Notebook Paper TREC 2005 Genomics Track

A Concept-Based Approach to Text Categorization

Bob J.A. Schijvenaars, Martijn J. Schuemie, Erik M. van Mulligen, Marc Weeber, Rob Jelier, Barend Mons, Jan A. Kors

Department of Medical Informatics
Erasmus University Medical Center, Rotterdam, The Netherlands

Wessel Kraaij
TNO, Delft, The Netherlands

Abstract

The Biosemantics group (Erasmus University Medical Center, Rotterdam) participated in the text categorization task of the Genomics Track. We followed a thesaurus-based approach, using the Collexis indexing system, in combination with a simple classification algorithm to assign a document to one of the four categories. Our thesaurus consisted of a combination of MeSH, Gene Ontology, and a thesaurus with gene and protein symbols and names extracted from the Mouse Genome Database, Swiss-Prot and Entrez Gene. To increase the coverage of the gene thesaurus, several rewrite rules were applied to take possible spelling variations into account.

Each document in the training set was indexed and the found concepts were ranked on term frequency, resulting in one concept vector per document. No particular care was taken to resolve ambiguous terms. For each of the four categories, two average concept vectors were computed, one by averaging the concept vectors of the documents in that category and the other by averaging all remaining concept vectors. The latter vector was then subtracted from the first, yielding a final category concept vector. The subtraction served to emphasize distinguishing concepts: high-ranked concepts in the final concept vector should, on average, occur relatively frequently in documents belonging to the category, while occurring infrequently or not at all in documents not belonging to the category.

For all documents in the training set, a matching score between the concept vector of a document and each of the category concept vectors was computed. A score threshold to discriminate between category and non-category documents was then determined per category by optimizing the performance measure (normalized utility). Different matching algorithms and different cutoffs for the number of concepts in the category vectors were evaluated. A standard cosine similarity score and a category vector with the 40 highest-ranking concepts proved to perform best on the training set. These settings and the score thresholds were subsequently used to categorize all documents in the test set.

Two runs were submitted: one based on the full text without any special treatment of particular sections, and one based on the Medline abstract, including the title and the MeSH headings.

In addition two runs were submitted by TNO for the ad-hoc search task. The ad-hoc system was based on the TREC 2004 system, with a small experiment trying to leverage information about the authority level of specific journals.

Introduction

Erasmus MC and TNO continued their collaboration concerning TREC. Like last year, Erasmus MC concentrated on the categorization task, whereas TNO concentrated on the ad-hoc search task. Erasmus MC participated in the categorization task of the Genomics track of TREC 2005 to determine the performance of a categorization approach based on concept vectors, utilizing straightforward information retrieval techniques and a simple combination of domain-specific thesauri. TNO investigated whether prior knowledge about the authority level

of publications could improve retrieval effeciveness. The working hypothesis here was that it might be beneficial to boost the retrieval score of “good” journals or journals from particular countries.

Categorization task

Methods

Indexing

Text documents were indexed with the Collexis indexing system (Geldermalsen, The Netherlands; http://www.collexis.com). For a given text, frequently occurring non-informative words are removed and the remaining terms are stemmed (using the LVG software which is part of the UMLS lexical tools: http://umlslex.nlm.nih.gov/lvg/current/). Subsequently, the document is searched for biomedical terms that occur in a thesaurus. Each found term is mapped to a unique identification code that denotes the preferred term, or concept, t_i and is assigned a relevance score or weight w_i that equals the term frequency normalized to the maximum term frequency of all terms in the document. A document can thus be represented by an M-dimensional vector $W = (w_1, w_2, \ldots, w_M)$, where M is the number of distinct concepts in the thesaurus, and $w_i = 0$ if t_i is not in the document. This weight vector W will subsequently be called the “concept vector” (CV) of the document, and is used for subsequent processing.

Thesaurus

The thesaurus used by the indexing system is a combination of the MeSH thesaurus (http://www.nlm.nih.gov/mesh), Gene Ontology (http://www.geneontology.org), and a thesaurus with gene and protein symbols and names extracted from the Mouse Genome Database (http://www.informatics.jax.org), Swiss-Prot (http://www.expasy.org) and Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=gene). No effort was made to detect and correct for terms occurring in more than one of these thesauri. To increase the coverage of the gene and protein thesaurus, several rewrite rules were applied to take possible spelling variations into account. Roman numerals were replaced with numbers and vice versa, hyphens were added or deleted between letters and numbers at the end of the gene and protein symbols, and several letters occurring at the end of a symbol were replaced with a possible long form (e.g. ‘a’ was replaced with ‘alpha’, 'r' was replaced with 'receptor'). Several gene synonyms such as ‘lobe’ and ‘peripheral’ were found in a large number of documents due to ambiguity, and were manually removed from the thesaurus. The final combination of thesauri contained about 98,000 concepts, and 270,000 distinct terms.

Document types

Concept vectors were generated for three types of content: (1) title and abstract of the articles as present in Medline, (2) title, abstract and MeSH headings as present in Medline, and (3) full-text articles in the TREC download package.
Tests to determine the performance on the training set were performed for all three types of content.

Categorization

For each category a “category concept vector” (CCV) was determined using the training set data. This was done by taking all documents in a specific category and averaging the corresponding vectors, resulting in a “positive” concept vector, CCVpos. The same was done for the remaining documents, yielding a “negative” concept vector CCVneg. The CCV was then computed by subtracting CCVneg from CCVpos, and keeping the N concepts with the highest rank. The objective of this procedure was to have the concepts with the most distinguishing power in the CCV.
We experimented with various values of N (20, 30, 40, 80) and found that the best results (highest normalized utility) were obtained for N=40 or N=80. We then settled for N=40.
The four category vectors obtained in this way were matched against all training document vectors using a standard cosine similarity matching algorithm [Salton1989]. A number of matching algorithms were tested (cosine similarity, Jaccard, Dice, and one that uses the dot

product of the two concept vectors and corrects using the number of overlapping concepts rather than the vector lengths as is the case in cosine similarity). The standard cosine similarity measure proved to give the best results. We also studied the effect of adjusting the weights with the inverse document frequency [Salton1988]. The effect per category was small, and did not increase the performance systematically.
For each category a matching score threshold T was determined. Documents with a matching score higher than T were considered to belong to the corresponding category, those with a lower score were not. The value of T was optimized by maximizing the normalized utility measure on the training set. This normalized utility U_{norm} is defined in the TREC 2005 protocol as

$$U_{norm} = \frac{u_r * TP - FP}{u_r * (TP + FN)}$$

where u_r is the relative utility of a relevant document in a particular category. Values of u_r are given in the TREC protocol and depend on the number of documents in a category (Table 1)

Table 1. Number of documents in the four categories for training set (N=5837) and test set (N=6043), and the corresponding relative utility parameters.

Category	Training set	Test set	u_r
Alelle	338	332	17
Expression	81	105	64
Go	462	518	11
Tumor	36	20	231

The category concept vectors and the thresholds derived from the training set were then applied to the test set.

Results

Using the training set, the performance was determined for different document types (Table 2). To gauge the value of a concept-based approach as opposed to a simple word-based indexing approach, performance was also computed when all words in the documents were indexed instead of using the thesaurus.

Table 2. Normalized utilities for various experimental settings.

		Normalized utility per category			
Documents	Thesaurus	Allele	Expression	Go	Tumor
Abstracts	Yes	0.754	0.574	0.420	0.906
Abstracts + MeSH headings	Yes	0.774	0.677	0.506	0.904
Full articles	Yes	0.835	0.636	0.469	0.880
Abstracts	No	0.757	0.654	0.351	0.917
Abstracts + MeSH headings	No	0.812	0.642	0.472	0.903
Full articles	No	0.734	0.421	0.248	0.802

When the thesaurus is used, abstracts+MeSH heading perform best, except for the Allele category where full articles score higher; abstracts alone clearly perform worse for the Expression and Go categories. When the concept-based approach is compared with the word-based approach, concept-based performs better for abstracts+MeSH headings for three of the four categories, and clearly outperforms word-based for the full articles; the comparison for abstracts alone gives mixed results.
Based on these results we decided to submit two runs on the test set, for concept-based indexing of the abstracts+MeSH headings and of the full articles. Unfortunately, after submission we discovered a bug that made the results of our full-text submission invalid. More detailed results are therefore only shown for our submission based on abstracts+MeSH headings (Tables 3 and 4).

Table 3. Results for concept-based indexing of abstracts+MeSH headings on the training set

Subtask	Precision	Recall	F-score	Norm Util
Allele	0.273	0.917	0.421	0.774
Expression	0.071	0.852	0.131	0.677
Go	0.219	0.749	0.338	0.506
Tumor	0.058	0.972	0.109	0.904

Table 4. Results for concept-based indexing of abstracts+MeSH headings on the test set

Subtask	Precision	Recall	F-score	Norm Util
Allele	0.241	0.892	0.380	0.726
Expression	0.084	0.819	0.153	0.680
Go	0.218	0.726	0.335	0.489
Tumor	0.031	0.950	0.061	0.823

The results on the training and test set are comparable, indicating that no overtraining took place in deriving the category concept vectors. Precisions are low, especially for those categories with a large relative utility parameter (Expression and Tumor). This is due to the fact that the categorization threshold was optimized for the normalized utility.

Discussion

Table 5 shows the median performance results of the runs submitted by all participants in the categorization task. Our results for the normalized utility are above the median for 3 of the 4 categories, but precision and F-score are lower than median for 3 of the categories. This is probably due to the fact that we optimized our categorization threshold for normalized utility only.

Table 5: Median results of all runs submitted by participants in the categorization task. Bold values indicate results where our system performed better than the median.

Category	Precision	Recall	F-score	Norm Util
Allele	0.3582	0.8946	0.5070	0.7785
Expression	0.1228	0.8190	0.1994	**0.6548**
Go	**0.2102**	**0.6506**	**0.3185**	**0.4575**
Tumor	0.0526	**0.9000**	0.0952	**0.7610**

Study of the category concept vectors showed that the concept “Mice” appeared as top concept in each of the four CCVs. This suggests that the documents in the training set that did not belong to any category, mainly consisted of non-mouse documents. A simple check confirmed this: the concept vectors of 94% of the documents that were assigned to one or more categories, contained the concept “Mice”. This was the case for only 35% of the documents that were not assigned to any category. Several highly unspecific concepts like “role” or “development” also ended up high in the CCV for some categories. Using an inverse document frequency correction however, did not improve the results on the training set. This needs to be investigated further.

Ad Hoc task

Method

This year's ad-hoc task in the genomics track was modeled slightly differently than classical ad-hoc tasks in the TREC tradition in the sense that all information needs were instantiations of one out of five "generic topic templates" (GTTs). Since the concepts that populate these templates refer to abstract concepts from the domain of genomics, TNO investigated whether the MeSH annotations of Medline articles could be exploited in order to improve search results. However, no straightforward one-to-one mapping to a MeSH term was found for all concepts in the GTTs.

We therefore explored instead whether our TREC2004 baseline full-text system based on language modeling techniques could be improved by looking at external indicators that might give a clue about the prior probability of relevance. We hypothesized for example that "important journals" have a higher prior probability of relevance than "low impact journals" or journals from certain less-advanced countries. For the combination of this "prior knowledge" and the retrieval score due to the query itself, we applied the same method as in [Kraaij2002] where we successfully exploited prior knowledge about a web document being a home page using link structure and URL length.

The basic TNO approach to retrieval tasks is based on generative language models (cf. [Hiemstra2005] for an overview of our TREC work using language models). In TREC2004, this fairly general approach yielded competitive results [Kraaij2005]. The model can be formalized as follows:

$$\log \frac{P(R \mid D,Q)}{P(\overline{R} \mid D,Q)} = \sum_{q_i \in Q} (\log \frac{\lambda P(q_i \mid R,D) + (1-\lambda) P(q \mid C)}{P(q_i \mid C)}) + \log \frac{P(R \mid D)}{P(\overline{R} \mid D)}$$

Where D denotes a document, Q a query,C a collection model, R is the binary relance variable, λ is a smoothing constant and q_i refers to the individual query terms.

This formula can be paraphrased as follows: the log-odds of relevance are composed of a query-dependent component (based on the difference between two cross-entropies) and a query-independent component: the prior odds of being relevant. In the actual experiments, the model was simplified by dropping the denominator of the query-specific term (since it does not contain document-specific elements) and the denominator of the prior odds (since it is close to 1). Queries were generated from the GTTs by deleting the template verbiage and applying a standard stoplist.

We computed the prior probability of relevance P(R|D) for each individual journal by computing the marginal probability of each journal in the set of relevant documents from the 2004 ad-hoc search task, smoothed by Lidstone's law [Manning1999].

Results

Two runs were submitted: tnog10, our baseline system, and tnog10p, using the journal based prior.

Run tag	Mean average precision
Tnog10	0.2346
Tnog10p	0.2332

Comparison with the results of other systems shows that most topics score above median, but that the system is less competitive than last year.

Unfortunately, adding prior knowledge about journals does not prove beneficial. This is probably due to the small set of relevance judgements in comparison with the number of unique journals in the Medline data collection.

Conclusion

The approach of applying general information retrieval techniques to vectors generated with a domain specific thesaurus appears promising. The use of category concept vectors that

describe the distinctive concepts within a category works quite well. Without any special finetuning techniques our results proved to be above the median performance of all the submitted results. An additional advantage of category concept vectors is that they can convey to users the category content without the need to present example documents, allowing users to evaluate and modify these vectors for their specific needs. Due to the optimization with respect to the normalized utility, precision values were low. It is therefore debatable whether the normalized utility is a useful measure to gauge text categorization results in daily practice.
Future improvements could be the application of additional information retrieval techniques like other matching algorithms or ways to determine the category vectors. Additional attention should be spent to the construction of the thesaurus which, due to the overlap of the parts used to construct it, now contains "technical" homonyms, i.e., identical terms from different parts denoting the same concept but not generating the same concept id. Disambiguation techniques could also have some influence [Schuemie2005, Schijvenaars2005], since the homonym problem in genomics research is large [Weeber2003].
Experiments with the use of global knowledge about the quality of different journals did not yield improvements for the ad-hoc task. This is probably due to the small size of the training collection.

References

[Hiemstra2005]
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.

[Kraaij2002]
W. Kraaij, T. Westerveld, and D. Hiemstra. The Importance of Prior Probabilities for Entry Page Search. In M. Beaulieu, R. Baeza-Yates, S. H. Myaeng, and K. Järvelin, editors, *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, pages 27-34, ACM Press, 2002.

[Kraaij2005]
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.

[Manning1999]
C. Manning and H Schütze. Foundations of Statistical Natural Language Processing, *MIT Press*, 1999.

[Salton1989]
Salton, G. *Automatic text processing: The transformation, analysis, and retrieval of information by computer*, Addison-Wesley, Reading, MA, 1989.

[Salton1988]
Salton, G. A simple blueprint for automatic boolean query processing, *Information Processing & Management,* Vol. 24, No. 3, pp. 269-280, 1988.

[Schuemie2005]
Schuemie M.J., Kors J.A., Mons B., Word sense disambiguation in the biomedical domain: an overview. *Journal of Computational Biology*, Jun, 12:554-65, 2005.

[Schijvenaars2005]
Schijvenaars B.J., Mons B., Weeber M., Schuemie M.J., van Mulligen E.M., Wain H.M., Kors J.A., Thesaurus-based disambiguation of gene symbols. *BMC Bioinformatics,* Jun, 16, 6:149, 2005

[Weeber2003]
Weeber M., Schijvenaars R.J.A., van Mulligen E.M., Mons B., Jelier R., van der Eijk C.C., Kors J.A., Ambiguity of Human Gene Symbols in LocusLink and MEDLINE: Creating an Inventory and a Disambiguation Test Collection. *Proceedings of AMIA Symposium,* 704-708, 2003