# Concept based document retrieval for genomics literature

Dolf Trieschnigg[*]
University of Twente
trieschn@ewi.utwente.nl

Wessel Kraaij[†]
TNO
kraaijw@acm.org

Martijn Schuemie[‡]
Erasmus MC
m.schuemie@erasmusmc.nl

**Abstract**

The 2006 TREC Genomics evaluation focuses on document, passage and aspect retrieval in the genomics domain. The Erasmus Medical Center, TNO and University of Twente collaborated on an approach combining concept tagging (named entity recognition) and information retrieval based on statistical language models. Experiments on the 2004 collection show that document retrieval based on concepts could not outperform the baseline based on words. However, experiments on the 2006 collection shows no significant difference between the two approaches. Further investigation has to show if and how these concept and word based language models can be effectively combined.

## 1 Introduction

The TREC Genomics Track focuses on information retrieval in the genomics domain. The 2006 task introduces a passage retrieval task which evaluates the retrieval performance at passage, aspect and document level [1]. This report discusses the collaborative work of the University of Twente, Erasmus Medical Center and TNO for the TREC Genomics Track 2006 in which we combine our experience from previous TREC evaluations [5, 8, 12].

Working with genomics literature is a challenge not only because of ambiguity inherent to natural language. The same vocabulary is used for different biomedical concepts, concepts are known by multiple names and some concept names have generic English meanings as well [13]. We expect that explicitly tagging documents with biomedical concepts can improve information retrieval. Ideally, correctly identified concepts in both documents and queries should improve precision in the search process. On the other hand, errors in the translation between words and concepts might lead to decreased performance. Furthermore, the specificity of the concepts in the queries should match those in the relevant documents.

In our approach we have focused on improving biomedical document retrieval by using statistical language models based on concepts. For the 2006 evaluation, an ad hoc algorithm was developed to extract relevant passages from relevant documents.

In this report we will try to answer the following research questions:

- Can a retrieval system based on concept language models outperform a system based on only words?
- What is the influence of tagging concepts at different levels of specificity?
- What is the influence of using full text papers on retrieval performance?
- What is the influence of mistagging concepts in the topics?

[*] Human Media Interaction group, University of Twente, Enschede, The Netherlands

[†] TNO Information and Communication Technology, Delft, The Netherlands

[‡] Biosemantics Group, Medical Informatics Department, ErasmusMC University Medical Center, Rotterdam, The Netherlands

The outline of this report is as follows. Sections 2 and 3 discuss the method used for preprocessing and tagging the documents and topic descriptions with biomedical concepts. Section 4 describes the document retrieval method. An ad hoc algorithm was applied to extract relevant passages from relevant documents, which is described in section 5. Section 6 describes the experiments and results carried out with the 2004 and 2006 collection and queries, followed by a concluding section.

## 2 Document and topic preprocessing

The HTML documents are split into sections with corresponding section titles, using several different templates to support the differences in document formatting. Texts within `<TABLE>`, `<A>` (hyperlink), and `<FONT>` tags are ignored, thus also ignoring figure captions. Sentences are split using a simple algorithm developed by our team, using an unsupervised sentence boundary detection approach (based on the work of A. Mikheev [10]). For each sentence the byte offsets of the first and the last character in that sentence are reported. This set is made available for download for other TREC participants [4].

Sections pertaining materials & methods, literature references and acknowledgements were removed.

From both documents and topics stop words were removed and words were stemmed to their uninflected form by means of the normalizer of the lexical variant generator [9].

## 3 Concept tagging

We use a thesaurus to identify concepts in the documents and topics. The use of a thesaurus allows the identification of multi-word terms and the mapping of synonyms to one concept. The thesaurus consists of the Unified Medical Language System (UMLS) [2], and a gene-thesaurus [14]. Concepts are identified in text by comparing (sequences of) words to the words in the thesaurus terms.

Two variations of concept tagging are used:

- Narrow concept tagging, and
- Full concept tagging.

If sequences of different length are identified at the same location, only the longest sequence is selected in case of narrow concept tagging. For instance, “Alzheimer’s disease” maps to both “disease”, and “Alzheimer’s disease”. In case of narrow concept tagging, the latter is selected as the correct tag. In case of full concept tagging both tags are selected.

Gene nomenclature contains many ambiguities [3], and we have therefore implemented a simple disambiguation algorithm similar to Koike & Takagi [7]. Using the gene thesaurus, our system could identify gene and protein names with a precision of 70.8% and recall of 70.8% on the Biocreative II Gene Normalization training set.

Figure 1 shows the result of the tagging process for an example document and topic.

## 4 Document retrieval

An inverted file index is created for the tagged collection to support document retrieval. The document retrieval is based on statistical language modeling using Jelinek-Mercer smoothing [6, 11]. This approach was used and proved to be succesful during previous TREC evaluations [5, 15, 8].

The probability that a query of a sequence of terms $T_1, \ldots, T_n$ is sampled from a document $D$ is defined by:

$$P(T_1, T_2, \ldots, T_n|D) = \prod_{i=1}^{n} \lambda P(T_i|D) + (1-\lambda)P(T_i|C)$$

```
<article pmid="10901333">
<section id="title">
  <sentence id="0" start="22" end="116">
    measurement
    low
    level
    arsenic
    exposure
    a
    comparison
    water
    toenail
    concentration
  <concept id="238690" tokens="3,4"/>
  <concept id="43047" tokens="7"/>
  <concept id="681563" tokens="7"/>
  <concept id="222007" tokens="8"/>
  <concept id="86045" tokens="9"/>
  </sentence>
</section>
<section ...>
   <sentence ...>...</sentence>
   ...
</section>
...
</article>
```

(a) Example of tagged document

```
<topic id="160">
<section id="genesymbols">
  <sentence id="0">
  prnp
  <concept id="2008214" tokens="0"/>
  </sentence>
</section>
<section id="additional">
  <sentence id="0">
    mad
    cow
    disease
    <concept id="85209" tokens="0,1,2"/>
  </sentence>
</section>
<section id="question">
  <sentence id="0">
    what
    be
    role
    prnp
    mad
    cow
    disease
    <concept id="35820" tokens="2"/>
    <concept id="85209" tokens="4,5,6"/>
  </sentence>
</section>
</topic>
```

(b) Example of tagged topic description

Figure 1: Examples of tagged document and topic description

$P(T_i|D)$ and $P(T_i|C)$ are the probabilities the term is sampled from a particular document $D$ or the complete collection $C$ respectively. These probabilities are based on maximum likelihood estimates. $\lambda$ is the Jelinek-Mercer smoothing parameter.

Two different features are used as terms for the index:

- The tokenized words from the documents.
- The concepts identified in the documents.

# 5 Relevant passage retrieval

An ad hoc algorithm is used to perform the relevant passage retrieval. For each retrieved document from the document retrieval step, one relevant passage is extracted. This passage consists of one or more "best scoring" sentences in the document, i.e. the sentence containing the most query terms expanded with adjacent sentences containing at least one query term.

More formally, a document $D$ is modeled as a list of sentences $S_1$ to $S_n$:

$$D = [S_1, \ldots, S_n]$$

Both a query $Q$ and a sentence $S$ consist of a bag of terms.

$$S = t_1, \ldots, t_s \qquad Q = t_1, \ldots, t_q$$

The score of sentence $S$ for a particular query $Q$ is defined as the overlap between its terms:

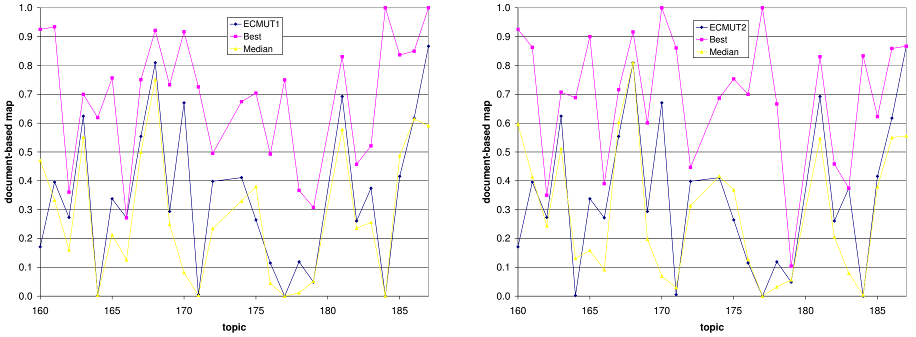
$$s(S, Q) = |S \cap Q|$$

The index $x$ of the best scoring sentence $S_x^*$ is defined as:

$$x = \arg \max_{x \in \{1..n\}} s(S_x, Q)$$

The relevant passage $P^*$ is a list of sequential sentences $[S_p, \ldots, S_q]$ (including $S_x^*$), having a score larger than 0.

$$i \in (p..q) \bullet s(S_i, Q) > 0 \ \wedge \ p \leq x \leq q$$



(a) automatic run (EMCUT1)

(b) manual run (EMCUT2)

Figure 2: Document based MAP per topic

# 6 Experiments & results

Initial experiments were carried out on the TREC genomics 2004 corpus with topics 1 to 50. These experiments were limited to document retrieval and the best configuration was used for submitting official TREC 2006 runs. No experiments were carried out with the passage retrieval algorithm because of a lacking test collection.

Section 6.1 describes the runs and results of our official runs. Section 6.2 describes the runs carried out before and after the official evaluation on both the 2004 and 2006 document collection.

## 6.1 Official runs

We submitted two official runs for the 2006 evaluation, `EMCUT1` and `EMCUT2`. `EMCUT1` is an automatic run in which the topics have been preprocessed the same way as the documents; `EMCUT2` is a manual run, in which incorrectly tagged concepts in the topics have been manually corrected.

Based on experiments on the 2004 collection, we carried out document retrieval using only query words (with $\lambda$ set to 0.35). The relevant passages were extracted from the documents by matching the concepts in the query with the tagged concepts in the documents.

The document retrieval process for both runs are identical and performs above median (0.35 MAP). The passage retrieval algorithm performs poor, witnessing the MAP scores for passage and aspect retrieval (0.01 and 0.09 MAP respectively). Remarkably, the character based map of the manual run is slightly higher than the map from the automatic run, although this difference is not significant.

Figures 2, 3 and 4 show the document, character and passage based retrieval performance respectively.

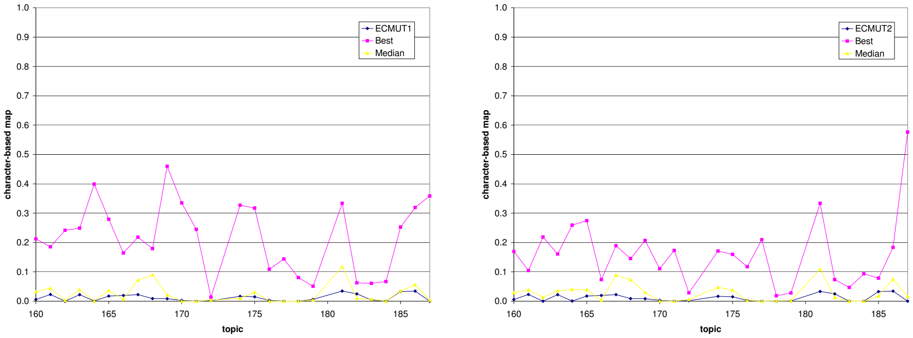
## 6.2 Additional runs

Before and after the TREC 2006 evaluation, experiments were carried out on the 2004 and 2006 collection. The 2006 collection contains full-text documents in contrast to the 2004 collection which contains only abstracts. To measure the influence of using full-text documents we also carried out experiments on the abstracts of the 2006 collection.

For each collection three indices were created: one based on words, one based on narrow concepts and one based on all concepts.

Uncertain if the smoothing parameter $\lambda$ would give different results when using index terms other than words [16], we varied $\lambda$ between 0.1 and 0.9.

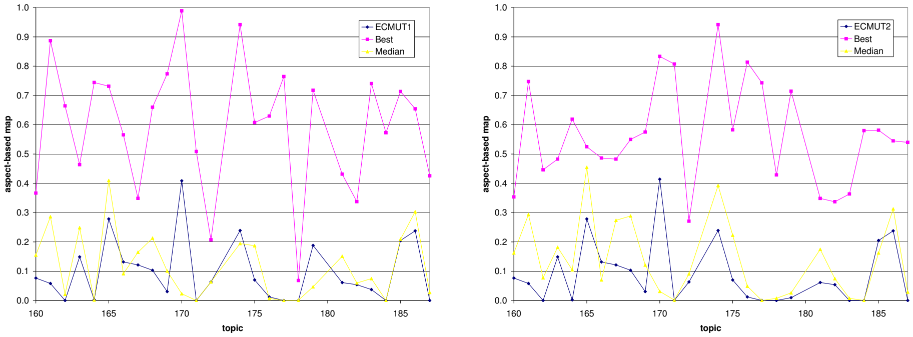
The following sections discuss the experiments on the three different collections.



(a) automatic run (EMCUT1)

(b) manual run (EMCUT2)

Figure 3: Character based MAP per topic



(a) automatic run (EMCUT1)

(b) manual run (EMCUT2)

Figure 4: Aspect based MAP per topic for automatic run.

|  | words (W) |  |  | narrow concepts (NC) |  |  | full concepts (FC) |  |  |
|---|---|---|---|---|---|---|---|---|---|
| smoothing | T | TN | TNC | T | TN | TNC | T | TN | TNC |
| 0.1 | 0.2741 | 0.3381 | 0.3240 | 0.1181 | **0.1358** | **0.1316** | 0.1379 | 0.1649 | 0.1470 |
| 0.15 | 0.2823 | 0.3469 | **0.3298** | 0.1189 | 0.1356 | 0.1314 | 0.1401 | 0.1650 | 0.1474 |
| 0.2 | 0.2878 | 0.3547 | 0.3266 | 0.1195 | 0.1357 | 0.1307 | 0.1412 | 0.1656 | 0.1473 |
| 0.25 | 0.2909 | 0.3575 | 0.3236 | 0.1201 | 0.1349 | 0.1299 | 0.1433 | 0.1663 | **0.1499** |
| 0.3 | 0.2926 | 0.3591 | 0.3202 | **0.1202** | 0.1348 | 0.1291 | 0.1451 | 0.1665 | 0.1494 |
| 0.35 | 0.2939 | 0.3597 | 0.3161 | 0.1200 | 0.1343 | 0.1277 | 0.1464 | 0.1670 | 0.1495 |
| 0.4 | **0.2952** | **0.3608** | 0.3104 | 0.1197 | 0.1335 | 0.1267 | 0.1472 | **0.1673** | 0.1489 |
| 0.5 | 0.2944 | 0.3581 | 0.3004 | 0.1189 | 0.1324 | 0.1243 | 0.1489 | 0.1663 | 0.1477 |
| 0.6 | 0.2926 | 0.3525 | 0.2901 | 0.1180 | 0.1309 | 0.1224 | 0.1497 | 0.1663 | 0.1447 |
| 0.7 | 0.2895 | 0.3450 | 0.2791 | 0.1166 | 0.1287 | 0.1188 | **0.1498** | 0.1659 | 0.1402 |
| 0.8 | 0.2849 | 0.3247 | 0.2661 | 0.1158 | 0.1277 | 0.1149 | 0.1492 | 0.1654 | 0.1375 |
| 0.9 | 0.2797 | 0.3087 | 0.2446 | 0.1149 | 0.1254 | 0.1088 | 0.1481 | 0.1645 | 0.1341 |

Table 1: MAP of document retrieval using different topic sections for query construction (combinations of Title, Need and Context) and varying the smoothing parameter (2004 collection). Values in **bold** denote the maximum in that column.

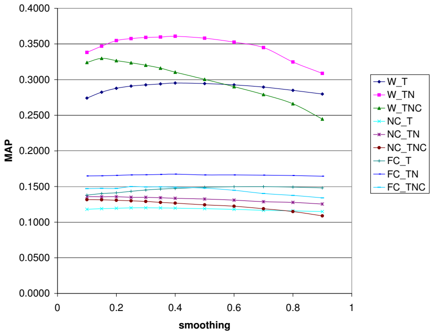


Figure 5: MAP of document retrieval using different topic sections and smoothing values (2004 collection)

### 6.2.1 2004 collection

The topics of the TREC Genomics 2004 collection consist of a Title, Need and Context section. We experimented with using the features (i.e. either words or tagged concepts) from only the Title section, a combination of Title and Need sections and the combination of all three sections.

Table 1 and figure 5 show the mean average precision of the runs. The runs based on words clearly outperform the runs based on both narrow and full concepts with a MAP ranging between 0.24 and 0.36. The runs based on full concepts (MAP between 0.13 and 0.17) outperform the runs based on narrow concepts (MAP between 0.11 and 0.14).

The optimal smoothing parameter for word based queries is correlated with the length of the query: a longer query requires a smaller smoothing parameter (i.e. giving more weight to the collection probability of term). This corresponds to conclusions from Zhai and Laverty [16]. The performance of the concept based queries does not seem to be strongly influenced by the smoothing parameter.

Both word and concept runs show the best performance when the query is constructed using features from the Title and Need sections of the query.

In order to understand the differences in results between a concept-index and a wordbased-index we compared the results per topic using the best concept runs and the best word based run (see Table 2 and Figure 6).

For a few topics the runs based on concepts have a larger average precision. We suspect that this is caused by the detection of two words as a single concept (e.g. "iron transporter" in topic 1). Since the combination of these two words is much more specific than each word individually, it increases the chance of selecting the correct phrase. We suspect that some of the poor scoring

| topic | W | NC | FC | topic | W | NC | FC |
|---|---|---|---|---|---|---|---|
| 1 | 0.3107 | 0.4363 | **0.4881** | 26 | **0.6979** | 0.0039 | 0.0046 |
| 2 | **0.0858** | 0.0187 | 0.0215 | 27 | **0.5506** | 0.0900 | 0.0724 |
| 3 | **0.0559** | 0.0207 | 0.0317 | 28 | **0.4478** | 0.1022 | 0.2144 |
| 4 | **0.0457** | 0.0084 | 0.0205 | 29 | **0.2959** | 0.0003 | 0.0002 |
| 5 | **0.0190** | 0.0082 | 0.0075 | 30 | **0.1952** | 0.0000 | 0.0000 |
| 6 | **0.3686** | 0.2904 | 0.2517 | 31 | 0.1786 | 0.0665 | **0.1930** |
| 7 | **0.3826** | 0.0898 | 0.3402 | 32 | **0.3293** | 0.1346 | 0.2308 |
| 8 | **0.1608** | 0.1042 | 0.1534 | 33 | **0.2870** | 0.0238 | 0.0013 |
| 9 | **0.8495** | 0.0000 | 0.0000 | 34 | **0.1597** | 0.0040 | 0.1047 |
| 10 | **0.7500** | 0.3603 | 0.2864 | 35 | **0.5913** | 0.0000 | 0.0000 |
| 11 | **0.6308** | 0.3981 | 0.4929 | 36 | 0.7942 | 0.7760 | **0.7968** |
| 12 | 0.4960 | **0.5784** | 0.5518 | 37 | 0.7678 | 0.7867 | **0.9104** |
| 13 | **0.0521** | 0.0090 | 0.0102 | 38 | **0.2548** | 0.0128 | 0.0536 |
| 14 | **0.0030** | 0.0000 | 0.0001 | 39 | **0.3339** | 0.0128 | 0.0120 |
| 15 | **0.3443** | 0.0257 | 0.0064 | 40 | 0.1544 | 0.0476 | **0.1735** |
| 16 | **0.3344** | 0.0000 | 0.0000 | 41 | 0.4048 | 0.1662 | **0.4221** |
| 17 | **0.2082** | 0.0113 | 0.0035 | 42 | **0.2245** | 0.0960 | 0.1155 |
| 18 | **1.0000** | 0.0000 | 0.0000 | 43 | 0.1066 | 0.0978 | **0.1207** |
| 19 | **1.0000** | 0.0385 | 0.0435 | 44 | 0.2642 | 0.0088 | **0.2666** |
| 20 | **0.2224** | 0.1474 | 0.1774 | 45 | **0.0092** | 0.0065 | 0.0077 |
| 21 | **0.4739** | 0.1794 | 0.0589 | 46 | 0.3064 | **0.3857** | 0.0327 |
| 22 | **0.0637** | 0.0157 | 0.0157 | 47 | **0.0743** | 0.0110 | 0.0340 |
| 23 | **0.4649** | 0.1500 | 0.2772 | 48 | **0.4610** | 0.0079 | 0.0036 |
| 24 | **0.9615** | 0.6724 | 0.9403 | 49 | **0.6546** | 0.2391 | 0.2204 |
| 25 | 0.0349 | **0.0681** | 0.0424 | 50 | **0.1751** | 0.0805 | 0.1531 |
|  |  |  |  | map | **0.3727** | 0.1452 | 0.1689 |

Table 2: Average precision of the best word (W), narrow (NC) and full concept (FC) runs (2004 collection)

results are caused by errors introduced by the concept tagging process: important concepts were missing from the thesaurus, and even after disambiguation many ambiguities remain.

### 6.2.2 2006 collection

For the 2006 collection we repeated the 2004 experiments. The 2006 topics do not consist of separate Title, Need and Context sections; each complete topic description is treated as a single query. The automatically concept tagged queries have been manually corrected to measure the influence of tagging errors. Table 3 and Figure 7 show the document MAP of the 2006 runs.

Again, the word based run shows the highest document MAP (0.36), comparable to the best MAP of the 2004 runs (also 0.36). The narrow and full concept runs perform much better compared to the 2004 runs. The best automatic run based on narrow concepts now achieves a MAP of 0.25 (vs. 0.14 for 2004); the best automatic run based on all concepts achieves a MAP of 0.27 (vs. 0.17

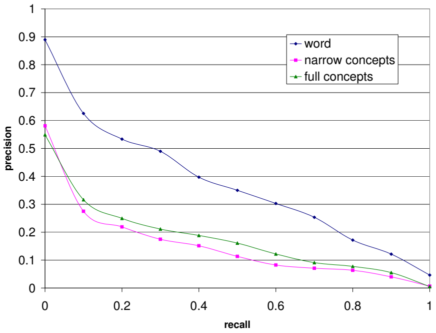


Figure 6: Interpolated recall precision curves for best word and concept runs (2004 collection)

| | words (W) | narrow concepts (NC) | | full concepts (FC) | |
|---|---|---|---|---|---|
| | | queries | | queries | |
| smoothing | | automatic | manual | automatic | manual |
| 0.1 | 0.2858 | 0.2325 | 0.2755 | 0.2386 | 0.2778 |
| 0.15 | 0.3010 | 0.2365 | 0.2846 | 0.2446 | 0.2855 |
| 0.2 | 0.3192 | 0.2396 | 0.2889 | 0.2484 | 0.2898 |
| 0.25 | 0.3260 | 0.2400 | 0.2946 | 0.2530 | 0.2997 |
| 0.3 | 0.3311 | 0.2400 | 0.2984 | 0.2574 | 0.3046 |
| 0.35 | 0.3375 | 0.2415 | 0.2995 | 0.2585 | 0.3082 |
| 0.4 | 0.3490 | 0.2414 | 0.3118 | 0.2592 | 0.3104 |
| 0.5 | 0.3543 | 0.2420 | 0.3169 | 0.2611 | 0.3248 |
| 0.6 | **0.3555** | 0.2461 | 0.3222 | 0.2631 | 0.3289 |
| 0.7 | 0.3526 | 0.2463 | 0.3285 | 0.2683 | 0.3394 |
| 0.8 | 0.3491 | 0.2468 | 0.3325 | 0.2713 | 0.3443 |
| 0.9 | 0.3423 | **0.2547** | **0.3383** | **0.2741** | **0.3523** |

Table 3: MAP of the best word (W), narrow (NC) and full concept (FC) runs (2006 collection)

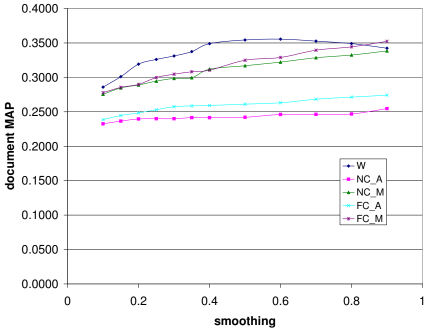


Figure 7: Document MAP for different smoothing values (2006 collection)

for 2004).

Manually correcting the concept tags of the topics also strongly improves the MAP of the concept runs. The best narrow concept run improves from 0.25 to 0.34 MAP (36% increase); the best full concept run improves from 0.27 to 0.35 (30% increase). The best manual concept runs come close to the best word based run. There is no significant [1] difference between these runs (also see Table 4).

However, the precision recall curve (Figure 8 shows that the word based runs achieve higher precision at lower recall levels: the top hits from these runs will more likely be relevant than the ones from the concept runs.

### 6.2.3 2006 abstract collection

Table 5 and Figure 9 show the document MAP of word and concept retrieval on the 2006 abstract collection. Both the concept and the word based approaches show a strong drop in retrieval performance compared to the runs with full-text documents (word based from 0.36 to 0.24, best concept based from 0.35 to 0.19). The concept retrieval approach seems to suffer most from shorter documents.

## 7 Conclusion

For the 2006 genomics task we have experimented with statistical language models based on concepts. On the 2004 collection, retrieval based on concepts (0.16 MAP) performed poor in

[1] Wilcoxon signed rank test, $p < 0.05$

| topic | W | NC | FC |
|---|---|---|---|
| 160 | **0.1562** | 0.0343 | 0.0417 |
| 161 | 0.3515 | 0.6852 | **0.7024** |
| 162 | **0.2989** | 0.2012 | 0.2199 |
| 163 | **0.5977** | 0.3530 | 0.3727 |
| 164 | 0.0010 | 0.6841 | **0.7198** |
| 165 | 0.1919 | 0.1701 | **0.2725** |
| 166 | **0.2783** | 0.0100 | 0.0090 |
| 167 | 0.5817 | 0.6420 | **0.6788** |
| 168 | 0.8026 | 0.8536 | **0.8632** |
| 169 | 0.3170 | 0.3500 | **0.4083** |
| 170 | 0.7260 | 0.7463 | **0.8056** |
| 171 | **0.0051** | 0.0008 | 0.0008 |
| 172 | **0.3127** | 0.1985 | 0.1854 |
| 174 | **0.4312** | 0.0754 | 0.1290 |
| 175 | **0.2840** | 0.0081 | 0.0092 |
| 176 | **0.1774** | 0.0293 | 0.0521 |
| 177 | **0.0000** | **0.0000** | **0.0000** |
| 178 | 0.3435 | 0.3507 | **0.3597** |
| 179 | 0.0419 | **0.0629** | 0.0617 |
| 181 | **0.6682** | 0.6199 | 0.6250 |
| 182 | **0.2614** | 0.1496 | 0.1553 |
| 183 | **0.2870** | 0.0612 | 0.0711 |
| 184 | 0.0000 | **0.5000** | **0.5000** |
| 185 | 0.6096 | 0.5918 | **0.6138** |
| 186 | 0.6017 | **0.6074** | 0.6037 |
| 187 | **0.9167** | 0.8095 | 0.6984 |
| map | **0.3555** | 0.3383 | 0.3523 |

Table 4: Average precision of the best word (W), narrow (NC) and full concept (FC) runs (2006 collection)

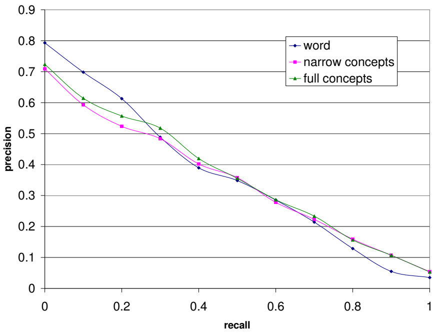


Figure 8: Interpolated recall precision curves for best word and concept runs (2006 collection)

|  | words (W) | narrow concepts (NC) |  |
|---|---|---|---|
|  |  | queries |  |
| smoothing |  | automatic | manual |
| 0.1 | 0.2324 | 0.1714 | 0.1785 |
| 0.15 | **0.2400** | 0.1715 | 0.1784 |
| 0.2 | 0.2391 | **0.1720** | 0.1785 |
| 0.25 | 0.2346 | 0.1696 | 0.1785 |
| 0.3 | 0.2315 | 0.1673 | 0.1915 |
| 0.35 | 0.2284 | 0.1653 | **0.1937** |
| 0.4 | 0.2292 | 0.1634 | 0.1930 |
| 0.5 | 0.2249 | 0.1659 | 0.1914 |
| 0.6 | 0.2183 | 0.1629 | 0.1910 |
| 0.7 | 0.2069 | 0.1577 | 0.1888 |
| 0.8 | 0.1985 | 0.1550 | 0.1871 |
| 0.9 | 0.1805 | 0.1509 | 0.1841 |

Table 5: Document MAP for different smoothing values (2006 **abstract** collection)

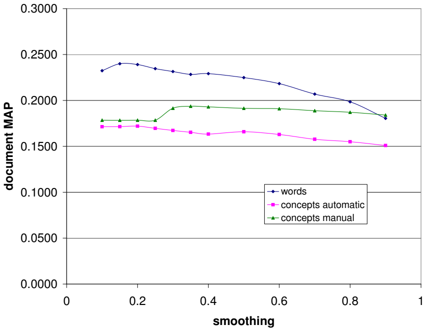


Figure 9: Document MAP for different smoothing values (2006 **abstract** collection)

comparison to our words baseline (0.35 MAP). On the 2006 collection, the same approach yields different results. Although document retrieval based on words (0.36 MAP) still outperforms the (manual) concept based approach (0.35 MAP), the difference in AP between the approaches is not significant.

Experiments on all three collection (2004, 2006 and 2006 abstracts) show that adding more concepts (full concepts in stead of narrow concepts) to our concept language model improves document retrieval performance.

The large improvement obtained from manually correcting automatically tagged topics suggest that (small) improvements in the concept tagger can have a strong impact on retrieval performance. This corresponds to our experience with different versions of the tagger.

Experiments on only the abstracts of the 2006 collection show large performance drops for both word and concept based approaches (words: 0.35 to 0.24; concepts: 0.36 to 0.19). The experiments confirm that searching full-text documents over abstracts improves retrieval performance.

We cannot conclude our approach based on a concept language model can outperform the classic word based language model. The approach seems sensitive to both collection and topics, and collection size. Further investigation has to show if perhaps a combination of concept and word language models can benefit information retrieval.

## 8 Acknowledgements

This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

## References

[1] TREC 2006 genomics track protocol. http://ir.ohsu.edu/genomics/2006protocol.html.

[2] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue):D267–D270, Jan 2004.

[3] Lifeng Chen, Hongfang Liu, and Carol Friedman. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21(2):248–256, Jan 2005.

[4] Erasmus Medical Center. Preprocessed TREC 2006 corpus. http://www.biosemantics.org/TREC2006/. Username and password same as for TREC data section.

[5] Djoerd Hiemstra and Wessel Kraaij. Twenty-one at TREC-7: Ad hoc and cross language track. In Ellen M. Voorhees and Donna K. Harman, editors, *The Seventh Text REtrieval Conference (TREC-7)*, volume 7. National Institute of Standards and Technology, NIST, 1999. NIST Special Publication 500-242.

[6] Djoerd Hiemstra and Wessel Kraaij. A language modeling approach for TREC. In Ellen M. Voorhees and Donna Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*. MIT press, 2005.

[7] A Koike and T Takagi. Gene/protein/family name recognition in biomedical literature. In *BioLINK 2004: Linking Biological Literature, Ontologies, and Databases*, pages 9–16. Association for Comutational Linguistics, 2004.

[8] Wessel Kraaij, Marc Weeber, Stephan Raaijmakers, and Rob Jelier. Mesh based feedback, concept recognition and stacked classification for curation tasks. In *Proceedings of TREC 2004*. NIST, 2005.

[9] A. T. McCray, S. Srinivasan, and A. C. Browne. Lexical methods for managing variation in biomedical terminologies. *Proc Annu Symp Comput Appl Med Care*, pages 235–239, 1994.

[10] Andrei Mikheev. Periods, capitalized words, etc. *Comput. Linguist.*, 28(3):289–318, 2002.

[11] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, New York, NY, USA, 1998. ACM Press.

[12] B.J.A. Schijvenaars, M.J. Schuemie, E.M. van Mulligen, M. Weeber, R. Jelier, B. Mons, J.A. Kors, and W. Kraaij. TREC 2005 genomics track a concept-based approach to text categorization. In *Proceedings of TREC 14*, 2006.

[13] Martijn J. Schuemie, Jan A. Kors, and Barend Mons. Word sense disambiguation in the biomedical domain: an overview. *Journal of Computational Biology*, 12(5):554–565, jun 2005.

[14] Martijn J. Schuemie, Barend Mons, Marc Weeber, and Jan A. Kors. Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification. *Journal of Biomedical Informatics*, In Press, 2006.

[15] Thijs Westerveld, Wessel Kraaij, and Djoerd Hiemstra. Retrieving web pages using content, links, urls and anchors. In *Proceedings of the tenth Text Retrieval Conference (TREC-10)*, pages 663–672. NIST Special Publication 500-250, 2002.

[16] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.