaccharide (LPS)-stimulated secretion of IL-18 (2)." and the fact that NOD (non-obese diabetic) mice were referenced twice elsewhere in this passage.

Query entered into PubMed to retrieve MEDLINE documents for question 227, "What [GENES] are induced by LPS in diabetic mice?":	(lipopolysaccharides OR lps) AND diabetes mellitus[mh] AND mice[mh] AND (gene OR genes OR ge[sh]) AND (free full text[sb]).
PubMed translation of this query:	(("lipopolysaccharides"[MeSH Terms] OR lipopolysaccharides[Text Word]) OR lps[All Fields]) AND "diabetes mellitus"[MeSH Terms] AND "mice"[MeSH Terms] AND ((("genes"[TIAB] NOT Medline[SB]) OR "genes"[MeSH Terms] OR gene[Text Word]) OR ("genes"[MeSH Terms] OR genes[Text Word]) OR "genetics"[Subheading]) AND "loattrfree full text"[sb]

Figure 7: Sample manual query to retrieve relevant documents used for improving passage scores.

4. Results

The performance of our base systems was not as uniform as in previous years (see Table 1). The significance in the differences in performance of our runs reported below was measured using Wilcoxon's Signed-rank Test.

Table 1: Average precision of the automatic base runs, knowledge-based runs, fusion and interactive runs. (Runs above the official mean in bold)

System	Average Precision		
	Document	Passage2	Aspect
EasyIR	0.0619	0.0133	0.0222
Essie	**0.2327**	**0.0698**	**0.2249**
Indri	**0.2209**	**0.0698**	**0.1790**
Terrier	**0.3008**	**0.0922**	**0.2493**
Theme	0.0568	0.0110	0.0552
SemanticFilter	0.0948	0.0137	0.0834
SemRepFilter	**0.1898**	**0.0470**	**0.1526**
LHNCBC	**0.2266**	**0.0679**	**0.2029**
NLMFusion	**0.3105**	**0.1097**	**0.2494**
NLMInter	**0.3286**	**0.1148**	**0.2631**

Focusing on the automatic runs, the fusion run (NLMFusion) significantly outperformed all but the Terrier base run on all evaluation levels. Our knowledge-based run (LHNCBC) did not outperform either its retrieval step (Essie), or the fusion run. Although both semantic post-processing steps performed significantly worse than the retrieval results on which they were based, there is no statistically significant difference between the combined run (LHNCBC) and the base run (Essie). Contrary to our expectations, our fusion run significantly outperformed the knowledge-based run on document, aspect and passage levels.

The exclusion of the reference sections of the full documents did not influence the results significantly. For example, document MAP of the Essie run with references was 0.2311, which is not significantly lower than the run without the references (Table 1).

Trimming of the passages also did not influence the results: scores for trimmed and untrimmed runs differ in the fourth digit after the decimal point (using both passage MAP metrics).

Despite the fact that our interactive run (NLMInter) amounted to a limited relevance feedback in the form of a list of documents potentially relevant to the question, its results are significantly better than the underlying fusion run on the document ($p < 0.001$) and passage levels ($p<0.05$), but not on the aspect level.

5. Conclusions

Our results suggest that for the tasks requiring identification of passages of text potentially relevant to biomedical questions, pure information retrieval approaches are sufficient. Adding knowledge-based methods (when extraction of entities and relations to answer the question is not required) does not improve the results.

Although our fusion approach does not significantly outperform one of the contributing base systems this year, it still reliably maintains an acceptable level of performance.

The improvements in retrieval due to relevance feedback were to be expected. However, it is worth noting that the knowledge about relevance of a document determined by virtue of its retrieval in an expert PubMed search and an examination of its abstract is sufficient to improve passage retrieval.

Acknowledgments

We would like to thank Tom Rindflesch, Marcelo Fiszman, Caroline Ahlers, and Halil Kilicoglu for their help with our knowledge-based approach.

References

1. Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting Semantic Predications from MEDLINE Citations on Pharmacogenomics. Pacific Symposium on Biocomputing 2007;12:209-220
2. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA 2001;17-21.
3. Aronson AR, Demner-Fushman D, Humphrey SM, Lin J, Liu H, Ruch P, Ruiz ME, Smith LH, Tanabe LK, Wilbur JW. Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. The Fourteenth Text Retrieval Conference, TREC-2005, Gaithersburg, MD.
4. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Res 2000 Jan 1;28(1):15-18.
5. Demner-Fushman D, Lin J. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. Computational Linguistics. 2007;33(1):63-104.
6. Fox EA, Shaw JA. Combination of multiple searches. Proceedings TREC-2. 1994;243-249.
7. Gobeill J, Ehrler F, Tbahriti I, Ruch P. Vocabulary-driven Passage Retrieval for Question-Answering in Genomics. The Sixteenth Text Retrieval Conference, TREC-2007, Gaithersburg, MD. (to appear)
8. Ide NC, Loane RF, Demner-Fushman D. Essie: A Concept Based Search Engine for Structured Biomedical Text. J Am Med Inform Assoc. 2007 May-June;14(3):253-263.
9. Kim W, Wilbur WJ. A Strategy for Assigning New Concepts in the MEDLINE Database. Proc AMIA Symp. 2005
10. Lin J. et al. The Sixteenth Text Retrieval Conference, TREC-2007, Gaithersburg, MD. (to appear)
11. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91.
12. Metzler D, Croft WB. Combining the Language Model and Inference Network Approaches to Retrieval. Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval. 2004;40(5):735-750.
13. Ounis I, Amati G, Plachouras V, He B, Macdonald C, Lioma C. Terrier: A High Performance and Scalable Information Retrieval Platform. In Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006). 10th August, 2006. Seattle, Washington, USA.
14. Rindflesch TC, Fiszman M. The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text. Journal of Biomedical Informatics. 2003;36(6):462-77.
15. Ruiz ME et al. The Sixteenth Text Retrieval Conference, TREC-2007, Gaithersburg, MD. (to appear)
16. Ruch P, Ehrler F, Gobeill J, Tbahriti I. Report on the TREC 2006 Experiment: Genomics Track." The Fifteenth Text ReTREC-2006, Gaithersburg, MD.

17. Smith L, Rindflesch T, Wilbur WJ. MedPost: a part-of-speech tagger for bioMedical text. Bioinformatics. 2004 Sep 22;20(14):2320.

18. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics. Aug 2002;18: 1124 – 1132.

19. Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ. GENETAG: a Tagged Corpus for Gene/Protein Named Entity Recognition. BMC Bioinformatics 6:Supp. 1, 2005.

20. The NCBI Taxonomy [Internet]. Bethesda (MD): National Library of Medicine (US). [date unknown]-[cited 2007 Oct 16]. Available from: http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/

21. Wilbur WJ. A thematic analysis of the AIDS literature. Pac Symp Biocomput. 2002:386-97

Appendix A: Entity types and relations allowed for a given answer type

Answer type	Semantic types	Relations
ANTIBODIES:	aapp, imft, rcpt, irda, moft, bacs	INTERACTS_WITH, INHIBITS, STIMULATES, AFFECTS, DISRUPTS, AUGMENTS
BIOLOGICAL SUBSTANCES:	bacs, aapp, imft, opco, horm, vita, mbrt, lbpr, eico, enzy, carb, lipd, nsba, orch, strd, nnon	INTERACTS_WITH, INHIBITS, STIMULATES, AFFECTS, DISRUPTS, AUGMENTS
PROTEINS:	aapp, bacs, biof	INTERACTS_WITH, INHIBITS, STIMULATES, AFFECTS, DISRUPTS, AUGMENTS, ASSOCIATED_WITH
CELL OR TISSUE TYPES:	bpoc, cell, celc, ffas, tisu, emst, anst	LOCATION_OF, PART_OF
DISEASES:	neop, mobd, dsyn, patf, acab, anab, cgab, inpo, sosy, comd	COEXISTS_WITH, ASSOCIATED_WITH, PREDISPOSES, CAUSES, PROCESS_OF, LOCATION_OF, AFFECTS
SIGNS OR SYMPTOMS:	sosy, patf, fndg, clna, mobd	COEXISTS_WITH, ASSOCIATED_WITH, CAUSES, AFFECTS, PROCESS_OF, PART_OF
DRUGS:	antb, phsu, horm, vita, orch, aapp, hops, strd, nnon	TREATS, ASSOCIATED_WITH, INTERACTS_WITH, PREDISPOSES, CAUSES
GENES:	aapp, genf, gngm, biof, celc, nnon, bacs, celf, rcpt, nusq	ASSOCIATED_WITH, PREDISPOSES, CAUSES
MUTATIONS:	genf, gngm, orga	ASSOCIATED_WITH, PREDISPOSES, CAUSES, PROCESS_OF, PART_OF
PATHWAYS:	celf, biof, orgf, enzy, aapp	ASSOCIATED_WITH, PREDISPOSES, CAUSES, AFFECTS, PROCESS_OF
TOXICITIES:	inpo, hops, sosy, comd, mobd, fndg	ASSOCIATED_WITH, PREDISPOSES, CAUSES
TUMOR TYPES	neop, patf, hops	ASSOCIATED_WITH, PREDISPOSES, CAUSES, LOCATION_OF, PROCESS_OF, AFFECTS
MOLECULAR FUNCTION:	biof, celf, comd, moft, orgf, ortf, phsf, genf, patf	PROCESS_OF, AFFECTS, DISRUPTS, AUGMENTS, ASSOCIATED_WITH, INHIBITS, STIMULATES
STRAINS:	inpr, bact, virs	PROCESS_OF, PART_OF, AFFECTS, DISRUPTS, AUGMENTS, PREDISPOSES, CAUSES

Answer semantic types are provided as four-letter abbreviations. Full names available at: http://semanticnetwork.nlm.nih.gov/Download/RelationalFiles/SRDEF