Entity-based Relevance Feedback for Genomic List Answer Retrieval *

Nicola Stokes, Yi Li, Lawrence Cavedon, Eric Huang, Jiawen Rong and Justin Zobel
National ICT Australia, Victoria Research Laboratory
Department of Computer Science and Software Engineering
The University of Melbourne, Victoria 3010, Australia.
{nstokes, yli8, lcavedon, ehuang, rongj, jz}@csse.unimelb.edu.au

Abstract

In this paper we present a system which uses ontological resources and a gene name variation generation tool to expand concepts in the original query. The novelty of our approach lies in our *concept-based normalization ranking model*. For the 2007 Genomic task, we also modified this system architecture with an additional dynamic form of query expansion called *entity-based relevance feedback*. This technique works by identifying potentially relevant entity instances in an initial set of retrieved candidate paragraphs. These entities are added to the initial query with the aim of boasting the rank of passages containing lists of these entities. Our final modification to the system, aims to maximizing the passage-level MAP score, by dropping sentences that do not contain any query concepts, from the beginning and the end of a candidate paragraph. Our TREC 2007 results show that our relevance feedback method can significantly improve baseline retrieval performance with respect to document-level MAP.

1 Introduction

The purpose of an information retrieval (IR) system is to retrieve documents or passages which are relevant to the user's information need. Genomic or biomedical IR focuses on retrieving passages which discuss genomic concepts (such as genes, proteins, biological processes) and other medical entities (such as diseases) over a collection of biomedical journal papers. The introduction of high throughput assays, and the subsequent dramatic increase in publications describing this data has made effective solutions to Genomic IR a high priority in the biomedical community.

The TREC Genomics Track[1] provides participants with a platform to test and evaluate their Genomic IR techniques and solutions. From 2006, the Genomics Track introduced a new task which focuses on retrieving short passages instead of the traditional documents or paragraphs (Hersh et al., 2006). Passages in this task are defined as text sequences that cannot exceed the paragraph (or *legal span*) boundaries, and are subsets of the original paragraphs in which they occur. Mean Average Precision (MAP) is used by the track to evaluate system performance at three different levels of information granularity: *Document*, *Passage* and *Aspect*.

Before building a system for this year's task, based on the Zettair[2] search engine, we implemented and evaluated a passage-level retrieval system that would have been ranked among the top 5 systems at the 2006 track. This system uses ontological resources (MeSH[3] and Entrez Gene[4]) and a gene name

*National ICT Australia is funded by the Australian Government's Department of Communications, Information Technology, and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Research Centre of Excellence programs.

[1] http://ir.ohsu.edu/genomics/

[2] http://www.seg.rmit.edu.au/zettair/

[3] http://www.nlm.nih.gov/mesh

[4] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

variation generation tool to expand concepts in the original query. The novelty of our approach lies in our modified Okapi ranking scheme which improves IR effectiveness by ensuring that passages that contain all of the concept terms in the original query are ranked higher than passages which contain multiple references to a subset of these concepts. In addition, the importance of the original query concepts is maintained after query expansion by using a geometric progression to normalize the contributed of the expansion terms.

In this paper, we introduce the query expansion and ranking methods used by the NICTA team at 2007 Genomics Track. We also report on our experimental results and analysis.

2 System Description

The TREC Genomics track was established in 2003 with the aim of supporting the evaluation of information retrieval systems capable of answering the types of questions typically posed by genomicists. This year's track focused on the retrieval of information supporting list-type answers to genomic queries. Here are some sample queries with general entity types in bold:

- What **DRUGS** have been tested in mouse models of Alzheimer's disease?
- What centrosomal **GENES** are implicated in diseases of brain development?
- What **MOLECULAR FUNCTIONS** does helicase protein NS3 play in HCV (Hepatitis C virus)?
- What **MUTATIONS** in apolipoprotein genes are associated with disease?
- Which **PATHWAYS** are possibly involved in the disease ADPKD?

Hence, relevant passages are those that contain both a list of specific instance of these general entity types, and the other biological concepts mentioned in the query. In all, 36 queries were evaluated on a collection of full-text journal papers, where the task was to retrieved relevant answer passages rather than full-text documents.

In the following section we describe our novel genomic retrieval system. The architecture of this system is shown in Figure 1. The retrieval of relevant answer passages involves a number of different processing steps: first the collection is preprocessed and indexed; then a user query is expanded with related terms extracted from two ontological resources MESH and Entrez Gene; an initial set of paragraphs is retrieved; these are then processed by an entity finder component which looks for entity instances in these paragraphs that match the general entity type in the user's query; these specific entity instances are then added to the query and a second set of candidate paragraphs is retrieved; these paragraphs are then reduced to passages and re-ranked before being presented to the user. A more detailed description of these steps is provided in the following subsections.

Collection Preprocessing

The TREC collection consists of full-text journal articles obtained by crawling the Highwire site[5]. The full collection contains 162,259 documents and is about 12.3 GB in size when uncompressed. After preprocessing, the whole collection becomes 7.9 GB and contains 8,920,137 paragraphs. We index these paragraphs rather than documents, since the task is a passage level retrieval task. The collection is preprocessed as follows:

Paragraph Segmentation: for evaluation purposes the Genomics Track requests that the ranked answer passages must be within specified paragraph boundaries.

Sentence Segmentation: all sentences within paragraphs are segmented using a tool called `sentence-boundary.pl`[6], with the original start and end offset information of sentences recorded.

Character Replacement: Greek characters represented by gifs are replaced by textual encodings; accented characters such as "À" or "Á" are replaced by "A"; Roman numbers are replaced by Arabic numerals. These replace-

[5] http://www.highwire.org

[6] http://l2r.cs.uiuc.edu/~cogcomp/atool.php?tkey=SS

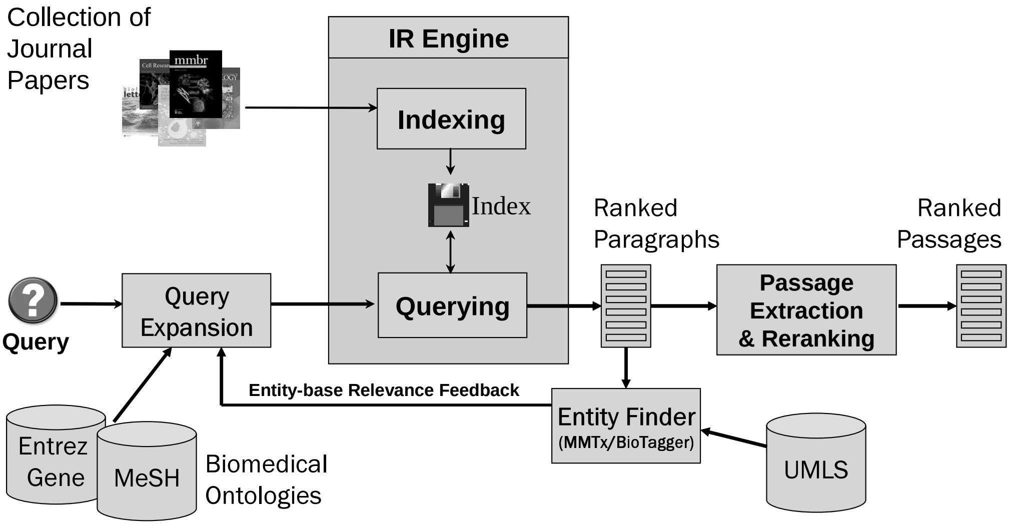

Figure 1: System architecture.

ments are very important for capturing variations in gene names.

Removal: all HTML tags, very short sentences, figures, tables, paragraphs with the heading *Abbreviations*, *Acknowledgement*, *Notes*, and some special characters are removed.

Query Expansion

Once the collection has been indexed, querying can begin. The 2007 Genomics Track focused on retrieving passages that respond to questions requiring list-type answers. Each topic contains one general biomedical entity type followed by at least one entity instance such as a gene, a disease or biological process. For example:

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

Topic 201: What **[MUTATIONS]** *in the* **Raf gene** *are associated with* **cancer***?*

Our query expansion process proceeds as follows. For each gene or protein in the query, it is expanded with entries from the Entrez Gene database. Since the same gene may occur in many different species, and many of their synonyms only differ with respect to capitalization, we choose the first entry retrieved that belongs to the species type *Homo sapiens*. Then, terms in the *Official Symbol*, *Name*, *Other Aliases* and *Other Designations* fields, for the gene, are added to the query. For example, “BRAF”, “v-raf murine sarcoma viral oncogene homolog B1”, “B-raf 1”, “BRAF1”, “MGC126806”, “MGC138284”, “RAFB1”, “B-Raf proto-oncogene serine/threonine-protein kinase”, and “Braf transforming” are added into Topic 201 as the *expanded terms* of the “Raf gene”.

For all other biological terms in the query, we use the MeSH taxonomy of medical terms to find their synonyms (using the *Entry Terms* and *See Also* fields). For example, in Topic 200, the phrase “serum protein” is expanded to “Blood Proteins”, “Proteins, Blood”, “Plasma Proteins”, “Proteins, Plasma”, “Serum Proteins”, “Proteins, Serum”, “Protein Binding”, and “Serpins”.

As a final step all the Roman digit numbers in the queries are replaced by their corresponding Arabic numbers. A similar operation was performed on the collection.

Gene Variant Generation

(Trieschnigg et al., 2006) report a significant improvement after using biomedically-tuned tokenization. As well as expanding with synonyms, we use a “gene variant” generation tool, which is based on the tokenization, to generate all the possible variants for both original query terms and expanded terms. Our segmentation rules are similar to those described in

(Buttcher et al., 2004). We describe them as follows:

Given a gene name containing a hyphen or punctuation, or a change from lower case to upper case, or from a character to a number (or vice versa), or a Greek character (e.g. "alpha"), we call this a *split point*. A word is split according to all its split points, and all variants are generated by concatenating all these split parts, optionally with a space inserted. Greek characters are also mapped to English variants, e.g. "alpha" is mapped to "a".

According to the rules, for the query term "Sec61alpha", we would generate the following lexical variants which are also commonly used forms of this term in the collection:"Sec 61alpha", "Sec61 alpha", "Sec 61 alpha", "Sec 61a", "Sec61 a", "Sec 61 a", and "Sec61a";

Concept-based Query Normalization

Our document ranking method is based on the Okapi model (Robertson et al., 1994), which is a widely used ranking metric. However, there are two fundamental problems with using this model on TREC Genomics queries.

The first problem regards Okapi not differentiating between concept terms and general query terms in the query. For example, consider two documents, one containing the terms "serum protein" and "lupus", and the other containing the terms "disease" and "lupus". Clearly, the first document containing the two biological concepts is more relevant to Topic 200. The second problem occurs because TREC 2007 Genomics topics contain more than one concept term. It is possible that a short paragraph that discusses only one concept will be ranked higher than a longer paragraph which mentions two concepts. Again this is an undesirable outcome.

To overcome these problems, a *Conceptual IR* model was proposed in (Zhou et al., 2006). In this paper we propose another method called *concept-based query normalization* which is based on the Okapi model and similar to the method introduced in (Li, 2007; Stokes et al., 2008) for geospatial IR.

The first problem is solved by dividing query terms into two types: *general terms* t_g and *concept terms* t_c. Given a query with two concept terms and a general term, the similarity between a query Q and a document D_d is measured as follows:

$$\mathit{sim}(Q, D_d) \quad = \quad \mathit{gsim}(Q, D_d) + \mathit{csim}(Q, D_d)$$

where $\mathit{gsim}(Q, D_d)$ is the *general similarity score* and $\mathit{csim}(Q, D_d)$ is the *concept similarity score*. The general similarity score is given by:

$$\mathit{gsim}(Q, D_d) \quad = \quad \sum_{t \in Q_g} \mathit{sim}_t(Q, D_d) = \sum_{t \in Q_g} r_{d,t} \cdot w_t \cdot r_{qt}$$

where Q_g is the aggregation of all general terms/phrases in the query. The concept similarity score is given by:

$$\begin{aligned} \mathit{csim}(Q, D_d) \quad &= \quad \sum_{C \in Q_c} \mathit{sim}_c(Q, D_d) \\ = \quad &\sum_{t \in C, C \in Q_c} \mathit{Norm}(\mathit{sim}_{t_1}(Q, D_d), \ldots, \mathit{sim}_{t_N}(Q, D_d)) \\ = \quad &\sum_{t \in C, C \in Q_c} (\mathit{sim}_{t_1} + \frac{\mathit{sim}_{t_2}}{a} + \cdots + \frac{\mathit{sim}_{t_N}}{a^{N-1}}) \end{aligned}$$

where Q_c is the aggregation of all concepts in the query, C is one concept in Q_c, and t_i is a term/phrase in the query, after expansion, which belongs to the concept C; the t_i are listed in descending order according to their okapi similarity scores sim_{t_1}, ..., sim_{t_N}:

$$\mathit{sim}_t(Q, D_d) \quad = \quad r_{d,t} \cdot w'_t \cdot r_{q,t}$$

where

$$\begin{aligned} r_{d,t} \quad &= \quad \frac{(k_1 + 1) \cdot f_{d,t}}{k_1 \cdot [(1 - b) + b \cdot \frac{W_d}{avgW_d}] + f_{d,t}} \\ w'_t \quad &= \quad \log \frac{N - \max(f_t, f_{t_q}) + 0.5}{\max(f_t, f_{t_q}) + 0.5} \qquad (1) \\ r_{q,t} \quad &= \quad \frac{(k_3 + 1) \cdot f_{q,t}}{k_3 + f_{q,t}} \end{aligned}$$

where k_1 and b are usually set to 1.2 and 0.75 respectively, and k_3 can be taken to be ∞. Variable W_d is the length of the document d in bytes; $avgW_d$ is the average document length in the entire collection; N is the total number of documents in the collection; f_t is the number of documents in which term t occurs; and $f_{\{d,q\},t}$ is the frequency of term t in either a document d or query q.

Note that (1) is an adjustment of the calculation for the weight w'_t of an *expansion* term t appearing in the query: for expansion term t, its own term frequency f_t and the corresponding original query term's frequency f_{t_q} are compared, and the larger value is used. This ensures that the term contributes an appropriately normalized "concept weight".

To solve the second problem, we use the following rules to ensure that for two passages P_1 and P_2, where one contains more unique concepts than the other, the number of concepts *ConceptNum(P)* will override the Okapi score *Score(P)* and assign a higher rank to the passage with more unique concepts:

```
**if** $ConceptNum(P_1) > ConceptNum(P_2)$ **then**
    $Rank(P_1) > Rank(P_2)$
**else if** $ConceptNum(P_1) < ConceptNum(P_2)$ **then**
    $Rank(P_2) > Rank(P_1)$
**else if** $Score(P_1) \geq Score(P_2)$ **then**
    $Rank(P_1) > Rank(P_2)$
**else**
    $Rank(P_2) > Rank(P_1)$
```

Entity-based Relevance Feedback

For the 2007 task, the topics are in the form of questions asking for lists of specific entities. Besides certain genes, diseases or biological processes, there are 14 general entity types. Each topic include one such entity type (e.g. "ANTIBODIES", "BIOLOGICAL SUBSTANCES" and so on). Unlike those certain gene or disease names, some relevant passages will be missing by our retrieval if they do not mention these general entity types by name. For example, for Topic 200, a relevant passage mentions "beta2glycoprotein" which is a antibody; however the term "antibody" is not used in the passage. Three tools are used to find instances of different entity types. They are: BioTagger[7] (for "GENE" or "PROTEIN"), MutationFinder[8] (for "MUTATION"), and MMTx[9] (for all other entity types, Table 1 shows a mapping relationship between MMTx headings and these entity types). We now describe our entity-based relevance feedback method:

1. Retrieve the first 100 paragraphs which include at least one instance of each concept in the query expanded by the ontologies and the variant generation tool.
2. Divide all the topics into three groups according to their entity types: Group A ("GENE" or "PROTEIN"), Group B ("MUTATION") and Group C (all other types). For topics in each group, run the corresponding tool against their top 100 paragraphs to find all instance names.
3. Among all the detected instance names, to avoid noise, discard all the names that only occur once. Then pick up the first 30 most frequent terms.
4. For all these instances, use the above generation tool to formulate all their lexical variants, and add them to the query. The expanded query is then re-submitted to the retrieval engine, and the passage extraction step, described below, is applied.

Passage Extraction

As already mentioned, in 2006, the Genomics Track defined a new question answering-type task that requires short full-sentence answers to be retrieved in response to a particular query. However, before answer passages can be generated, we first retrieve the top 1000 ranked paragraphs for each topic, and use a simple passage extraction rule to reduce these paragraphs to shorter answer spans. Any sentence in paragraphs which mention any of the query terms/phrases, including expanded ones, is called a *relevant sentence*; otherwise it is called an *irrelevant sentence*. Our method is described as follows:

1. From the first sentence to the last one, keep removing the irrelevant sentences until a relevant sentence is found;
2. Repeat this process starting from the last sentence.

This method only reduces the size of retrieved paragraphs; however, it does not split a paragraph into multiple passages.

After passage extraction has been applied for a particular topic, we re-rank passages by re-indexing them, and re-querying the topic against this new index, using the global statistics from the original indexed collection, i.e. using term/phrase frequency f_t and the average paragraph length $avgW_d$.

[7] `http://www.seas.upenn.edu/~strctlrn/BioTagger/BioTagger`

[8] `http://mutationfinder.sourceforge.net/`

[9] `http://mmtx.nlm.nih.gov/`

Table 1: Mappings between MMTx Headings and 2007 Entity Types.

Entity Types	MMTx Headings
ANTIBODIES	Concept: Antibodies Semantic Type: Immunologic Factor
BIOLOGICAL SUBSTANCES	Semantic Type: Hormone Semantic Type: Enzyme Semantic Type: Element, Ion, or Isotope Semantic Type: Carbohydrate Semantic Type: Carbohydrate Sequence Semantic Type: Lipid Semantic Type: Amino Acid, Peptide, or Protein Semantic Type: Amino Acid Sequence Semantic Type: Nucleic Acid, Nucleoside, and Nucleotide Semantic Type: Biologically Active Substance Semantic Type: Steroid Semantic Type: Eicosanoid
CELL OR TISSUE TYPES	Semantic Type: Cell Semantic Type: Tissue
DISEASES	Semantic Type: Disease or Syndrome Semantic Type: Neoplastic Process
DRUGS	Pharmacologic Substance Semantic Type: Antibiotic
MOLECULAR FUNCTION	Semantic Type: Molecular Function
PATHWAYS	N/A
STRAINS	Concept: Strain Semantic Type: Virus Semantic Type: Bacterium Concept: Sterotype
SIGNS OR SYMPTOMS	Semantic Type: Sign or Symptom Semantic Type: Finding
TOXICITIES	Concept: Toxic effect Concept: Toxicity aspects Semantic Type: Hazardous or Poisonous Substances
TUMOR TYPES	Semantic Type: Neoplastic Process

Table 2: Table showing improvement in MAP score obtained over baseline MAP when query expansion and normalization ranking methods have been used.

Run	Passage2 MAP			Aspect MAP			Document MAP		
MuBase	0.0604			0.1427			0.1896		
MuMshNfd	0.0776†	+28.5%	$P = 0.01$	0.2156†	+51.1%	$P = 0.01$	0.2724†	+43.7%	$P < 0.001$
MuMan	0.0747	+23.7%	$P = 0.1$	0.1937	+35.8%	$P = 0.2$	0.2438	+28.6%	$P = 0.1$

Table 3: Table showing improvement in MAP score when entity-based feedback is performed.

Run	Passage2 MAP			Aspect MAP			Document MAP		
MuMshNfd	0.0776			0.2156			0.2724		
MuMshFd	0.0895	+15.3%	$P = 0.2$	0.2068	−4.08%	$P = 0.5$	0.2906†	+6.68%	$P = 0.03$

Table 4: Table showing MAP scores of runs with (MuMshfd) and without feedback (MuMshNfd) when passage extraction is used.

Run	Passage2 MAP			Aspect MAP			Document MAP		
MuMshNfd	0.0776			0.2156			0.2724		
MuMshNfdRsc	0.0809	+4.25%	$P = 0.4$	0.2079†	−3.57%	$P = 0.05$	0.2682	−1.54%	$P = 0.08$
MuMshFd	0.0895			0.2068			0.2906		
MuMshFdRsc	0.0893	−0.22%	$P = 0.4$	0.2016†	−2.51%	$P = 0.02$	0.2880	−0.89%	$P = 0.2$

3 Experimental Results and Analysis

This year NICTA participated in the TREC Genomics Track and submitted three official runs:

- MuMshNfdRsc: ontology-based (MESH and Entrez Gene) query expansion and passage extraction
- MuMshFd: ontology-based query expansion and entity-based relevance feedback
- MuMshFdRsc: ontology-based query expansion, entity-based relevance feedback, and passage extraction

In this paper, for comparison purposes, we also discuss the results of three unofficial post-TREC runs:

- MuBase: baseline run using the original query terms and phrases without passage reduction
- MuMshNfd: ontology-based query expansion without passage extraction
- MuMan: a manual query expansion run without passage extraction

Table 2 presents the MAP scores of the MuBase, MuMshNfd and MuMan runs[10]. The purpose of this comparison is to evaluate the effectiveness of the query expansion (manual and automatic) with a baseline run. A paired one-sided Wilcoxon signed-rank test at the 0.05 confidence level is also reported. This table shows that both manual and ontology expansion outperform a baseline system that just retrieves paragraphs based only on the original query terms. But only the ontology expansion has statistically significant improvements in all three MAPs. It is interesting to observe that our automatic expansion method outperforms manual expansion of the queries performed by a PhD student studying Genetics. The student was instructed to add additional terms using any freely available ontologies or biomedical databases, or based solely on their knowledge of the topic.

Table 3 shows the effectiveness of our entity-based feedback method by comparing the MAP scores of the runs MuMshNfd and MuMshFd. Our dynamic feedback run (MuMshFd) shows improved Passage2 and Document MAP scores; however, only the Document MAP increase is statistically significant. There is also a small drop in Aspect MAP. This may be explained by the fact that we make no attempt to optimize for this metric.

Table 5: Table showing performance of our best Passage2 MAP scoring run MuMshFd with the maximum and median average scores on the Genomics Track.

Run	Passage2 MAP	Aspect MAP	Document MAP
MAX	0.1817	0.4156	0.4353
MEDIAN	0.0278	0.1078	0.1871
MuMshFd	0.0895	0.2068	0.2906

Table 4 compares the performance of our passage extraction strategy. Four runs are investigated: MuMshNfd, MuMshFd and their passage extraction extensions MuMshNfdRsc and MuMshFdRsc. As observed in our TREC Genomics 2006 experiments, while the Passage2 MAP increases, the Aspect and Document MAP scores drop slightly when passage extraction is applied (Stokes et al., 2007). However, in our 2007 experiments no significant improvement on Passage2 MAP occurs with passage extraction. This can be explained by the difference in gold standard span lengths between the 2006 and 2007 tasks. We found that on average, these spans are much shorter in 2006 than in 2007: 1.7 sentences for 2006 compared with 3.9 sentences per span for 2007. Hence, this year's gold standard passages are closer in length to actually paragraphs, thus reducing the need for passage reduction.

Our final set of results in this section, shows how our best run (MuMshFd) performs with respect to systems that participated in the official TREC 2007 Genomics Track. Table 5 lists the MAP scores of the best and median results of all the participants. The maximum and median results in this table were calculated according to the average scores of the best and median runs for each topic, which belong to different systems, so these scores are somewhat inflated. Hence, our best run's (MuMshFd) MAPs are lower than the maximum scores listed here, but this run does outperform the median average scores in all three MAPs, as does our baseline (MuBase).

[10] The PASSAGE2 MAP score is used as the primary passage retrieval evaluation measure in 2007

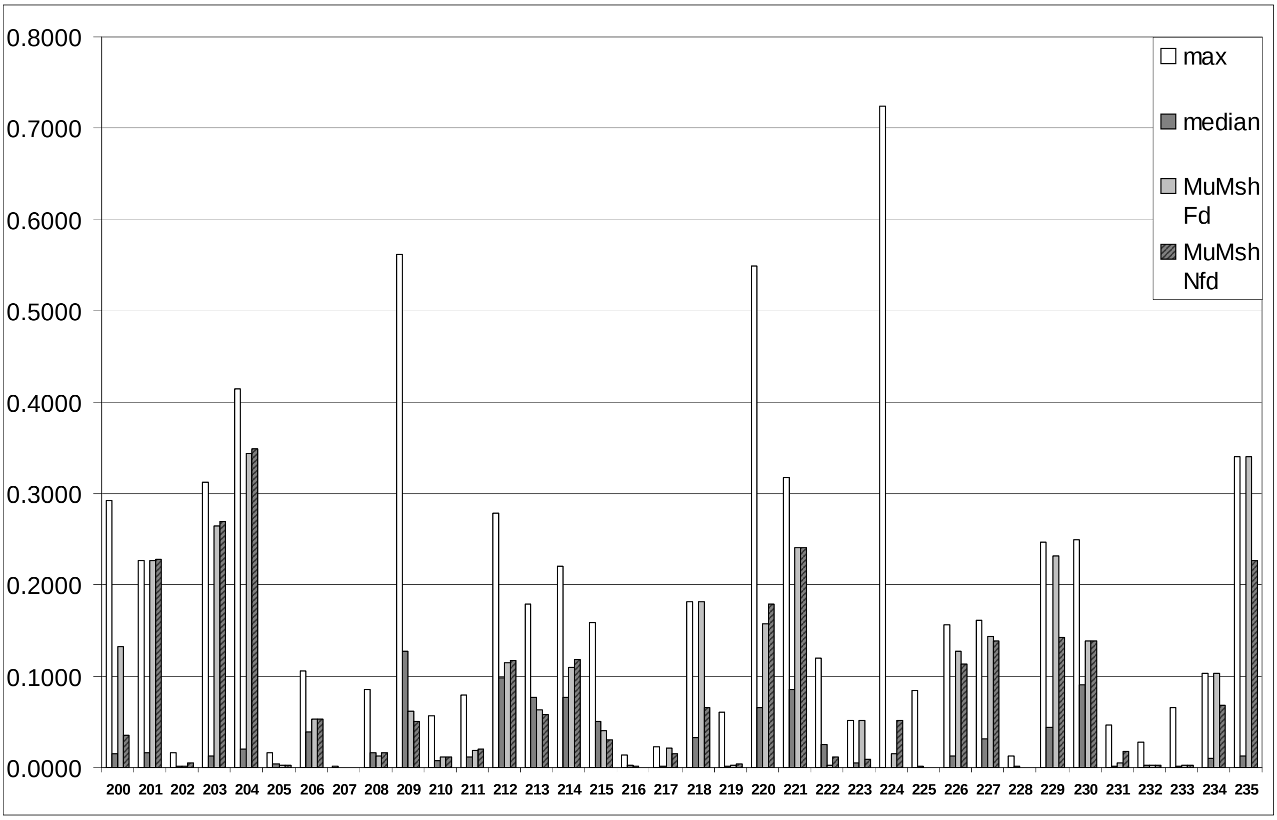

Figure 2: Passage2 MAPs of the maximum, median results, as well as our `MuMshFd` and `MuMshNfd` runs on the basis of each topic.

4 Discussion

In this section, we will discuss some of the pros and cons of our entity-based feedback method and some observations on the performance of our system at a topic level.

Given the wide-variety of entity types under investigation at this year's TREC Genomics Track, it was impossible to annotate enough training data in time to build a classifier that could tag all instances of these entities in the TREC collection before indexing. Apart from gene, protein and mutation entities there are no freely available machine learning entity classification tools to perform full entity annotation. The MMTx tool is a rule-based lookup system that provides extensive entity coverage, at the cost, however of efficiency. Runtime estimates for annotating the entire collection far exceeded system development time allowed by the Genomics Track — around 63 days for the MMTx tool. Consequently, we chose an intermediate solution where a subset of an initial set of retrieved paragraphs was annotated automatically by a set of open source taggers, and all annotated terms matching the entity type in the query are then added to the initial query, which is followed by a second retrieval step.

We have not performed a detailed analysis of the miss and false alarm rates of the entity taggers we used; however, we have confirmed that there is a reasonable correlation between a high *entity feedback/gold standard passage* overlap score[11] and an increase in MAP score.

Figure 2 shows the per topic Passage2 AP (average precision) score of the maximum and median participant result scores, as well as our `MuMshFd` and `MuMshNfd` runs. Five of our feedback run (`MuMshNfd`) topics match the AP score of the maximum system performance for that topic (see topics 201, 218, 223, 234, and 235). In contrast four of our topics for the feedback run have AP scores that are slightly less than the median scores (see topics 209, 213, 215 and 222). Apart from topic 209 these dif-

[11] the entity feedback/gold standard passage overlap score is calculated as the number of entity feedback terms that are mentioned in the gold standard passages for a given topic divided by the total number of entity feedback terms added to the query.

ferences are only very slight. There are only two topics (224 and 231) where our feedback method is obviously outperformed by the non-feedback run. However, our Passage2 MAP scores in Table 3 show that our feedback run improvement is not statistically significant compared to the non-feedback run. We can explain this outcome by looking at the topic results in more detail.

In general, we have observed that the feedback run returns many new relevant passages that were not retrieved by either the non-feedback or baseline system; however, in turn many of the previously retrieved relevant passages have dropped out of the top 1000 candidate passages in the feedback run. We have identified the following reasons for this. In some cases, entity feedback terms, while relevant, are also either general English terms (e.g. “Net”, “Bad”) or ambiguous abbreviations (e.g. “rD” and “HI”) — this is particularly evident in gene and protein entity feedback. For example, the feedback method identified the gene “HI”, which is also a commonly used abbreviation for *histamine*, *haemophilus influenzae*, *hemagglutination inhibition* and many other biological concepts. These ambiguous terms may be responsible for retrieving off-topic passages. We noticed that there are many instances where a candidate passage only shares a single entity feedback term with the query; hence, in future system development we need to account for this, by either filtering out these passages or adjusting the weighting scheme to boost the relevance of passages which mention concrete concept instances that were in the original query.

5 Conclusion

In conclusion then, our results concur with those reported by previous TREC genomics participants, where query expansion using ontological resources and a lexical variant generation tool have provided statistically significant improvements over baseline systems. However, our passage reduction strategy did not significantly improve Passage MAP this year, since gold standard passages tended to be nearer the length of actually paragraphs, which differs from gold standard passages in 2006. Our results also show that for this task a manual expansion of the queries does not produce significant performance improvements over the baseline.

Our major contribution to this year’s track, was the evaluation of an entity-based feedback method that captured entity instances from an initial set of retrieved paragraphs that were then used to expand the original query. This method improved our results; however, the difference was only statistically significant at the document MAP level. In general our system achieved higher scores than the median participant score for the document, passage and aspect MAPs. Our intentions for future work is to address the problems with our feedback technique that were outlined in Section 4.

References

S. Buttcher, C.L.A. Clarke, and G.V. Cormack. 2004. Domain-specific synonym expansion and validation for biomedical information retrieval. In *The Thirteen Text REtrieval Conference (TREC 2004) Proceedings*.

W. Hersh, A. Cohen, P. Roberts, and H. Rekapalli. 2006. Trec 2006 genomics track overview. (Voorhees and Buckland, 2006).

Yi Li. 2007. Probabilistic toponym resolution and geographic indexing and querying. Masters thesis, The University of Melbourne.

S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. 1994. Okapi at TREC-3. In *The Third Text Retrieval Conference (TREC 3) Proceedings*, Gaithersburg, Maryland, November.

N. Stokes, Y. Li, L. Cavedon, and J. Zobel. 2007. Exploring abbreviation expansion for genomic information retrieval. In *Proceedings of Australasian Language Technology Workshop 2007*, Melbourne, Australia, December.

N. Stokes, Y. Li, A. Moffat, and J. Rong. 2008. An empirical study of the effects of NLP components on Geographic IR performance. *International Journal of Geographical Information Science*. To appear.

D. Trieschnigg, W. Kraaij, and F. de Jong. 2006. The influence of basic tokenization on biomedical document retrieval. In *SIGIR 2007 Proceedings*, Amsterdam, The Netherlands, July.

E. M. Voorhees and Lori P. Buckland. 2006. *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*. NIST, Gaithersburg, Maryland.

W. Zhou, C. Yu, V. Tovik, and N. Smalheiser. 2006. A concept-based framework for passage retrieval in genomics. (Voorhees and Buckland, 2006).