DUTIR at TREC 2007 Genomics Track

Zhihao Yang, Hongfei Lin, Baojin Cui, Yanpeng Li, Xiao Zhang
Department of Computer Science and Technology, Dalian University of Technology
No 2 LingGong Road Shahekou District, Dalian 116023, China.
{yangzh, hflin}@dlut.edu.cn, cuibaojin@gmail.com, lyp_8218@163.com,
zhangxiao830416@163.com

Abstract

This paper describes our experiments on TREC 2007 Genomics Track which is concerned with question answering extraction from full-text biomedical literatures. In our experiment, named entities were recognized at the preprocessing stage using a two-view method. MeSH was used to expand the terms. We performed passage retrieval by using sentence-level half overlapped sliding windows. Indri structured query language operators were also used to construct queries.

1. Introduction

TREC 2007 Genomics Track is a modification of the question answering extraction task used in 2006, where the answers, in part, are lists of named entities of a given type. The new challenge is how to recognize the named entities and how to combine them into the retrieval model. We use a simple entity-level retrieval method to solve this problem. Fourteen types of entities mentioned in 2007 protocol [1] are recognized at the preprocessing stage. The entity class name is appended to each named entity as a special token. If terms of a query are concerned with some class name of named entity, the token that represents the entity class is added to the query as a common term and then retrieve the index. Our method for passage retrieval is almost identical to our system at 2006 Genomics Track [2], which uses a sentence-level half overlapped sliding window and a linear combination of the passage and paragraph score. We used MeSH (Medical Subject Headings) for query expansion. Indri structure query language model [3] is used in our two interactive runs (DUTgen1 and DUTgen2). BM25 algorithm is used in the automatic run (DUTgen3).

The other sections are organized as follow: Section 2 presents the method for named entity recognition. Section 3 describes query expansion and retrieval. Section 4 is the result discussion. Section 5 is the conclusion.

2. Named Entity Recognition

In this year's task, most works in preprocess are similar to our early work [2]. In addition, Named Entity Recognition (NER) is used in this stage due to the specialty of the new topics.

NER in biomedical domain has attracted the attention of numerous researchers in resent years. The official evaluation results of JNLPBA [4] and BioCreative 2004 [5] show that the state-of-the-art performances are between 70%-85% varying with different evaluation measures. The general approaches can be categorized into dictionary-based, rule-based and machine learning based methods. Successful systems always combine the three methods to obtain a high overall performance. In this task, fourteen types of named entities are required to recognize. We employ a two-stage method. In the first stage, four types of named entities are recognized using a

Conditional Random Field (CRF) model. The training data is derived from GENIA corpus [6], where 36 classes of entities are labeled by biologists. In our experiment, we merged the 36 classes into 4, i.e., PROTEINS, GENES, CELL OR TISSUE TYPES and OTHER NAMES. Features are words, N-grams, regular expressions and so on. Table 1 shows the relationship between the original GEINA labels and the four classes.

Entity Class	**GEINA label**
PROTEINS	protein_subunit, protein_substructure, protein_molecule, protein_family_or_group, protein_domain_or_region, protein_complex, protein_N/A
GENES	RNA_molecule, RNA_family_or_group, RNA_domain_or_region, RNA_N/A, DNA_substructure, DNA_molecule, DNA_family_or_group, DNA_domain_or_region, DNA_N/A
CELL OR TISSUE TYPES	cell_type, tissue
OTHER NAMES	polynucleotide, peptide, other_organic_compound, other_name, other_artificial_source, nucleotide, multi_cell, mono_cell, lipid, inorganic, cell_component, carbohydrate, body_part, atom, amino_acid_monomer

Table 1: Relationship between the original GEINA labels and entity classes

The next stage focuses on the names labeled as OTHER NAMES in the first stage. Dictionary and rule based methods are used to categorize OTHER NAMES into proper classes. For the entity type DISEASES, BIOLOGICAL SUBSTANCES and SIGNS OR SYMPTOMS, we search the texts in OTHER NAMES recognized in the previous stage using a dictionary extracted from MeSH. If the term is found in the dictionary, it will be assigned a label of corresponding class. For DRUGS, the dictionary is obtained from http://www.rxlist.com/drugs/alpha_a.htm. For the rest entity types, manual rules are applied to extract names according to the last token of the name or its prefix or suffix. These key words for each class are listed in Table 2.

Entity Class	**Key Words or Affixes**
ANTIBODIES	antibody ,antibodies, anti~
MOLECULAR FUNCTIONS	activity, binding
MUTATIONS	mutation, mutations, mutants, mutant, variants, variant
PATHWAYS	pathway, pathways, metabolism
STRAINS	strain, strains
TOXICITIES	toxicitiy, toxic~
TUMOR TYPES	~oma, ~omas, ~tumor, ~tumos

Table 2: Key words or affixes for rule-based entity recognition

3. Query Expansion and Retrieval

In the two interactive runs (DUTgen1 and DUTgen2), we pick up all the name entities in topics at beginning, such as disease, gene and other entities. In DUTgen3, noun phrases are extracted automatically using GENIA Tagger [7] and articles (such as “the” and “a”) are removed. After obtaining a list of noun phrases, our system expands the phrases into lists of synonyms by searching the MeSH database. We download MeSH files in ASCII format, and program an interface to search any noun phrase. All SY terms in SCR2007 and entry terms in Descriptor Records2007 related to a name entity are considered to be synonyms. These terms are combined with the original name entity to form a preliminary synonym list. We remove anything after the first comma, parenthetical strings and all punctuations.

The passage retrieval method is similar to our earlier work [2], which uses a sentence-level half overlapped sliding window approach. The passage length is also 60. In DUTgen3 we use BM25 algorithm and a linear combination of the passage and paragraph score. In the two interactive runs (DUTgen1 and DUTgen2) queries are constructed manually and interactively by utilizing several of the Indri structured query language operators, such as #syn, #1, #band, #filreq and #combine, which is similar to the work [8]. For some terms that are selected manually, fuzzy match which is implemented by #uwN operator is applied in DUTgen1. In limited N window all terms must appear within current context in any order. There is an example in topic 212. If you retrieve “#1(insect segmentation)” in Indri, you will get only a few relative documents, but if you input “#uw5 (insect segmentation)”, you will get much more relative documents involving “segmentation of insect” or some patterns else. In addition, DUTgen1 uses only the passage ranking, while DUTgen2 uses a linear combination of the paragraph score and passage score, where paragraph retrieval uses BM25 algorithm.

4. Results

Table 3 shows the results of the three submitted runs. It can be seen that the performances of the two interactive runs are slightly better than the automatic run in Passage, Passage2 and Document MAP but have a significant improvement in Aspect MAP. The possible reason is that the two interactive runs employ the Indri structure query operators which will bring more “precise” rank than retrieval with separated tokens. When examining the results, we find that the performances for the topics about the entity class PROTEINs, GENEs, and CELL OR TISSUE TYPES are much better than other topics. The performances of NER in these classes are relatively high due to large amount of training data, while for other types simple rules may not work well since we are not biologists. It indicates that NER is an important step in this task.

Run ID	Passage	Passage2	Document	Aspect
DUTgen1	0.06195488	0.03844992	0.18175141	0.18654663
DUTgen2	0.05949058	0.03386782	0.18510934	0.14110326
DUTgen3	0.05866095	0.03137043	0.17051470	0.08828482

Table 3: Performance of official runs

5. Conclusions

In this year's task, it is difficult to pick up fourteen types of entities mentioned in corpora while all possible methods are applied into this phase. The precision of recognition directly affect that of retrieval in the next phase. Also it is the key phase. If entities are not recognized, much post-process is not helpful. Compared with other teams' approaches, NER is a straight way to solve this year's task, but it is not promising enough for many types if its performance is not guaranteed.

Indri structured query language is verified in our experimented of last year's task. Due to its efficiency and effect for specifically key terms, it is applied in this year's task to improve NER's performance.

Question and answer extraction (passage retrieval) system can give more specific answer to users than document retrieval. However it is a difficult task. In our opinion, the main challenge is how to improve the performance of understanding the meaning of the given sentences or passages, which is just the task of text mining. Retrieval technique based on information of surface words can be used as a preliminary step, but it will not bring significant improvement to this kind of task.

References

[1] http://ir.ohsu.edu/genomics/2007protocol.html.

[2] Z Yang, H Lin and Y Li, et al. DUTIR at TREC 2006 Genomics and Enterprise Tracks. The Fifteenth Text Retrieval Conference Proceedings (TREC 2006).

[3] Trevor Strohman, Donald Metzler, Howard Turtle, and W. B. Croft. Indri: A language model based serach engine for complex queries. In Proceedings of the International Conference on Intelligence Analysis, 2004.

[4] Kim JD, Tomoko O and Yoshimasa T, et al. Introduction to the Bio-Entity Recognition Task at JNLPBA. In the Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-04), 2004:70--75.

[5] Lynette Hirschman, Alexander Yeh, Christian Blaschke and Alfonso Valencia. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 2005, 6(Suppl 1):S1.

[6] Kim, J., Ohta, T., Tateisi, Y. and Tsujii, J. GENIA corpus - a semantically annotated corpus for bio-text mining. Bioinformatics. 19(suppl. 1). pp. i180-i182.

[7] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/.

[8] Andrew B. Goldberg, David Andrzejewski, Jurgen Van Gael, Burr Settles and Xiaojin Zhu. (2006). Ranking Biomedical Passages for Relevance and Diversity: University of Wisconsin, Madison at TREC Genomics 2006. Proceedings of the Text REtrieval Conference.