# Learning Domain-Specific Knowledge from Context --THUIR at TREC2005 Genomics Track

Jiao Li, Xian Zhang, Yu Hao, MinLie Huang, Xiaoyan Zhu*

State Key Lab of Intelligent Technology and Systems (LITS), Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

*Contact author: zxy-dcs@tsinghua.edu.cn

## Abstract

We(Tsinghua University) participated both Ad Hoc Retrieval Task and Categorization Task in TREC2005 Genomics Track, in which we designed and implemented a serious of methods encompassed learning domain-specific knowledge from context. In Ad Hoc Retrieval Task, internal resource is introduced to expand query, different granularity indexing provides more flexible retrieval space, and pattern discovering imports Information Extraction (IE) concept into Information Retrieval (IR). In Categorization Task, instead of the single word feature, we presented Seed-based Loose N-gram Feature, which achieved success in the four subtasks.

## 1. Introduction

Based on the experience of TREC2004 Genomics Track, in this year, our research focused on taking advantage of document collection itself to mine useful information, contributing to Information Retrieval (IR) and Text Categorization (TC) in Genomics.

Ad Hoc retrieval task is to search relevant documents from 4,591,008 Medline citations, about the 50 topics which are organized in 5 generic topic templates (GTTs). The semantic expression makes each genomic entity in the topic more senseful, especially in the topics (110-149). In this task, we engaged in finding more knowledge about the genomic entities and the relations between them to enhance the retrieval performance.

Categorization Task, a traditional TC problem, is to classify full-text articles from three journals for four categories (**A**lleles of mutant phenotypes, **T**umor biology, **E**mbryologic gene expression, and **G**O annotation). We made a thorough modification to the feature selection of our classification system: a new type of feature, which contains more semantic information, is proposed, and to generate this feature, a new recursive incremental method is designed.

The rest of this paper is organized as follow: Section 2 and Section 3 indicate the methods we employed, and the results we got in two tasks of TREC2005 Genomics respectively, and Section 4 contains the conclusions and future works.

## 2. Ad Hoc Retrieval Task

### 2.1 Task Overview

In this task, the retrieval object is the 10-year MEDLINE subset, each record of which contains necessary bibliographical information such as <AuthorList>, <JournalIssue>, <PubDate> and nearly 60 other fields, however, not all the data are related with the task. Thus, it's needed to

filter non-text information in document collection before index, ensuring our index as clean and slim as possible. Finally, the remained fields involve <PMID>, <ArticleTitle>, <AbstractText>, <MeSHList> and the other two 3 fields, moreover, we split the content of <AbstractText> into much smaller and senseful units, i.e. UniSen and BiSen, to support different granularity retrieval (See Section 2.3).

The 50 structured topics can be expressed within the biomedical entity tagged, for example, topic110 can be expressed as following:

**Example1:** <110> Provide information about the role of the gene <Gene>Interferon-beta</Gene> in the disease <Disease>Multiple Sclerosis</Disease>.

Therefore, “Interferon-beta” and “Multiple Sclerosis” become the main elements in the query space. To reduce the dismatch between query and document, internal resource for the query expansion purpose is introduced (See Section 2.2). Four of the five GGTs emphasize the relationship among two (or three) entities, which may co-occur in one document within some scope and some pattern, respectively, we present the different granularity retrieval (section 2.3) and pattern extraction and match (Section 2.4).

### 2.2 Internal Resource

In the domain of biomedical publication, synonyms and homonyms are omnipresent and post a great challenge for document retrieval systems. Many works contributed to query expansion, through integrating biological database (MeSH, LocusLink, AcroMed) and pseudo-relevance feedback [2], with the purpose of finding the actual form of query in the document. Our method in query expansion, named internal resource, is try to extract a glossary for 10-year subset of Medline data, just as most biological journal articles have a section named glossary. And the extracted internal resource would be used for expanding queries.

Different from some dictionary based on MEDLINE [1], the extraction of our internal resource is trigged by query term. After collecting sentence candidates, sentence containing the query term, pattern detector and rule matcher are applied to them, and the extracted glossary may be the strict or the loosen one according to the user configuration. Take topic125 for instance, the median of MAP in this topic is 0.0000, which means most groups failed to find relevant documents about it. Our group found 9 documents out of 11 in qrels.

**Example2:** <125> Provide information on the role of the gene <Gene>Nurr-77</Gene> in the process of <BiologicalProecess>preventing auto-immunity by deleting reactive T-cells before they migrate to the spleen or the lymph nodes</BiologicalProcess>.

Our success on this topic attributes to the internal resource, which finds the more common format (Nur-77) of the gene Nurr-77 in the collection.

### 2.3 Different Granularity Retrieval

To describe the relationship between two (or more) entities, the positions of them in one document could not be far from each other. We calculate the cooccurrence rate of sample topic entities within one sentence, a senseful unit in the abstract text (see Table1), the results of which prove the above assumption, with average cooccurrence rate 0.5 in DR, 0.2963 in PR, and 0.1364 in NR.

**Table1. Cooccurrence rate in one sentence**

| TOPIC_ID | DR(Definitely Relevant) | PR(Possibly Relevant) | NR(Not Relevant) |
|---|---|---|---|
| 92 | 0.0000 | 0.0000 | 0.0000 |
| 93 | 0.3929 | 0.0000 | 0.0455 |
| 94 | 0.9091 | 0.6061 | 0.4375 |
| 95 | 0.2308 | 0.2941 | 0.3125 |
| 96 | 0.0000 | 0.0000 | 0.0000 |
| 97 | 0.0000 | 0.0000 | 0.0000 |
| 98 | 0.5408 | 0.1905 | 0.2449 |
| 99 | 0.2368 | 0.1154 | 0.2308 |
| Average | 0.5000 | 0.2963 | 0.1364 |

Correspondingly, we divide the abstract text into three granularities referring to sentence boundary, naming UniSen (unique sentence), BiSen (two near sentences), and Abstract (content in <AbstractText> field). In the second type of GGT, i.e. Gene-Disease template, gene name in the query is G, and the items in the internal resource for G are G1, G2, and G3, so is disease name D and its D1, D2 and D3. If the logical formula, (G or G1 or G2 or G3) and (D or D1 or D2 or D3), is satisfied in UniSen, BiSen or Abstract, the document containing the above unit will be weighted properly in our ranking algorithm.

### 2.4 Pattern Extraction and Match

In section 2.3, our work focuses on the cooccurrence scope of genomic entities in the context, however, the cooccurrence might not illustrate the relation required by topic. Furthermore, we want to dig out the actual description of "the role in" relation in MEDLINE collection.
As the lack of corpus which provides training data about relation expression, we try to extract the patterns from the sentences including all the objects involved in one relation automatically, with the help of sentence alignment algorithm [3]. After (Part-Of-Speech) POS and pattern extraction at token level, we find some meaningful verbs for relation expression, such as *"effect", "treat", "active", "associate", "bind"* and so on, and all these words are used to expanding queries. Moreover, the patterns, for instance [Gene] {effect, active} [Disease], are used to evaluate the sentences within both gene name and disease name, and the score given by the pattern match algorithm are fused into document ranking program.

### 2.5 Official Runs & Results

In this year's Ad Hoc Retrieval task, we submitted two official runs generated automatically, focusing on the latter four types of topics, where THUIRgen1S is based on the cooccurrence of entities at different retrieval granularity, and THUIRgen2P is based on pattern idea.

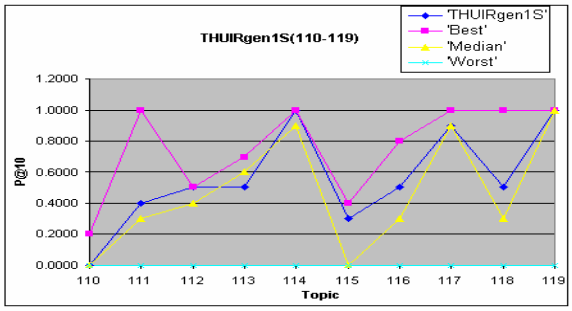


**Fig1. Gene-Disease performance**

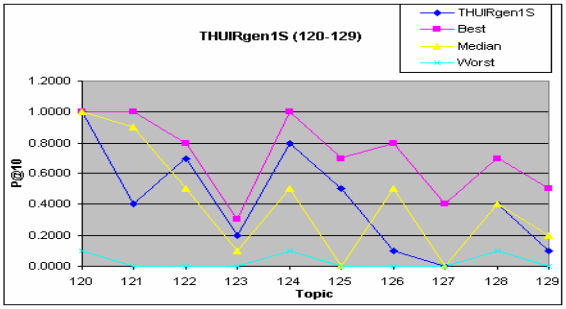


**Fig2. Gene-Biological Process performance**

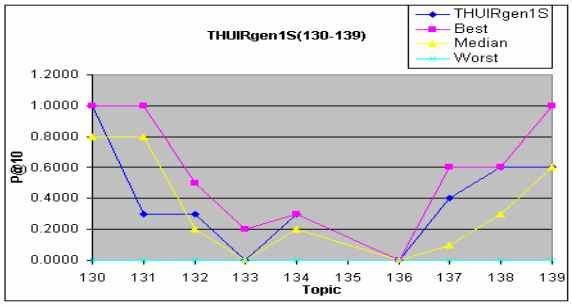


**Fig3. Gene-Organ Function-Disease performance**

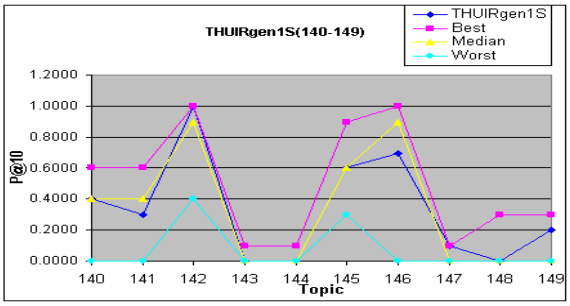


**Fig4. Gene-Biological Impact performance**

Fig1--4 show our P@10 performances at Topic110-149 in blue line, with several points at the Best, most points above the Median.

## 3. Categorization Task

### 3.1 Task Overview

The TREC Genomics 2005 Categorization Task is a traditional Text Categorization (TC) problem. In most TC applications, a Bag-of-Words method is implemented, which treats a document as an unordered set of words. Thus a document can be represented by a high dimensional vector, with each dimension giving the weight of a certain word. Training can be performed on the training set, which is now a set of high dimensional vectors, to get a classification model on traditional classifiers, such as Support Vector Machine, or Naïve Bayes classifier.

In the triage subtask of TREC Genomics 2004 Categorization Task, we did all the above work, like most researchers had done, and got a median result among all the participants. In this year's task, we made a thorough modification to our classification system: a new type of feature, which can contain more semantic information, is proposed, and to generate this feature, a new recursive incremental machine learning method is employed. We still use Support Vector Machine, a common, simple yet powerful tool, as the classifier. Experiment results show that our new idea on the feature is successful at least in this field.

Beside this, on GO annotation subtask, we made use of the thesaurus MeSH library [5] to enrich the feature set.

### 3.2 Seed-based Loose N-gram Feature

TREC protocol provided a cheat-sheet explaining how the positive documents are different from the negative ones. We first extract the "meaningful" words from the cheat-sheet as the seeds. We assume that the sentences containing these seeds are good for classification. And all such sentences are extracted. Between each seed and each single word in the seed's host sentence, a word pair is made as a feature candidate. All the feature candidates are then filtered by the $Chi^2$ measure:

$$Chi(t_k, c_i) = \frac{n[P(t_k, c_i)P(\overline{t_k}, \overline{c_i}) - P(t_k, \overline{c_i})P(\overline{t_k}, c_i)]}{P(t_k)P(c_i)P(\overline{t_k})P(\overline{c_i})}$$

where $t_k$ is for term k, and $c_i$ is for category I, n is the number of all the documents.

Those pairs who have high $Chi^2$ values are selected as features, and those who have a high *DF* value, while $Chi^2$ value are not so high, are selected as new seeds. Thus a recursive procedure can be built, which would stop when no new good features are generated.

Table2 shows some example features for the four subtasks. It is obvious that this kind of new feature is very expressive, and has much more semantic information than a single word. The seeds are carefully selected so that most parts of the documents are covered, which avoids the loss of information. The generation method is carefully designed to guarantee that the number of features is under control, which prevents from the data sparse problem.

**Table2. Sample features generated by seed-based loose n**

| Allele | Expression | GO | Tumor |
|---|---|---|---|
| mice homozygous | embryonic hybridization | protein heterozygous | *mouse cell line histology* |
| generation genotype | *gene expression embryo* | protein embryo | tumor hyperplasia |
| analysis genotype | expression arch | protein mouse | adenoma polyp |
| mice compare | embryonic developmental | expression northern | lymphoma loss |
| mice comparable | adult hybridize | abnormal embryo | neoplasia age |
| generation heterozygous | blot adult | expression section | tumor age |

**(* Note: the Italic features contain more words)**

### 3.3 MeSH Library

The result of the GO subtask wasn't good enough at first. From the experience of former participants [4], we turned to the existing biological knowledge for help. In the feature selection stage, we select the words which occur in our MeSH library as the patch to the GO feature space. We didn't make any change to all the other three topic's feature sets.

### 3.4 Official Runs & Results

As Table3 shows, among all the participants, our result is close to the best result, and better than the median. Considering that our result is achieved without traditional word features, this is an amazing success.

**Table3. THUIRgen official Runs and Results**

| Subtask | Measurement | Best | Median | Our Run |
|---|---|---|---|---|
| **Allele** | **Precision** | 0.7957 | 0.3582 | 0.4902 |
| | **Recall** | 0.9578 | 0.8946 | 0.9006 |
| | **F-score** | 0.6667 | 0.5070 | 0.6348 |
| | **Normalized Utility** | 0.8710 | 0.7785 | 0.8455 |
| **Expression** | **Precision** | 1.0000 | 0.1228 | 0.1322 |
| | **Recall** | 0.9905 | 0.8190 | 0.9238 |
| | **F-score** | 0.4333 | 0.1994 | 0.2312 |
| | **Normalized Utility** | 0.8711 | 0.6548 | 0.8290 |
| **GO** | **Precision** | 0.5542 | 0.2102 | 0.2107 |
| | **Recall** | 0.9363 | 0.6506 | 0.6776 |
| | **F-score** | 0.4230 | 0.3185 | 0.3214 |
| | **Normalized Utility** | 0.5870 | 0.4575 | 0.4468 |
| **Tumor** | **Precision** | 1.0000 | 0.0526 | 0.0213 |
| | **Recall** | 1.0000 | 0.9000 | 0.9500 |
| | **F-score** | 0.4375 | 0.0952 | 0.0417 |
| | **Normalized Utility** | 0.9433 | 0.7610 | 0.7610 |

## 4. Conclusions and Future Work

In the TREC2005 Genomics Track, we have tried to explore the potential of document collection itself, and then import the explored domain-specific knowledge into both Ad Hoc task and Categorization Task. The evaluation results encourage our research work, as we are able to get above the median merely depending on the limited resource. However, there are lots of problems left for us thanks to the experience of TREC2005 Genomics Track, and worth of considering seriously. The external database, which can bring in lots hints to internal resource for query expansion, should emerge in our system. The arbitrarily assigned weights, which lead to unsatisfied MAP with good performance at P@10 and Recall, should be modified. THUIRgen2P failed because of some unpredictable reasons still under analysis. Seed-based loose n-gram Feature is proved useful, and expected to become an aid of the traditional word features.

In next phase, we would solve all the problems mentioned above and work at combining internal and external resource together, and fusing seed-based loose n-gram feature with typical word feature to enhance the performance of our IR and TC system.

## Acknowledgements

The work was supported by Chinese Natural Science Foundation under grant No.60272019 and 60321002.

We also would like to thank Dr. Min Zhang for her good advice at our work, and Hao Yu, Shilin Ding, Xin Sun, Xiaozhe Li , and Zhiyuan Liu for their works at internal resource part and pattern part.

## References

[1] Chang JT, Schütze H, and Altman RB. Creating an Online Dictionary of Abbreviations from MEDLINE. The Journal of the American Medical Informatics Association. 9(6): 612-20.

[2] Buttcher S, Clarke CLA, and Cormack GV.Domain-specific synonym expansion and validation for biomedical information retrieval(MultiText experiments for TREC 2004). TheThirteenth Text Retrieval Conference: TREC2004. 2004. Gaithersburg, MD: National Institute of Standards and Technology.

[3] M.L. Huang, X.Y. Zhu, Y. Hao, D.G. Payan, K. Qu and M. Li. Discovering patterns to extract protein-protein interactions from full-texts. Bioinformatics, Dec, 2004; 20(18):3604-3612.

[4] William R. Hersh, Ravi TB, Laura R, Phoebe J, Aaron MC, Dale FK. TREC 2004 Genomics Track Overview. TheThirteenth Text Retrieval Conference: TREC2004. 2004. Gaithersburg, MD: National Institute of Standards and Technology.

[5] U.S. National Library of Medicine-Medical Subject Headings home page. http://www.nlm.nih.gov/mesh/, 2005.

[6] Metzler, D. and Croft, W.B., "Combining the Language Model and Inference Network Approaches to Retrieval," Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval, 40(5), 735-750, 2004.