Finding relevant passages in scientific articles:

fusion of automatic approaches vs. an interactive team effort

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

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

[b]University of Maryland, College Park, Maryland

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

[d]State University of New York at Buffalo, Buffalo, New York
meruiz@buffalo.edu

Abstract

This paper presents our approach to retargeting the information retrieval systems designed and/or optimized for retrieval of MEDLINE citations to the task of finding relevant passages in the text of scientific articles. To continue using our TREC 2005 fusion approach, we needed a common representation for the full text biomedical articles to be shared by the four base systems (Essie, SMART, EasyIR and Theme.) The base systems relied upon previously developed NLP, statistical and ML methods. The fusion of the automatic runs improved the results of three contributing base systems at 99.9% significance level on all metrics: document, passage, and aspect precision. The fusion run outperformed Essie, the best of the base systems, at 94% to 99% significance level, with the exception of passage precision.

The novelty of the task and the lack of training data inspired our exploration of an interactive approach. The prohibitive cost (in time and domain expert effort) required for a truly interactive retrieval led to a team interaction with one of the base systems – Essie. The initial queries were developed by a computational biologist and a medical librarian. The librarian merged and then refined the queries upon inspecting the initial search results. The refined queries were submitted as a batch without further interaction with the system. The interactive results, the best we achieved, improved over the baseline automatic Essie run at the 91% significance level. The difference between the fusion and the interactive results is not statistically significant.

Keywords: Genomics; MEDLINE/PubMed; MeSH; Interactive Information Retrieval; Vector Space Models; Statistical Natural Language Processing; Machine Learning; Thematic Analysis.

1 Introduction

The single task of the TREC genomics track 2006 focused on finding passages containing information relevant to topics expressed as questions. The task posed several challenges for our base information retrieval systems optimized for indexing and retrieval of MEDLINE® citations. The documents for this task were scientific articles in HTML format. The expected results were passages -- spans of text that cannot cross an HTML paragraph boundary. As it was not clear if finding the best documents and then the relevant passages would be better than finding the spans with the high density of relevant terms first, a hierarchical XML representation of each citation ranging from sentences to documents was created (see Section 2.1). After experimenting with the two sample topics, we decided to use the maximum-length legal spans (the passages delimited by the HTML paragraph tags) as a retrieval unit. The spans retrieved by the base systems were merged using the sum fusion method (Fox and Shaw, 1994) described in Section 2.3.

Observing that the results obtained automatically for two sample topics were not satisfactory, we decided to experiment with manual and interactive query refinement, in addition to automatic query generation. Our approach to interactive retrieval is described in Section 3. To our surprise, the difference in document average precision between the fusion of automatic runs and the interactive run is not statistically significant. The difference between these

runs in passage and aspect precision is significant only at the 90% level. Section 4 provides a detailed description of our results. Section 5 contains some conclusions about the work.

2. Automatic Retrieval

In the automatic fusion run we combined the retrieval results of four systems (Essie, EasyIR, SMART, and Theme) each of which is known to perform well for some IR tasks. Our fusion approach consisted of normalizing the scores from each system on a query by query basis and then using these normalized scores to compute a new combined score for the union of all results returned by all four systems. The top 1,000 results were selected based on the combined score.

2.1 Document preparation

The Highwire collection was processed to convert the HTML source to XML coded paragraph text. The Perl programming language was used to manually construct conditions (discovered by inspection and sampling of the collection) in order to remove non-displayed HTML tags, record internal section names, recognize titles, headers, and likely text, recognize and omit bibliographic sections, and convert escape symbols and glyph images to approximate ASCII equivalents. The result was then coded in UTF-8 and saved in a hierarchical XML structure representing the document, internally defined sections, paragraphs (as determined by the <P> tags), and titles, headers, and paragraphs.

```
<DOC FILE="file-name">
  <SECTION NUMBER="num" NAME="name">
    <PARAGRAPH NUMBER="num">
      <TITLE ID="id" NUMBER="num"
        OFFSET="offset" LENGTH="len">…</TITLE>
      <HEADER1 ID="id" NUMBER="num"
        OFFSET="offset" LENGTH="len">…</HEADER1>
      <HEADER2 ID="id" NUMBER="num"
        OFFSET="offset" LENGTH="len">…</HEADER2>
      <PARAGRAPH ID="id" NUMBER="num"
        OFFSET="offset" LENGTH="len">…</PARAGRAPH>
    </PARAGRAPH>
  </SECTION>
</DOC>
```

2.2 Basic automatic approaches

2.2.1 Essie

Essie is a concept-based search engine for structured biomedical text. Searches use token adjacency indexes to find sequences of tokens (a phrase search instead of just a word search.) Fine-grained tokenization treats every punctuation mark as significant.

Queries are automatically expanded before search. UMLS® terms are recognized and expanded with their synonyms. Then the original term and its synonyms are both automatically expanded with a conservative set of word variants from the SPECIALIST Lexicon: plurals, possessives, hyphenation, compound words, and alternative spellings. Stemming and inflection (other than plural/singular) are not used.

A relaxation strategy breaks long phrases into sub-phrases and individual words. Each piece is expanded with synonymy and word variants. All combinations of pieces that span the original query are evaluated. Thus a search for "heart attack in older adults" will find "myocardial infarction in seniors". A lossy strategy allows some fragments to be missing.

Essie ranks results with an "all the right stuff in all the right places" strategy. Terms generated through relaxation, synonymy, and word variants are discounted relative to the original query. Terms found in high value regions of a document, such as a title, are weighted more than those found in low value regions, such as addendums or footnotes. This latter strategy has been shown to be very effective but relies on a structured XML document format that is not always available.

Essie was first evaluated in the TREC 2003 Genomics Track, where it produced the high score of 0.42 mean average precision (MAP). Essie is in production use as the search engine for ClinicalTrials.gov.

Data. The TREC 2006 corpus consisted of full text HTML journal articles. The HTML was converted to XML with a sentence parser which inserted tags at sentence boundaries, paragraph boundaries, and around the full text as described in Section 2.1. There was some error in this process, since not all sentences fell within the legal ranges provided later.

Essie was configured to treat paragraph elements as documents. Searches found relevant paragraphs, which were then trimmed as described in the *Passages* section below to produce relevant passages. Two preliminary trimming strategies, 1) find relevant sentences and merge to form passages and 2) find relevant full text documents and trim to form passages, were considered and abandoned.

Queries. The automatic query formulation took advantage of two factors: 1) Essie query expansion capabilities, especially based on the UMLS-derived synonymy, and 2) the structure of the Generic Topic

Types (GTTs) underlying the genomics track topics that allows for in-depth semantic analysis.

The topics for the 2006 genomics track are questions based on the Generic Topic Types (GTTs). The GTTs contain: 1) one or more biological objects (genes, proteins, gene mutations, etc.), 2) one or more biological processes (physiological processes or diseases), and 3) a relationship between the objects and the processes. The first two components of GTTs were identified automatically using named entity extraction tools. All topics were processed using MetaMap (Aronson, 2001) to identify biological processes and objects. Based on the sample queries and template descriptions, the following semantic types were manually categorized as "related to processes": all children of the UMLS semantic type *Event* (for example, *Disease or Syndrome*, *Neoplastic Process*, *Cell Function*, etc.), children of the semantic type *Anatomical Structure* excluding *Gene or Genome* (for example, *Cell*, *Cell Component*, *Body Part*, *Organ, or Organ Component*, etc.), and children of the semantic type *Substance*, excluding *Amino Acid, Peptide, or Protein*, and Nucleic *Acid, Nucleoside, or Nucleotide*. Terms extracted from the topics using this strategy were dubbed ***clinical***.

Extraction of "biological objects" was treated as gene and protein name recognition. In addition to the UMLS semantic types *Gene or Genome*, *Amino Acid, Peptide, or Protein*, and *Nucleic Acid, Nucleoside, or Nucleotide*, gene names were recognized in the topics using ABGene (Tanabe and Wilbur, 2002), and then expanded automatically using the Entrez Gene[1] synonymy. Terms extracted from the topics using this strategy were named ***gene***. A third group of query terms, named ***other*** was identified using a heuristic for recognizing potential interesting terms missed by the first two tools. According to this rule all words containing upper and lower case, numbers and/or punctuation were extracted, if not present in the other two groups.

No question understanding or recognition of relations was attempted. To use the potentially important terms identifying a relationship, the whole original question was added to each query.

The final queries were assembled as a union of the original question with the intersection of three groups of the extracted terms (terms within groups were ORed): (CLINICAL_TERM(s) AND GENE(s) AND OTHER) OR QUERY. To retrieve documents containing some but not all terms, each term was ORed with a small constant weight. Figure 1 presents the query automatically generated for topic 172.

CLIP[0.00011,1.0] ((CONST[0.01] OR apoptosis) **AND** (CONST[0.01] OR p53 OR WEIGHT[0.4] TERM_EXPAND (CG10873 OR CG31325 OR CG33336 OR DmP53 OR dp53 OR prac))) **OR** WEIGHT[0.0001] LOSSY_EXPAND (How does p53 affect apoptosis)

Figure 1. Automatic query for Essie search. Groups of query terms are combined using operators shown in bold.

In Figure 1, a ***clinical*** term *apoptosis* identified as *Cell Function* is ANDed with the ***gene*** *p53*. Terms identified as UMLS concepts are not expanded in the queries because of Essie's built-in morphologic and UMLS derived expansion. When no expansion type is indicated, the concept based expansion is applied by default. Synonyms found using Entrez Gene (*CG10873*, *CG31325*, *CG33336*, *DmP53*, *dp53*, and *prac*) are ORed with the original query terms. A term expansion operator that expands a term using its morphological variants was applied to synonyms. Lossy expansion described above is applied to the question. Because lossy expansion results might be quite noisy, a small weight assigned to this part of the query ensures that documents containing all extracted terms are promoted to the top of the ranked list.

In all topics, only three terms were identified as ***other***: *Nurr-77*, *HPV11*, and *GFs*. For the most part the ***clinical*** and the ***gene*** terms were identified correctly, however in several cases the terms were misplaced due to the UMLS ambiguity. For example, *APC* and *adenomatous polyposis coli* were assigned to different groups, *APC* was recognized as ***gene***, but *adenomatous polyposis coli* was mapped to *Neoplastic Process*; *L2* was recognized by MetaMap as *Body Location or Region*, but *L1* was recognized as ***gene***.

Manual Retrieval. The placement of the above mis-tagged terms was corrected in the query formulation for our manual run. In addition, queries containing the term *Nurr-77* were augmented with the term *Nur-77*. The remaining manual queries do not differ from the automatic queries.

Passages. Essie produced a ranked set of paragraphs which were trimmed to produce passage candidates. The trimming process maintained the relevancy ranking, but dropped problematic paragraphs.

Due to inconsistent HTML tagging and inaccessible document structure, some highly ranked paragraphs

[1]http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

were inappropriate. Paragraphs were found with more than 50 sentences. Inspection revealed that some of these were miss-tagged unions of many paragraphs, and some were reference sections. References have many interesting terms as part of citation titles and are therefore ranked highly in the Essie search results. Any paragraph with over 30 sentences was dropped. Paragraphs were also found with a single sentence, more than 1,000 characters long. This case was caused by tables and other data structures tagged as paragraphs. Paragraphs with a single sentence longer than 500 characters were dropped. Note that paragraphs with one or two sentences are often titles and captions, and are not necessarily bad passages.

The remaining paragraphs consist of sentences, some of which have query terms in them. Leading and trailing sentences without query terms were dropped. If the remaining sentences did not contain gaps, defined as a series of 4 sentences without query terms, they were kept as a passage. When gaps were present, the sentences were divided at the gaps into multiple candidate passages. Again, leading and trailing sentences without query terms were dropped. If one of the candidates was clearly the best, containing at least two more query terms than any other, it was kept as the passage. Otherwise the paragraph was abandoned and no passage extracted.

2.2.2 SMART

This year the SMART run involved indexing each of the paragraphs with SMART (Salton, 1971), using the pre-parsed XML topics described in Section 2.1 above. See (Ruiz, 2006) for details. Due to the size of the indexes produced by SMART, we split the collection into three subsets that would generate indexes less than 2GB in size. This meant that we had to create three sub-collections (A, B, and C) and submit the queries to each of the sub-collections. The final SMART run was generated by merging the results obtained from the three sub-collections. This merging process used the same algorithm as the general fusion code created to combine runs from all participating systems in the NLM fusion run.

We use the pivoted length normalization (*Lnu.ltu*) weighting scheme (Singhal et al., 1996) with slope = 0.25 and pivot = average document term in each sub-collection (where pivot_a = 32.435, pivot_b = 30.4005, and pivoc_c = 33.4440). Finally, we expanded the original query with terms generated by MetaMap and the Theme method.

2.2.3 EasyIR

The official topics of the 2006 Genomics track are a subset of the official 2005 topics. Topics from 2005, which were not in the test set this year, were used to tune the parameters of the EasyIR engine. See (Ruch et al., 2006) for details.

Indexes were generated based on a pre-processed document collection (cf. 2.1. Document preparation). Tuning was performed on the collection of abstracts. The best weighting was obtained using a slightly modified *dtu.dtn* schema (Singhal, 2001; Ruch et al., 2004), with slope = 30 and using a modified Porter stemmer.

While pivoted normalization (and related length normalization factors) has shown some effectiveness in previous TREC Genomics experiments, probably due to the bi-Gaussian distribution of document length in MEDLINE (cf. Ruch et al., 2005), the simple Gaussian distribution of document length in the 2006 collection may favor simpler normalization strategies (cosine, Boolean, etc.)

2.2.4 Theme

This approach was virtually identical to the Theme approach described in our TREC 2005 paper (Aronson et al., 2005), and readers may refer to that document for details. There were, however, minor differences in the collection and the format of query topics.

Each query was tagged using ABGene and MetaMap to identify gene names and known medical theme phrases. Gene synonyms were obtained from Entrez Gene and other synonyms were taken from the MetaMap output and obvious singular/plural variants. All remaining phrases, excepting stop words, were lumped together with the theme phrases. All phrases were queried separately, expanded using the theme-based query expansion, and the results from query synonyms were combined with OR. All of the phrases were then combined with AND. Finally, the result was rescored to prefer paragraphs containing the explicit gene name (if one was found), and paragraphs mentioning genes and proteins generally.

As an example, consider topic 161 “What is the role of IDE in Alzheimer's disease?”. The gene “IDE” was found by ABGene and was expanded to the synonym “insulysin”. The MetaMap phrase “Alzheimer’s disease” was found with synonyms “alzheimers disease”, “alzheimer’s diseases” and “alzheimers disease”. The resulting query can be described as

```
(Query("ide" OR "insulysin")
        AND1
        Expand(Query("alzheimer's disease" OR "alzheimers disease"
                        OR "alzheimer's diseases")
                OR1
                (Expand("alzheimers") AND1 Expand("diseases"))))
AND2 (Query("IDE") OR2 0.5)
AND2 Expand("gene" OR "genes" OR "protein" OR "proteins")
```

The Query function returns a fuzzy set in which every document scores 1 if it is in the query result, and 0 otherwise. The Expand function returns a fuzzy set that results from performing the query followed by query expansion. In this process, the original query output is guaranteed to appear in the result with a score of 1. In this query, the AND/OR operations with subscript 1 correspond to the min/max fuzzy set operators, and the subscript 2 correspond to multiplicative fuzzy set operators. Non-subscripted operators within queries are Boolean, in the usual sense. The AND_2 operators effectively rescore the result as described, and the 0.5 is the fuzzy set in which every document has score 0.5.

2.3. Fusion of automatic approaches

The fusion approach was designed to merge paragraphs retrieved from the four systems described in the previous section that comprised the automatic retrieval effort at NLM:

- Essie, a search engine developed specifically for biomedical text supporting flexible query expansion;
- SMART, the traditional retrieval engine with the *Lnu.ltu* weighting scheme;
- EasyIR, a method using term and document weightings as well as pivoted normalization; and
- Theme, a method that performs selective query expansion based on theme analysis.

The fusion formula that we used this year is similar to the one we used for our fusion runs last year except that the fusion was performed at the paragraph level. Each system returned a ranked list of paragraphs with the same format as the official genomic runs (including document ID, paragraph offset, and paragraph length).

Since all systems participating in the fusion runs were using the same set of paragraphs we did not have to deal with merging partially overlapping segments. The fusion program did verify that the generated paragraphs were contained within the maximum-length legal spans (no span could cross paragraph tags). In doing this we discovered a problem with our XML processing which involved crossing the legal span boundaries. To solve the problem, the fusion program resized the returned paragraphs by discarding material that crossed the paragraph tags. This is probably not the best solution, but it was the easiest way to solve the problem.

The fusion program normalizes and merges the scores of each system according to the following equation:

$$rsv_{fusion}(i) = \sum_{s \in S} \lambda_s \frac{rsv_s(i) - \min_s}{\max_s - \min_s}$$

where $rsv_{fusion}(i)$ represents the final score assigned to paragraph i, S is the set of systems participating in the fusion, $\min_s$ and $\max_s$ are the minimum and maximum scores reported by system s, and λ_s is a factor that weights the contribution of system s to the final fusion run. This year, all participating systems were assigned the same weight for their contribution to the fusion run.

3. Interactive Retrieval

The initial queries for our interactive run were developed independently by a computational biologist and a medical librarian based on their combined domain expertise and knowledge of the literature. The queries were subsequently merged and refined by the librarian based on interaction with Essie.

The Essie search engine accepts queries in the form of Boolean searches. The following is an example of the manual translation of a query into an Essie search:

Query 165:
How do Cathepsin D (CTSD) and apolipoprotein E (ApoE) interactions contribute to Alzheimer's disease?

Essie search for query 165:
(apoe OR apoproteins-e OR ad2) AND alzheimer AND ("cathepsin d" OR cpsd)

Synonyms of concepts in the queries came from regular MeSH® descriptor records and MeSH supplementary concept records (SCRs), text word variants of query concepts, Entrez Gene, and synonyms/variants built into Essie. These built-in Essie synonyms/variants greatly streamlined the

searches by making it unnecessary to add some of the MeSH/Entrez Gene synonyms, as well as numerous variants, to queries.

For example:

- alpha7nachr (query 173 "How do alpha7 nicotinic receptor subunits affect ethanol metabolism?") is a synonym in MeSH SCR for alpha-bungarotoxin receptor.
- Synonyms for colon cancer (query 163 "What is the role of APC (adenomatous polyposis coli) in colon cancer?") were expressed as the text word Boolean strategy (colon OR colonic OR colorectal) AND (neoplasm OR neoplastic OR tumoral OR carcinoma OR carcinogenesis); Essie had further built-in synonyms/variants for the cancer concept, e.g., malignancy, which therefore did not need to be specified in the search.
- Entrez Gene contributed the synonym tp53 for p53 (query 172 "How does p53 affect apoptosis?").
- Essie already knew several synonyms/variants for TGF-beta1 (query 166 "What is the role of Transforming growth factor-beta1 (TGF-beta1) in cerebral amyloid angiopathy (CAA)?" thus allowing their removal and reducing the TGF-beta1 parameter to (trfb1 OR TGF-beta1 OR dpd1).

Searches were refined based on Essie retrievals, specifically when retrievals would be too small or when synonyms would likely result in many irrelevant passages or obviously have meanings unrelated to the query. This includes using a feature of Essie which suppresses specified terms that would normally be included in Essie term expansion.

For example:

- The CFTR parameter (query 170 "How does COP2 contribute to CFTR export from the endoplasmic reticulum?") was eliminated because intersecting the COP2 terms with this parameter resulted in 0 documents; thus, it was decided to intersect the COP2 parameter with the endoplasmic reticulum parameter, eliminating CFTR as a third parameter.
- The Entrez Gene synonym n10 for Nurr-77 (query 171 "How does Nurr-77 delete T cells before they migrate to the spleen or lymph nodes and how does this impact autoimmunity?") was removed because of retrieval of passages containing "N 10" where this refers to the number of subjects in a study numbering 10.
- Essie normally expands the BSE synonym for mad cow disease (query 160 "What is the role of PrnP in mad cow disease?") to include breast self-examination, but expansion to this obviously undesirable synonym was suppressed by an "exclude" feature in term expansion by Essie.

The finalized search strategies were submitted to Essie. The results of the interactive query formulation were trimmed automatically and submitted without further inspection or manual processing.

4. Results

The overall performance of the base automatic systems either slightly exceeds or approaches the median (see Table 1 and Figure 2).

Table 1. Average precision of the base automatic, fusion, manual and interactive runs.

System	Average Precision		
	document	**passage**	**aspect**
Median (auto)	0.279	0.024	0.117
Essie	0.342	0.041	0.192
EasyIR	0.277	0.030	0.167
SMART	0.278	0.035	0.159
Theme	0.225	0.019	0.122
Fusion	**0.379**	**0.047**	**0.262**
Median (manual)	0.288	0.028	0.139
Manual	0.365	0.047	0.266
Interactive	**0.473**	**0.083**	**0.405**

The fusion run significantly outperformed the contributing base runs (see Table 2).

Table 2. Improvements of the fusion and interactive runs over the baseline systems. Statistical significance evaluated using the Wilcoxon signed ranks test**.**

	P <		
Systems	**doc.**	**passage**	**aspect**
Essie -- Fusion	0.06	0.17	0.02
EasyIR -- Fusion	0.0007	0.0003	0.0008
SMART -- Fusion	0.0004	0.008	0.0001
Theme -- Fusion	0.0002	0.0003	0.002
Manual -- Fusion	0.2	1	0.6
Interactive-Fusion	0.6	0.1	0.1
Essie -- Manual	0.08	0.001	0.07
Essie - Interactive	0.09	0.002	0.002
Manual-Interactive	0.06	0.03	0.01

The manual and the interactive runs also exceed median performance, with the interactive run being the best or very close to best on seven topics. The interactive run outperformed the manual run at 94% to 99% significance level.

Although the manual and the interactive results obtained using Essie are significantly better than the baseline run for this system, the contribution of the other three systems to the fusion run made the differences between the manual run and fusion and between the fusion and the interactive results not statistically significant (see Figure 2).

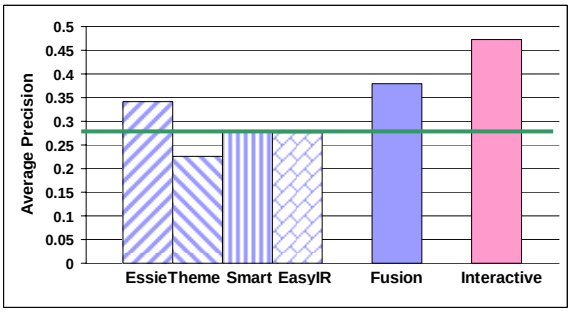

Figure 2. Document Average Precision of the base systems, fusion results, and interactive retrieval. The horizontal line represents median document AP for automatic runs (0.279)

Figure 3 shows the difference in the results of the interactive and the fusion runs. For a few topics the interactive and fusion results are similar, but for 17 topics the difference in document AP is 0.1 or more, with seven topics for which fusion results outperform interactive and ten topics for which the interactive results are better.

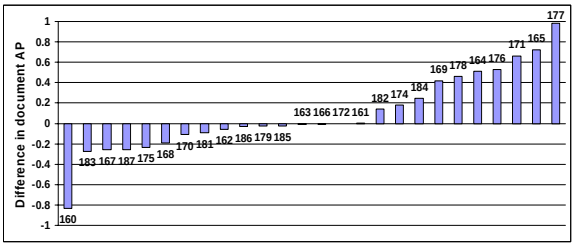

Figure 3. Difference between the interactive and fusion results on the test topics.

The poor performance of the interactive run for topic 160 (What is the role of PrnP in mad cow disease?) can be attributed to the absence of gene synonyms in the final query. The automatically constructed query that includes gene synonyms achieved 0.8 precision in the automatic Essie run, whereas precision is 0.08 for the interactive run. For topic 183, however, the unexpanded automatic query achieved precision 0.3, but the precision of the expanded interactive query was only 0.06, probably because the over-expanded "clinical" part of the query skewed the results. The same reasoning might be appropriate for topic 187: the term "hippocampal" was not identified by any of the automatic methods, and is therefore missing in the automatic query. Adding the term to the interactive query brought its precision down to 0.33 from the original 1. Topic 168 also is over-expanded in the interactive query. For the remaining topics in which fusion outperforms the interactive run, the success can be attributed to fusion itself.

Dropping of the leading and trailing sentences without query terms was clearly beneficial for the passage AP: precision improved from 0.03 to 0.04 for Essie automatic run. Although our passage retrieval seems to be above the median, the median itself is so low that we are probably one of the many groups that need to improve relevant passage identification in the future.

The aspects that reflect the breadth of coverage of the topic by a system were assigned to the passages during the evaluation. As it was not clear what aspects will be of a particular interest to the users, no special treatment of the potential aspects was undertaken. Any success in our aspect retrieval can be attributed to search for relevant passages.

5. Conclusions

Our results confirm our previous findings, showing that the fusion approach represents a significant improvement over the base runs and over the median even though some of the contributing runs used in the fusion were not significantly better than the median. The comparison of the interactive results with the baseline performance and with fusion shows that over-expansion and/or unbalanced expansion of a query degrades the performance. This of course could be quickly corrected in a real interactive retrieval.

It is quite interesting that the fusion approach seems to be able to *mimic* the improvement in results obtained through user intervention. This suggests that the different automatic approaches highlight different shades of domain knowledge.

Finally, we are convinced that the interactive approach should have produced better results than it did. The fact that it did not points towards the difficulty of the task and possible relevance judgment biases/inconsistencies.

References

Aronson A.R. (2001) “Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.” *Proc AMIA Symp*., 17-21.

Aronson A.R., Demner-Fushman D., Humphrey S.M., Lin J., Liu H., Ruch P., Ruiz, M.E., Smith L.H., Tanabe L.K. and Wilbur J.W. (2005) “Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents.” *The Fourteenth Text Retrieval Conference, TREC-2005*, Gaithersburg, MD.

Fox E.A. and Shaw J.A. (1994). Combination of multiple searches. In *Proceedings TREC-2*, (pp. 243-249). Gaithersburg: NIST Publication #500-215.

Ruch, P., Chichester, C., Cohen, G., Ehrler, F., Fabry, P., Marty, J., Muller, H. and Geissbuhler, A. (2004) “Report on the TREC 2004 Experiment: Genomics Track.” *The Thirteenth Text Retrieval Conference, TREC-2004*, Gaithersburg, MD.

Ruch P., Ehrler F., Abdou S. and Savoy J. (2005) “Report on the TREC 2006 Experiment: Genomics Track.” *The Fourteenth Text Retrieval Conference, TREC-2005*, Gaithersburg, MD.

Ruch P., Ehrler F., Gobeill J. and Tbahriti I. (2006) “Report on the TREC 2006 Experiment: Genomics Track.” *The Fifteenth Text Retrieval Conference, TREC-2006*, Gaithersburg, MD. (to appear)

Ruiz M.E. (2006) “UB at TREC-Genomics 2006: Using passage retrieval and pre-retrieval query expansion for genomics IR.” *The Fifteenth Text Retrieval Conference, TREC-2006*, Gaithersburg, MD. (to appear)

Salton, G (1971) *The SMART Retrieval System - Experiments in automatic Document Processing*. NJ: Prentice Hall.

Singhal A., Buckley C. and Mitra, M. (1996) Pivoted document length normalization. *SIGIR 1996*, 21-29.

Singhal A. (2001) Modern information retrieval: A brief overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 24(4):35--43, 2001.

Tanabe, L. and Wilbur, W.J. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics*, Aug 2002; 18: 1124 – 1132.