The Value of an In-Domain Lexicon in Genomics QA

Yutaka Sasaki* **John McNaught** **Sophia Ananiadou**
National Centre for Text Mining
School of Computer Science, University of Manchester
MIB, 131 Princess Street, Manchester, M1 7DN, UK
{Yutaka.Sasaki,John.McNaught,Sophia.Ananiadou}@manchester.ac.uk

Abstract

This paper demonstrates that a large-scale lexicon tailored for the biology domain is effective in improving question analysis for genomics *Question Answering (QA)*. We use the TREC Genomics Track data to evaluate the performance of different question analysis methods. It is hard to process textual information in biology, especially in molecular biology, due to a huge number of technical terms which rarely appear in general English documents and dictionaries. To support biological Text Mining, we have developed a domain-specific resource, the BioLexicon. Started in 2006 from scratch, this lexicon currently includes more than four million biomedical terms consisting of newly curated terms and terms collected from existing biomedical databases. While conventional genomics IR/QA systems provide query expansion based on thesauri and dictionaries, it is not clear to what extent a biology-oriented lexical resource is effective for question pre-processing for genomics QA. Experiments on the genomics QA data set show that question analysis using the BioLexicon performs slightly better than that using n-grams and the UMLS Specialist Lexicon.

1 Introduction

Recently, the focus of biological research has shifted from individual genes and proteins to entire biological systems (Jensen *et al.*, 2006). Due to this paradigm shift, domain experts now have to relate their experimental data to a number of conventional results already published in biological literature. In fact, biologists spend more than half of their research time surveying relevant publications.

This is because, in the biomedical domain, a large number of journal and conference papers are published every year. For example, more than 15 million MEDLINE abstracts have been published in the past with more than 600,000 abstracts being added annually. Researchers into the cell cycle need to read more than 7,000 new papers a year (Jensen *et al.*, 2006), which is beyond the limit of the number of papers that domain experts can carefully read.

In the light of this, clearly, biology is one of the fields that can benefit from information retrieval technology. Currently, most biologists use traditional search engines, such as PubMed, Google, or Google Scholar. It would be much more helpful for domain experts if they could find literature of interest using natural language questions since users are not sure about what kind of keywords are effective in order reach relevant documents quickly.

Question Answering (QA) has been actively studied since the start of the QA Track at TREC-8 (Voorhees, 1999). It started then as an evaluation campaign on retrieval of 50 and 250 byte passages from newspaper corpora. The research target shifted to questions that require the spotting of named entities (*e.g.*, "Who was the first Prime Minister of the UK?"). Recently, the target of QA studies has been widened to address more general questions, such as why and how questions.

Biomedical Question Answering is in its early stages in terms of the trend in general QA technology. Today, due to the lack of training and test data, biomedical QA can be evaluated as passage-based QA.

The TREC Genomics Track (Hersh *et al.*, 2007) has evaluated *passage retrieval* performance in the biomedical domain. Whereas TREC Genomics Tracks 2004 and 2005 targeted retrieval of MEDLINE abstracts, TREC Genomics Tracks 2006 and 2007 targeted passage-based QA based on full papers. In 2006, queries were generic, such as "What is the role of PrnP in mad cow disease?". As questions became entity-oriented questions in 2007, such as "What [GENES] regulate puberty in humans?", the 2007 data set is the most suitable for QA perfor-

Proceedings of the 3rd International Symposium on Languages in Biology and Medicine,
Long Papers, pages 47-55, Jeju Island, South Korea, 8-10 November 2009.

mance evaluations in this domain.

2 BioLexicon Overview

In this section, we provide a brief summary of the BioLexicon (Rebholz-Schuhmann *et al.*, 2008; Sasaki *et al.*, 2008). The BioLexicon is a collective achievement by EBML-EBI, CNR-ILC, and the University of Manchester in the EC BOOTStrep Project. The BioLexicon has been constructed in the following steps:

1. EBML-EBI collected biomedical terms from existing biomedical databases, such as ChEBI[1] and Gene Ontology[2]. The University of Manchester extracted new synonyms of gene/protein names from MEDLINE abstracts.

2. The University of Manchester manually curated terminological and general English verbs, adjectives, adverbs, and nouns. Inflections of general words are manually curated based on the MedPost dictionary (Smith *et al.*, 2004).

3. CNR-ILC generated verb subcategorization frames which are linked to semantic bio-event frames created by the University of Manchester.

4. CNR-ILC also devised the database model of the lexicon which follows the Lexical Markup Framework (LMF) (Francopoulo *et al.*, 2006).

2.1 Entries in the BioLexicon

The terminologies in the lexicon are fivefold:

(1) Terminological verbs (*e.g.*, *repress*): 759 base forms (4,556 inflections) of terminological verbs with automatically extracted verb subcategorization frames.

(2) Terminological adjectives (*e.g.*, *repressive*): 1,258 adjectives.

(3) Terminological adverbs (*e.g.*, *repressively*): 130 adverbs.

(4) Nominalized verbs (*e.g.*, *repression*): 1,771 nominalized verbs.

(5) Biomedical terms: Currently, the BioLexicon contains the following biomedical terms in each category:

- Cell (842 entries, 1,400 variants)
- Chemicals (19,637 entries, 106,302 variants)
- Enzymes (4,016 entries, 11,674 variants)
- Diseases (19,457 entries, 33,161 variants)
- Genes and proteins (1,640,608 entries, 3,048,920 variants)
- Gene ontology concepts (25,219 entries, 81,642 variants),
- Molecular role concepts (8,850 entries, 60,408 variants)
- Operons (2,672 entries, 3,145 variants)
- Protein complexes (2,104 entries, 2,647 variants)
- Protein domains (16,940 entries, 33,880 variants)
- Sequence ontology concepts (1,431 entries, 2,326 variants)
- Species (482,992 entries, 669,481 variants),
- Transcription factors (160 entries, 795 variants).

In addition to the collected gene/protein names, 70,105 new variants of gene/protein names have been extracted from 15 million MEDLINE abstracts.

3 Genomics QA Test Collection

3.1 Document collection

The corpus for the TREC Genomics Track 2007 is a collection of full papers obtained from 49 biomedical journals. The corpus is 13.3 GB in total and contains 162,259 full papers. Each of the full papers has a unique ID, called a PubMed ID (PMID).

The full papers are provided as HTML documents in which metadata are represented using formatting tags, unlike XML. Therefore, it is not straightforward to correctly analyze the organization of journal papers and to extract various kinds of metadata (such as author names and publication dates) embedded in HTML format. Many full papers in biomedicine are available only in HTML or PDF formats and the HTML formats differ from journal to journal.

Full papers have different characteristics from abstracts. As full papers contain general descriptions relevant to the targeted research topics, the range of contents in full papers is more broad than that found in abstracts.

3.2 Questions

The official test run questions for the Genomics Track 2007 ask about specific biomedical entities.

Targeted biomedical entities are as follows (Hersh *et al.*, 2007):

[1] `http://www.ebi.ac.uk/chebi/`
[2] `http://www.geneontology.org/`

ANTIBODIES Immunoglobulin molecules having a specific amino acid sequence by virtue of which they interact only with the antigen (or a very similar shape) that induced their synthesis in cells of the lymphoid series (especially plasma cells).

BIOLOGICAL SUBSTANCES Chemical compounds that are produced by a living organism.

CELL OR TISSUE A distinct morphological or functional form of cell, or the name of a collection of interconnected cells that perform a similar function within an organism.

DISEASES A definite pathologic process with a characteristic set of signs and symptoms. It may affect the whole body or any of its parts, and its etiology, pathology, and prognosis may be known or unknown.

DRUGS A pharmaceutical preparation intended for human or veterinary use.

GENES Specific sequences of nucleotides along a molecule of DNA (or, in the case of some viruses, RNA) which represent functional units of heredity.

MOLECULAR FUNCTIONS Elemental activities, such as catalysis or binding, describing the actions of a gene product or bioactive substance at the molecular level.

MUTATIONS Any detectable and heritable change in the genetic material that causes a change in the genotype and which is transmitted to daughter cells and to succeeding generations

PATHWAYS A series of biochemical reactions occurring within a cell to modify a chemical substance or transduce an extracellular signal.

PROTEINS Linear polypeptides that are synthesized on ribosomes and may be further modified, crosslinked, cleaved, or assembled into complex proteins with several subunits.

SIGNS OR SYMPTOMS A sensation or subjective change in health function experienced by a patient, or an objective indication of some medical fact or quality that is detected by a physician during a physical examination of a patient.

STRAINS A genetic subtype or variant of a virus or bacterium.

TOXICITIES A measure of the degree and the manner in which which something is toxic or poisonous to a living organism.

TUMOR TYPES An abnormal growth of tissue, originating from a specific tissue of origin or cell type, and having defined characteristic properties, such as a recognized histology.

Table 1: Classification of official questions

Entity type	# of questions
ANTIBODIES	1
BIOLOGICAL SUBSTANCES	3
CELL OR TISSUE	2
DISEASES	1
DRUGS	2
GENES	11
MOLECULAR FUNCTIONS	2
MUTATIONS	1
PATHWAYS	2
PROTEINS	5
SIGNS OR SYMPTOMS	2
STRAINS	1
TOXICITIES	2
TUMOR TYPES	1
Total	36

Table 1 shows the number of official questions for each entity type in the test set.

As an acceptable simplification of genomics questions for the first large-scale genomics QA evaluation, *Question types* and *question focuses* are explicitly given in question sentences, which makes question analysis much easier. Moreover, the forms of questions (unintentionally) fall into the following pattern:

```
{<WH>|<PREPOSITION> <WH>} <MODIFIER>*
[<ENTITY TYPE>] <WORD>+
```

where ⟨WH⟩ is "what" or "which", ⟨ENTITY TYPE⟩ is one of the entity types in Table 1 and ⟨MODIFIER⟩ is a noun, verb, or adjective phrase that restricts the range of entities in question.

This means that some natural variations of question styles, such as "What are the names of genes that regulate puberty in humans?" or "Could you name genes that regulate puberty in humans?", do not appear in the training and test data. As entity types are predefined, synonymic paraphrases, such as "What gene products are involved ...", do not appear in the Genomics Track questions whereas *gene product* can be used as a synonym of *protein*.

Due to this question style, the most crucial task of question analysis in the Genomics Track is to find query terms that are effective in finding passages relevant to questions.

3.3 Genomics QA Task

The task of the TREC Genomics Track 2007 was to retrieve passages from a full paper corpus and return a ranked list of at most 1,000 passages for each question.

Passages are defined as follows (Hersh *et al.*, 2007):

> Retrieved passages could contain any span of text that did not include any part of an HTML paragraph tag (i.e., one starting with ⟨P or ⟨/P).

In this paper, we report the results of retrieving passage spans with the maximum length.

Each passage can be identified by triples (PMID, offset, and length), where the offset is the starting position of the passage in a document in terms of the number of bytes from the top of the document.

3.4 Gold standard relevance judgement

A total of 66 runs was submitted by 29 groups. Each of the runs returns at most 1,000 passages for each question and judges with domain expertise manually checked the relevance of the submitted passages. Relevant passages are pooled in the gold standard. Sometimes the number of texts relevant to a question is very small. For example, the numbers of passages/documents relevant to Topic 224 and 225 are only three and one, respectively. This is a typical phenomenon in entity-oriented Question Answering. In general, topics for information retrieval evaluations are specifications of more general requirements (such as "the current situation of SARS") than entity-oriented question answering. In this paper, we explore what kind of keyword generation methods are advantageous for passage retrieval in genomics QA.

3.5 Evaluation metrics

Three kinds of *Mean Average Precision (MAP)* are officially used in the Genomics Track 2007 (Hersh *et al.*, 2007):

Passage2 MAP: The original Passage MAP for the 2006 track was found to be problematic in that splitting passages into shorter units had substantial positive effects on Passage MAP. To avoid this, Passage2 MAP calculates MAP as if each character in each passage were a ranked document.

Aspect MAP: Passages in the gold standard are grouped into aspects identified by one or more *Medical Subject Headings (MeSH)* terms. The aspect retrieval MAP is the average precision for the aspects of a topic, averaged across all topics.

Document MAP: Any document ID that had a passage associated with a topic ID in the set of gold standard passages was considered a relevant document for that topic.

4 Passage-based QA system

4.1 Passage retrieval methods

We adopted the probabilistic IR toolkit Xapian[3] as our retrieval platform. We have created two indexes, one for document retrieval and the other for passage retrieval. We use Xapian's built-in components, tokenizer and standard English stemmer. The stemmer performs simple word normalization, such as hyphen removal, but no complex normalization, such as Greek letter normalization, is applied to tokens. The employed IR model is a variant of Okapi BM25 (Robertson *et al.*, 1992).

$$\frac{(k_3+1)q}{(k_3+q)} \cdot \frac{(k_1+1)f}{(K+f)} \cdot \log \frac{(r+0.5)(N-n-R+r+0.5)}{(n-r+0.5)(R-r+0.5)}$$

where
k_1, k_3: constants
K: $k_1(bL+(1-b))$
q: within query frequency
f: within document frequency
n: the number of documents in the collection indexed by this term
N: the total number of documents in the collection
r: the number of relevant documents indexed by this term
R: the total number of relevant documents
L: the normalized document length

We used the default parameter setting, $k_1 = 1$, $k_3 = 1$, $b = 0.5$.

To capture local information in a passage and global characteristics of its full paper, retrieved passages are ranked by the score that is a weighted sum of BM25 scores of passage p and its full paper d.

$$BM25_{d,p} = \alpha\ BM25(p) + (1-\alpha) BM25(d),$$

The baseline passage retrieval algorithm to compare usefulness of lexical resources is as follows:

1. Analyze a question sentence using a dictionary-based Part-of-Speech (POS) tagger based on the BioLexicon, the UMLS Specialist Lexicon, or an n-gram collection.
2. Create a list of query terms from the question.
3. Retrieve N_d full papers using the query terms.
4. Retrieve N_p passages using the query terms.
5. Rerank the passages according to the $BM25_{d,p}$ score based on the scores of the retrieved documents and passages.

[3] http://xapian.org/

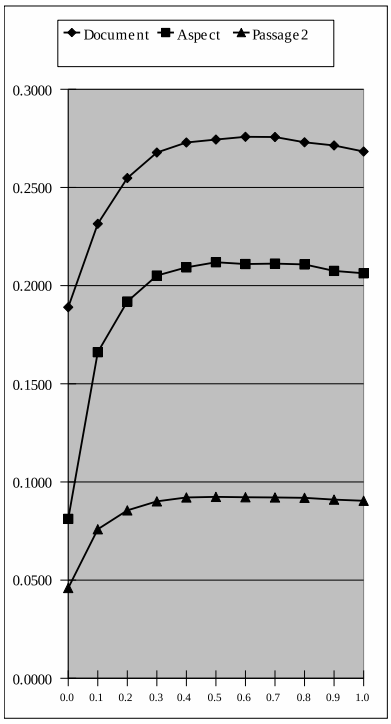


Figure 1: Trade-off of document and passage BM25 scores

6. Output the top 1,000 passages from the ranked passages.

For conciseness, we call technical terms that are extracted from a sentence based on lookup of the BioLexicon BL terms and terms extracted from a sentence based on lookup of the Specialist Lexicon SL terms. Here, N_d is set to 1,000. N_p is set to a large number, 1,000,000.

Since the goal of this paper is to estimate usefulness of lexical resources, first, we decide the value of the parameter α based on the baseline model without using external resources.

We remove stop words from a question, and the remaining words are used as query terms. The Document, Aspect, and Passage2 MAP are measured using the TREC Genomics Track 2007 test set. Figure 1 shows the results for $\alpha \in \{0.0, 0.1, 0.2, ..., 1.0\}$. As a result, we set $\alpha = 0.5$ which is the peak of the Aspect and Passage2 MAP curves.

4.2 Question Analysis

We compare the following question analysis methods.

[w1] Word uni-grams: The baseline question analysis method is to use tokens (i.e., uni-grams) which are not stop words, *e.g.,* "T-cell", "growth", and "factors".

[w12] Word uni- and bi-grams: In addition to w1, bi-grams of consecutive non-stop words are used, *e.g.,* "T-cell", "growth", "factors", "T-cell growth", and "growth factors".

[w123] Word uni-, bi-, and tri-grams: In addition to w12, tri-grams of consecutive non-stop words are used, *e.g.,* "T-cell", "growth", "factors", "T-cell growth", "growth factors", and "T-cell growth factors".

[b1] Lemma uni-grams: Uni-grams of lemmas (i.e., base forms) of tokens which are not stop words, *e.g.,* "T-cell", "growth", and "factor".

[b12] Lmma uni- and bi-grams: In addition to lemma uni-grams, bi-grams of consecutive lemmas of non-stop words, *e.g.,* "T-cell", "growth", "factor", "T-cell growth", and "growth factor".

[b123] Lemma uni-, bi-, and tri-grams: In addition to the above, tri-grams of consecutive lemmas of words which are not stop words, *e.g.,* "T-cell", "growth", "factor", "T-cell growth", "growth factor", and "T-cell growth factor".

[w1∪BL/SL] Multi-word terms and words that are not in BL or SL terms: If the multi-word terms in a question are BL or SL terms, the terms are added to the query term list. Then, word uni-grams that are not in the query term list are added to the query term list, *e.g.,* "T-cell growth factors".

[w1+BL/SL] Words and BL or SL terms: First, word uni-grams are added to the query term list. Then, if lexicon terms are found in a question, the terms are added to the query term list, *e.g.,* "T-cell", "growth", "factors", and "T-cell growth factors".

[b1∪BL/SL] Multi-word terms and lemmas that are not in BL/SL terms: If multi-word terms in a question are BL or SL terms, the terms are added to the query term list. Then, lemma uni-grams that are not in the BL terms are added to the query term list, *e.g.,* "T-cell growth factor".

[b1+BL/SL] Lemmas and BL or SL terms: First, word uni-grams are added to the query term list. Then, if multi-word terms in a question are BL or SL terms, these terms are added to the query term list, *e.g.,* "T-cell", "growth", "factor", and "T-cell growth factor".

Table 2: Experimental results on the TREC Genomics Track 2007 data

Method		Document MAP	Aspect MAP	Passage2 MAP
(a) TREC Top 6 (w.r.t. Doc MAP)	1st (NLMinter)	0.3286	0.2631	0.1148
	2nd (NLMfusion)	0.3105	0.2494	0.1097
	3rd (MuMshFd)	0.2906	0.2068	0.0895
	4th (MuMshFdRsc)	0.2880	0.2079	0.0893
	5th (UniNE1)	0.2777	0.2189	0.0988
	6th (UniNE3)	0.2710	0.2043	0.0970
(b) TREC statistics (automatic run)	Min	0.0329	0.0197	0.0008
	Median	0.1954	0.1272	0.0391
	Mean	0.1891	0.1286	0.0421
	Max	0.3105	0.2494	0.1097
(c) n-gram	w1	0.2744	**0.2119**	0.0924
	w12	0.2257	0.1955	0.0760
	w123	0.2156	0.1697	0.0737
	b1 (BL)	0.2272	0.1773	0.0768
	b12 (BL)	0.2190	0.1674	0.0688
	b123(BL)	0.2137	0.1601	0.0666
	b1 (SL)	0.2483	0.1811	0.0765
	b12 (SL)	0.2217	0.1707	0.0650
	b123(SL)	0.2134	0.1514	0.0628
(d)BL	w1∪BL	0.2747	0.2069	0.0923
	w1+BL	**0.2763**	0.2018	**0.0931**
	b1∪BL	0.2274	0.1717	0.0766
	b1+BL	0.2369	0.1668	0.0779
(e) SL	w1∪SL	0.2665	0.1967	0.0855
	w1+SL	0.2759	0.1959	0.0887
	b1∪SL	0.2440	0.1667	0.0722
	b1+SL	0.2536	0.1637	0.0755

5 Experiments

Experiments on the TREC Genomics Track 2007 data have been conducted with the different question analysis methods described in the previous section.

Table 2 shows the results. The top 6 official runs of the TREC Genomics Track 2007 are presented in Table 2 (a). Table 2 (b) contains statistics of automatic runs. Table 2 (c), (d), and (e) show the results from testing n-grams, the BioLexicon, and the Specialist Lexicon.

The best document MAP is 0.2763 when the BL terms are added to query word uni-grams. It is clear that adding bi-grams and/or tri-grams generated from a noun, verb, and adjective phrases is not effective, as the MAP scores decrease to 0.2257 and 0.2156.

The Document MAP of the queries consisting of the Specialist Lexicon terms and word uni-grams is 0.2759, which is better than the n-gram-based approach but not as good as the BioLexicon-based query analysis.

The best Passagae2 is 0.0931 when the BL terms are added to query word uni-grams whereas the best Aspect MAP is 0.2119 when a set of word uni-grams is used as a query.

These best performances are comparable to the performances of the top 6 automatic IR systems in the TREC Genomics Track 2007.

6 Discussion

Figure 2 shows the Document, Aspect, and Passage2 MAP scores of w12, w123, w1+BL, w1+SL per question. Simply adding bi-grams and tri-grams is mostly disadvantageous. There is only a little difference between w1+BL and w1+SL approaches in terms of the Document MAP, but this is not the case for the Passage2 MAP.

As stated before, the topics of the TREC Genomics Track 2006 are more IR-oriented queries than QA questions. Due to this, the top performing systems are different when comparing 2006 and 2007. When we applied the same query pre-processing to the 2006 data, we found different trends than those of the experimental results. Table 3 shows that using BL-terms with lemmas is the best way for the Document MAP. The best Passage2 MAP was ob-

Table 3: Experimental results on the TREC Genomics Track 2006 data

Method		Document MAP	Aspect MAP	Passage2 MAP
(c) n-gram	w1	0.3206	0.1746	0.0273
	w12	0.2852	0.1272	0.0175
	w123	0.2771	0.1172	0.0166
	b1 (BL)	0.3301	0.1838	**0.0284**
	b12 (BL)	0.2902	0.1216	0.0176
	b123(BL)	0.2781	0.1144	0.0168
	b1 (SL)	0.3293	0.1859	0.0250
	b12 (SL)	0.3094	0.1276	0.0167
	b123(SL)	0.2986	0.1221	0.0159
(d)BL	w1∪BL	0.3291	0.1838	0.0238
	w1+BL	0.3233	0.1641	0.0214
	b1∪BL	**0.3367**	0.1904	0.0249
	b1+BL	0.3311	0.1729	0.0225
(e) SL	w1∪SL	0.3105	**0.1917**	0.0260
	w1+SL	0.3075	0.1539	0.0208
	b1∪SL	0.3255	0.1778	0.0233
	b1+SL	0.3222	0.1498	0.0198

tained when we used lemmas of question terms as query terms.

7 Related work

The top 6 official runs are generated by the following approaches. Demner-Fushman *et al.* (2007) experimented with three models: an interactive model (NLMinter), a fusion model (NLMfusion), and a knowledge-based model (LHNCBC). Two of them, NLMinter and NLMfusion, achieved excellent performance. NLMinter used *manually* constructed queries consisting of a conjunction of topic terms and additional terms. NLMfusion is the equally-weighted fusion of the results of four automatic IR methods. Whereas LHNCBC attempted to exploit semantic types and synonyms, the performance was not comparable to the leading runs. Both MuMshFd and MuMshFdRsc employ an automatic query expansion with entities and ontological terms (Stokes *et al.*, 2007). In addition, MuMshFdRsc applies passage reduction and re-ranking. UniNE1 is a retrieval system based on Divergence from Randomness with WordNet (Fellbaum, 1998) expansions and UniNE3 is a fusion of three IR models incorporating WordNet expansions (Fautsch and Savoy, 2007).

In this paper, we investigated whether using an in-domain dictionary is meaningful in the IR tasks. Although it is potentially effective to use external resources to gain new information, in reality, however, it is well known that query expansion could degrade IR performance as described in Jimeno and Pezik (2007).

This paper shows that adding technical terms to query terms improves Document and Passage2 MAP. Constructing resources is very costly and time consuming, which requires long-term steady efforts. Whereas our biology lexicon is constructed independently from the track, it has been shown that the lexicon provides technical terms that effectively improve genomics IR performance.

8 Conclusion and Remarks

This paper reveals results that show that a large-scale lexicon tailored for the biology domain is effective in question analysis for genomics Question Answering. TREC Genomics Track data were used to evaluate the effect of various question analysis methods. Experiments on the genomics QA data set show that question analysis using the BioLexicon performs slightly better than that using n-grams and the UMLS Specialist Lexicon.

Our future work includes applying the BioLexicon to other parts of QA, such as answer spotting and learning to rank. The BioLexicon is available from the ELRA catalogue (ref T0373)[4].

9 Acknowledgements

This research is partly supported by EC IST project FP6-028099 (BOOTStrep) and the JISC sponsored National Centre for Text Mining. The authors also thank anonymous reviewers for their helpful comments.

[4] http://catalog.elra.info/product_info.php?products_id=1113

References

Dina Demner-Fushman, Susanne M. Humphrey, C. Ide, Russel F. Loane, James G. Mork, Patrick Ruch, Miguel E. ruiz, Lawrence H. Smith, W. John Wilbur, Alan R. Aronson, Combining resources to find answers to biomedical questions, Proc. of *TREC-16, NIST Special Publication*, 2007.

Claire Fautsch, Jacques Savoy IR-Specific Searches at TREC 2007: Genomics & Blog Experiments, Proc. of *TREC-16, NIST Special Publication*, 2007.

Francopoulo, G., M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria, Lexical Markup Framework (LMF), Proc. of *LREC 2006*, Genova, Italy, 2006.

Fellbaum, C., editor, *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA., 1998.

William Hersh, Aaron Cohen, Lynn Ruslen, Phoebe Roberts, TREC 2007 Genomics Track Overview, Proc. of TREC 2007, 2007.

Jensen, L. J., Saric, J., and Bork, P., Literature mining for the biologist: from information retrieval to biological discovery, Nat. Rev. Genet. 7 (2006), pp. 119-129.

Antonio Jimeno and Piotr Pezik, Information Retrieval and Information Extraction in TREC Genomics 2007, Proc. of *TREC-16, NIST Special Publication*, 2007.

McCray, A.T., Srinivasan, S. and Browne, A.C., Lexical methods for managing variation in biomedical terminologies, SCAMC'94, pp. 235-239, 1994.

Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau, Okapi at TREC, Proc. of *Text REtrieval Conference*, pp. 21-30, 1992.

Dietrich Rebholz-Schuhmann, Piotr Pezik, Vivian Lee, Jung-Jae Kim, Riccardo del Gratta, Yutaka Sasaki, Jock McNaught, Simonetta Montemagni, Monica Monachini, Nicoletta Calzolari, and Sophia Ananiadou, BioLexicon: Towards a Reference Terminological Resource in the Biomedical Domain, *16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB-2008)* (Poster), Toronto, Canada, 2008.

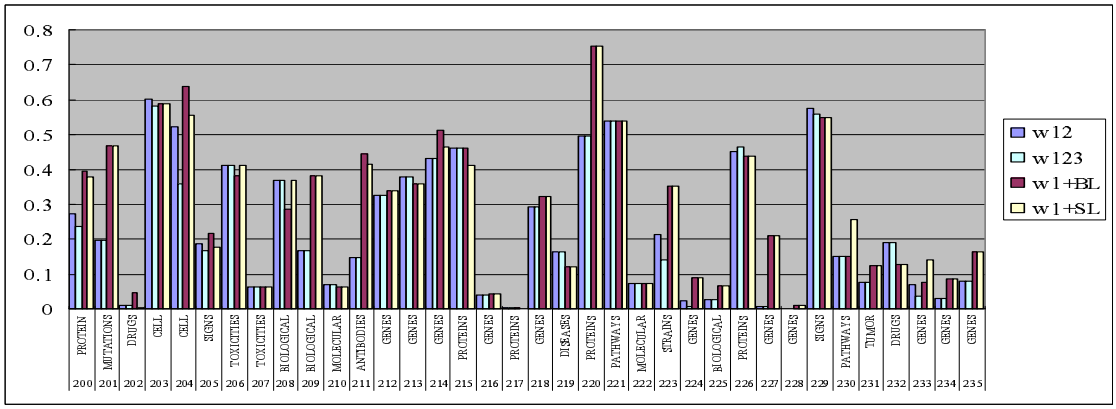
Yutaka Sasaki, Simonetta Montemagni, Piotr Pezik, Dietrich Rebholz-Schuhmann, John McNaught, and Sophia Ananiadou, BioLexicon: A Lexical Resource for the Biology Domain, Proc. of the *Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, 2008.

Smith, L., T. Rindflesch, and W. J. Wilbur, MedPost: a Part-of-Speech Tagger for BioMedical Text, *Bioinformatics*, 20:2320-2321, 2004.

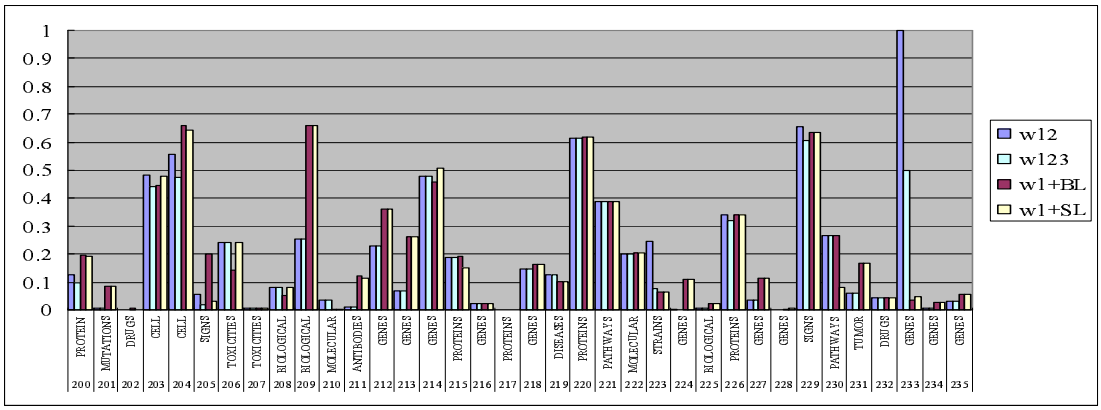
Nicola Stokes, Yi Li, Lawrence Cavedon, Eric Huang, Jiawen Rong and Justin Zobel, Entity-based Relevance Feedback for Genomic List Answer Retrieval, Proc. of *TREC-16, NIST Special Publication*, 2007.

Ellen M. Voorhees, The TREC-8 Question Answering Track Report, Proc. of *Eighth Text REtrieval Conference (TREC-8)*, pp. 77-82, 1999.

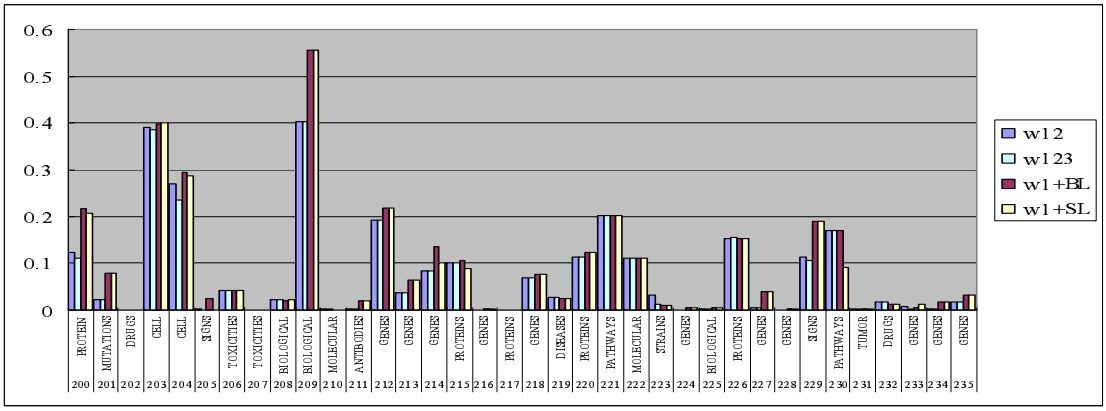
Figure 2: Per question MAP scores of word uni/bi-grams (w12), word uni/bi/tri-grams (w123), word uni-gram + BL-terms (w1+BL), word uni-grams + SL-terms (w1+SL)



(a) Document MAP



(b) Aspect MAP



(c) Passage2 MAP