TREC Genomics 2004:

Gail Sinclair **Bonnie Webber**
School of Informatics,
University of Edinburgh
{csincla1, bonnie}@inf.ed.ac.uk

1 Introduction

The TREC Genomics track started in 2003 as the first domain specific track of the Text Retrieval Competition. The aim of the track is to develop various IR tasks specific to the biomedical field. One task of the first year involved the retrieval of documents given a specific gene, while the second task required the extraction a brief description of gene function from documents. This year sees a foray into ad hoc retrieval and a curation and categorization task.

2 Ad hoc Retrieval

2.1 Task description

It was important that the retrieval tasks mirrored the real life current information needs of biologists. Examples of these needs were ascertained via interviews which were then formulated into queries suitable for ad hoc IR. 50 of these queries were then chosen as the test topics. 5 queries and their selected relevant documents were also later given as training examples. As is standard in TREC ad hoc retrieval, each topic has a title, need and context field, e.g.
TITLE: DNA repair and oxidative stress
NEED: Find correlation between DNA repair pathways and oxidative stress
CONTEXT: Researcher is interested in how oxidative stress effects DNA repair.

The purpose of the task was then to retrieve only those documents relevant to the queries. Since analysing every document in the document set is so resource intensive and the track has a limited timeline, standard practice was used to evaluated based on sampling. The top 100 documents of each run submitted by the track participants were analysed and judged on relevance. For evaluation, each document judged as relevant was counted as a positive instance and each document judged as not relevant plus all those documents not analysed were counted as negative instances. The nature of this evaluation technique means that there is likely to be some positive examples treated as true negatives.

2.2 Methods

Since there was initially no training data for this task, we decided to use an existing dataset from a related domain.

The MuchMore[1] corpus contains 25 medical queries and their relevance judgments with respect to almost 8,000 abstracts from 41 journals. These queries differed from the track queries in that they only contained the need and had no title or context field. The abstracts are in English but translated from German. MuchMore is a parallel corpus, with abstracts in both English and German. It is often used for cross lingual IR. The relevance judgments are supplied in a format amenable to TREC evaluation. A version of the corpus is annotated with various linguistic information such as part-of-speech, morphology, UMLS semantic classes. However, for our purposes, the plain version without any annotation was used as this was most similar to the test TREC Genomics queries.

For retrieval, we used the same system as we did in the 2003 track(Osborne et al., 2003), the Lucene retrieval engine and expansion of gene names on the queries. This system performed well in the 2003 track, and so it seemed reasonable to use it again.

Using default Lucene retrieval and the MuchMore queries, the baseline evaluation gave us TREC metrics of

MAP	27.7%
Recall	55.3%
Ave Prec at 0.1 recall	64.6%
Prec at 10 docs	51.6%

This year we focused on how we could differently expand and weight the queries.

[1] http://muchmore.dfki.de/

2.2.1 Expansion

The UMLS provides a knowledge server[2] that, given a term or phrase, will search the UMLS according to certain criteria, e.g. exact string match, normalised string match.

Expanding the queries using UMLS-sourced synonyms for each word in the query increased MAP and recall, while decreasing precision at 0.1 and average precision at 10.

MAP	29.3%
Recall	65.9%
Ave Prec at 0.1 recall	61.7%
Prec at 10 docs	49.6%

2.2.2 Weighting

Our query expansion from last year used a weighting scheme for the different type of gene representations e.g. official symbol was weighted higher than alias product. Since this year involved sentences rather than just different representations of the same entity, a strategy was required to find out which terms should be weighted higher than others.

We decided first to weight noun phrases higher than the other words in the query – how much so was left to experimentation. Whereas last year it was found that the weight 2.9 was most useful in performance with regards to gene representations, with MuchMore's type of queries, weighting the nouns 7 times more relevant than the rest of the query was found to best increase performance. However using any weighting at all for this subset of terms significantly bettered the default of no weighting.

It then seemed appropriate to order the query terms in order of their ability to discriminate between documents. For this, we used a term's frequency in the literature (via PubMed) or on the Web (via a Google API). This strategy was used as a follow on from our success in the BioNLP task at Coling 2004(Finkel et al., 2004).

The Google API[3] was used to find the frequency of each term across the Web. The terms were then weighted according to this frequency with respect to the other terms in the query - the lower the frequency the greater the weight. The highest frequency term received a weighting of 1.0, with the weights being incremented with each next lower frequency term.

Similarly each term was individually used as a search term in Pubmed and the number of documents retrieved was automatically recorded. This number of documents was then used to weight the terms in a similar way as above.

Additionally, to incorporate the fact that a term may have a higher relative frequency to another in a PubMed search than in a Google search, the term orderings determined by the two search strategies were merged, with the weights of each term in a query being averaged, e.g if Google gave term **X** in a query a weighting of 3 and PubMed gave the same term in the query a weighting of 2, the combined weighting was 2.5.

Using these weighting schemes, the merged weighting scheme achieved slightly better performance compared to the individual schemes, as follow:

	Google	PubMed	Both
MAP	36.1%	36.0%	36.3%
Recall	61.5%	61.1%	61.5%
Ave Prec at 0.1 recall	76.1%	73.1%	75.6%
Prec at 10 docs	62.4%	60.8%	61.6%

The combination of the two weighting decisions was done manually and so for simplicity only Google was used for the test data, as it was the best performer when the two were compared. The test data had many more words and so manually combining the weightings would have been time consuming although automating the process would not have been difficult.

When synonyms were used, their associated weighting was the same as the original term from which the synonym was obtained. Frequencies of these new terms could have been found independently however this could have given false emphasis to an irrelevant synonym if it happened to be a rarer term than the original.

The official runs submitted to TREC involved a combination of the techniques described above. One run used the individual terms, noun phrases and synonyms of both the terms and noun phrases which were then weighted with respect to usage frequency (according to Google). The second run also included the use of stemming.

Although both runs performed similarly overall, the former technique performed significantly better than the second more often on individual queries than vice versa. This would lead us to surmise that stemming can often be a hindrance, echoing our findings in (Sinclair and Webber, 2004)

[2] http://umlsks.nlm.nih.gov/

[3] http://www.google.com/apis/

3 Categorization Task

3.1 Task Description

The Mouse Genomics (MGI) team currently manually curate new articles for annotation with Gene Ontology (GO) codes. The Gene Ontology consists of 3 separate vocabularies - one for each of *biological process, cellular component* and *molecular function*. The MGI first decides whether each new article is relevant to mouse genomics and so possibly amenable to GO annotation, then any relevant GO codes are assigned together with the evidence for that code.

The categorization task attempts to imitate this process in three parts:
1) triage task - decision on whether each document contains experimental evidence of mouse genomics and can be considered for annotation.
2) decision on which vocabularies each article could be annotated with
3) the evidence on which the above decision is based.

Only our efforts for the first triage part of this task was entered into the track competition.

3.2 Triage Task

3.2.1 Division of Text

To further existing work we have been doing with categorization according to GO codes(Sinclair and Webber, 2004), for this task we wanted to compare the information content of the different sections of full text articles.

The documents were initially classified according to whether they were about mice. The same species classifier used in the 2003 track was used for this(Osborne et al., 2003). This classification went further than the initial MGI retrieval in that one simple mention of mouse, mice or murine within the article was not sufficient to classify as being 'about mice'. Any documents considered not about mice were then removed from the document set as they were perceived as not curatable according to MGI's curation process.

The remaining dataset of full text articles was divided up according to sections. A sample set of the SGML of the articles for the three journal publications used in the dataset was studied and a division strategy devised accordingly, so that appropriate sections could be kept together in the groupings ***Abstract, Introduction, Methods, Results, Conclusions***. (***Abstract*** also includes the article titles.) Unfortunately, any articles not formatted in this way were then lost to the categorization task. Any document omitted in this manner is then deemed to have had a "do not curate" decision made upon it. Although these omissions result in a reduced subset of documents that may be curated, it does not take away from the overall intention of our experimentation - i.e. how 'useful' each section is for decision making.

The ***MeSH*** and ***RN*** annotations for the articles were also retrieved and combined with the ***Abstract*** documents since these are publicly available. This data was merged so that the less publicly available full text sections could be compared for 'usefulness' with what is already widely available and most extensively used in biomedical information retrieval, i.e PubMed annotations.

In the training set of 5837 documents, 100% of the articles had an ***Abstract*** section, 11 documents did not have any distinct ***Introduction***, ca. 100 documents did not have a distinct ***Results*** section, ca. 500 documents did not have a distinct ***Discussion*** or ***Conclusions*** section. However more than half of the training documents did not have a distinct ***Methods*** section.

In the test document set, all section groupings contained approximately 85% of the articles, except the ***Discussion*** grouping which contained 79% of the full test set. It is unclear why so may articles in the training set lack a ***Methods*** section. This is particularly regrettable since the ***Methods*** section seemed to be the most informative vis-a-vis the curation decision.

These subsets were then searched for indicators of GO code.

3.2.2 GO Code Identification

All GO codes were extracted from the ontology and synonymous phrases were looked for in the UMLS Knowledge Server used in the retrieval task. The number of instances of each GO term and associated synonyms were recorded for each article. The training triage decisions were then analysed according to these counts to see if there was any trend in the uses of GO terms in the documents. A lower bound was formulated so that if any document did not contain at least that number of GO terms and/or synonyms then the triage decision would be not to curate.

At this point the ratio of remaining documents deemed not worthy of curation and those already discarded was significantly different to that of the same ratio in the training set. There was much discussion across Track participants about the ratio of curated to not curated documents in the training and test sets. The ratio in both sets was deemed to be not significantly dissimilar. According to this further filtering was considered appropriate.

Comparison of Sections

Precision	Methods	Public	Results	Discussion	Introduction
Recall	Results	Public	Methods	Discussion	Introduction
F-Score	Methods	Results	Public	Discussion	Introduction
Utility	Results	Methods	Public	Discussion	Introduction

Table 1: Comparison of article sections with respect to the Track metrics, from highest scoring section (leftmost) to lowest.

Although our method of species classification from the 2003 track proved to be very useful, it was not ascertained whether it was significantly better than several other teams' classification by MeSH term as each strategy generated different false positives and negatives. To this end (and at the last minute), the documents were further filtered according to whether the documents had Mice as a MeSH heading. This reduced the aforementioned ratio to a number not so significantly different to that of the training data. In hindsight this was probably a mistake, but the decision was panic-driven.

As can be seen in Table 1, in all 4 metrics used in this task, ***Introductions*** proved the least useful, closely followed by the ***Discussion*** sections. The ***Methods*** sections proved most informative with respect to Precision and F-score, with the ***Results*** sections outperforming the rest with respect to Recall and Utility Measure.

Although the publicly available data performed well, these results show that it is worthwhile to invest resources in analysing full text.

References

Jenny Finkel, Shipra Dingare, Hy Nguyen, Malvina Nissim, Christopher Manning, and Gail Sinclair. 2004. Exploiting context for biomedical entity recognition: From syntax to the web. In *NLPBA/BioNLP 2004,Joint Workshop on Natural Language Processing in Biomedicine and its Applications*.

M. Osborne, M. Cumiskey, G. Sinclair, M. Smillie, B. Webber, J. Chang, N. Mehra, V. Rotemberg, and R.B. Altman. 2003. Edinburgh-stanford trec-2003 genomics track. In *Proc. of the Twelfth Text REtrieval Conference (TREC-12)*.

Gail Sinclair and Bonnie Webber. 2004. Classification from full text: A comparison of canonical sections of scientific papers. In *NLPBA/BioNLP 2004,Joint Workshop on Natural Language Processing in Biomedicine and its Applications*.