Using IR-n for Information retrieval of Genomics Track

María Pardiño, Rafael M. Terol, Patricio Martínez-Barco, Fernando Llopis, Elisa Noguera
Grupo de Investigación en Procesamiento del Lenguaje Natural y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos
University of Alicante, Spain
maria,rafamt,patricio,llopis,elisa@dlsi.ua.es

Abstract

Nowadays there is a big amount of biomedical literature which uses complex nouns and acronyms of biological entities thus complicating the task of retrieval specific information. The Genomics Track works for this goal and this paper describes the approach we used to take part of this track of TREC 2007. As this is the first time we participate in this track, we configurated a new system consisting of the following diferenciated parts: preprocessing, passage generation, document retrieval and passage (with the answer) extraction. We want to call special attention to the textual retrieval system used, which was developed by the University of Alicante. Adapting the resources for the propouse, our system has obtained precision results over the mean and median average of the 66 official runs for the Document, Aspect and Passage2 MAP; and in the case of Passage MAP we get nearly the median and mean value. We want to emphasize we have obtained these results without incorporating specific information about the domain of the track. For the future, we would like to further develop our system in this direction.

General Terms

Measurement, Performance, Experimentation

Keywords

Information Retrieval, Genomics Track of TREC, Document/Aspect/Passage/Passage2 MAP

1 Introduction

Like the last year [7], this edition of the Genomics Track consists of retrieval of passages from HTML documents where systems will return passages of text. But, in this case, the question answering extraction task used in 2006 was modified, so that, instead of categorizing questions by generic topic type (GTT), the questions were based on biologists' information needs and the answers, in part, are lists of named entities of a given type.

For this task, the documents used come from a new full-text biomedical corpus in HTML format assembled with permission from Highwire Press. Therefore, we had to pre-process the texts removing the HTML spans and adapting them to the format our information retrieval system needs (creating one passage from one paragraph). We split each document of the corpus into paragraphs (HTML text bounded by the HTML tag <P>) and indexed them. Moreover, empty spans were removed from the texts because they do not give any information for retrieval.

```
<DOCNO>12401806-19047-204</DOCNO>
<DOCID>12401806-19047-204</DOCID>
<TEXT>Materials-- Cell culture materials, restriction enzymes, and PCR primers
were from Invitrogen. All other chemicals were reagent grade or better. </TEXT>
</DOC>
```

Figure 1: Example of text format accepted by IR-n [2]

The next figure shows the overall architecture of our system. Firstly, we prepared the corpus by removing HTML spans and adapting it to the information retrieval system. Since we used IR-n [2], which is based on retrieval passages, we indexed all the passages obtained in the previous step. We also pre-processed the topics and lastly, the system gave the answer to the topics formulated.

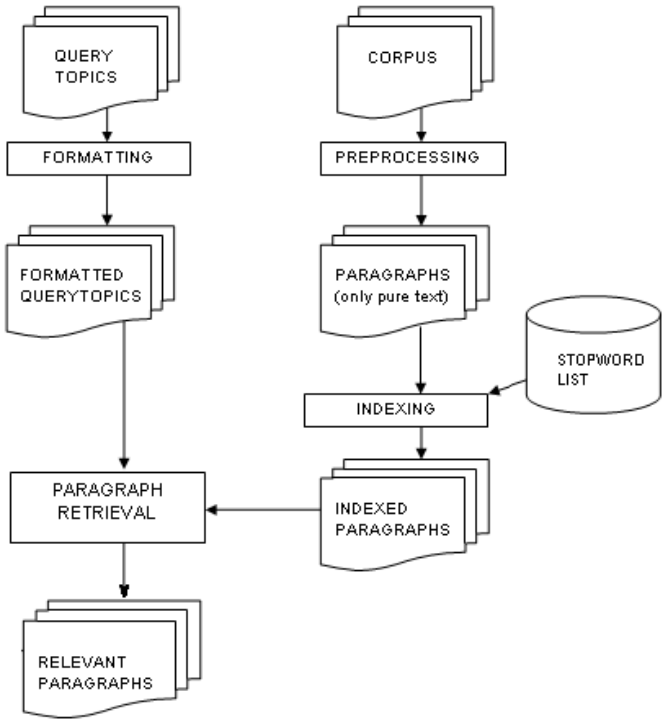

Figure 2: Configuration of our system

In the original topics, the list entity types are incorporated into the questions as capitalized phrases within square brackets. Therefore, we pre-processed the topics, adapting them to the suitable format for the information retrieval system used, IR-n. [2].

```
<top>
<num> C200 </num>
<EN-title> What serum change expression in association with high disease
activity in lupus? </EN-title>
<EN-desc>  </EN-desc>
<EN-narr>  </EN-narr>
</top>
```

Figure 3: Example of topic format accepted by IR-n [2]

Similar to the last edition of Genomics Track, after sending TREC our results, relevance judges assigned the relevant passages "answers" or items belonging to a single entity class. Passages must contain some named entities of the given type with supporting text that answers the given question to be marked as relevant.

This paper contains the following parts: Firstly, we present the background and the state-of-the-art; the following section explains the main characteristics of the IR-n system, further, we describe the experiments done and finally, in the last section we present the conclusions reached.

2 Background

In this section we talk about different systems which participated in previous editions of the track [7] [6]. We have carried out this analysis to study several configurations which took part in the Genomics Track. In particular, we analyse cases of Genomics Track 2006 such as: University Hospital of Geneva-University of Geneva [3], National Taiwan University [1], State University of New York at Buffalo [4] and University of Neuchatel [5].

Firstly, we review the work made in Geneva (Universities, Institute of Bioinformatics and Artificial Intelligence Lab.) They used a module to tokenize the query to normalize and expand only diseases, chemical substances and body parts using UMLS resources (but not gens and proteins). Moreover, they declared that by using MeSH categories it was possible to improve retrieval effectiveness in the corpus used.

Next, we mention the National Taiwan University. They used Lemur as the information retrieval system and worked at passage level. They used KL-divergence as the retrieval model. They made query expansion and post-processed the results.

The State University of New York at Buffalo used the SMART retrieval system and a pre-retrieval expansion method using the ABGene and MetaMap tools and two different weighting schemes: pivoted length normalization and augmented tf-idf, getting results above the median in the last case.

Finally, we describe the approach developed by the University of Neuchatel. They used a 5-gram indexing scheme, which peformed worse than the word-based scheme. Thus, orthographic variants for search words did not improve retrieval accuracy.

3 IR-n System

For this track we have used IR-n [2], which is an information retrieval system based on passages where each document is considered like a set of passages and where a passage is defined as a portion of contiguous text block. This kind of systems, unlike those which are based on documents, allows considering the proximity of appearance of the words in a document.

IR-n [2] uses the phrase as the unit to define the passages. So, the passages are defined by a number of consecutive phrases of the document.

In the following section, we describe the main characteristics of the IR-n system [2] and the techniques we have used for the Genomics Track.

3.1 Resources: stemmers and stopword list

The stopword list of each language contains those words whose presence does not help to determine if a document or a passage is relevant or not (even if these words appear in the query), while stemmers obtain the root of a word removing their suffixes and prefixes, for their index and search. Thus, both of them remove what is not helpful in information retrieval.

The stemmers and the stopword list used by IR-n are available at `www.unine.ch/info/clef`.

3.2 Weighting models

The weighting models allow to quantify the similarity between a text (a complete document or a passage) and the query. These measures are based on the terms that are shared by the text and the query, as well as on the help to discriminate between the different documents that each term can provide us. IR-n uses several weighting models, but for this competition we used okapi. The document ranking produced by each weighting model is obtained using the same general expression, defined as the product of the weight of a term in the document and the weight of the term in the query.

$$sim(q,d) = \sum_{t \in q \wedge p} w_{t,p} \cdot w_{t,q} \tag{1}$$

Variables List The variables used in the following formulas have the following signifficance 2:

- $f_{t,p}$ is the frequency of the term t in the passage p,
- $f_{t,q}$ is the frequency of the term t in the query q,
- n is the number of documents in the collection,
- n_t is the number of documents in which t appears,
- k_1, b and k_3 are constant values,
- ld is the length of the document,
- avg_{ld} is the average of the length of the documents

Okapi Using the okapi model, the weight of a passage p for a query q is given by:

$$\begin{aligned} w_{t,p} &= \frac{(k_1+1)\cdot f_{t,p}}{K \cdot f_{t,p}} \\ w_{t,q} &= \frac{(k_3+1)\cdot f_{t,q}}{k_3 \cdot f_{t,q}} \cdot w_t \\ K &= (1-b) + b \cdot \frac{ld}{avr_l d} \\ w_t &= log_2 \frac{n - n_t + 0.5}{n_t + 0.5} \end{aligned} \tag{2}$$

4 Experiments and Results

In the experiment phase, since IR-n is a parametrizable system, we find a concrete configuration of the input parameters for our data collection. In addition, we describe the input parameter of the system:

- **Size of the Passage (sp)**: Number of phrases that form the passage.
- **Weight model (wm)**: We can use two weighting models: **okapi** y **dfr**.
- **Opaki parameters**: k_1, b and avg_{ld} (k_3 is fixed as 1000).
- **Dfr parameters**: c and avg_{ld}.
- **Query expansion parameters**: If **exp** is equal to 1, this denotes we use relevance feedback based on passages in this experiment. But, if **exp** is equal to 2, the relevance feedback is based on documents. In our case, we do not use Query expansion. **num** denotes the number of passages or documents that the expansion will use, and **term** indicates the k terms extracted from the best ranked passages or documents from the original query
- **Evaluation Measure**: Mean average precision (**MAP**) is the evaluation measure used in order to evaluate the experiments.

In the next table, we specify the values of the parameters we established to be used by IR-n. We want to emphasize that due to lack of time we were not able to find the best configuration and therefore we chose the one that gave good results in previous experiments.

sp	wm	avgld	k1	b
4	okapi	300	1	0.5

Table 1: Configuration for the best results at Genomics Track 2007

Finally we present the results obtained for the Genomics Track. We have only sent TREC one automatic run, IR-n. In many answers, the system outperforms the median and mean results for the 66 official runs for the track The results are shown in Table 2.

Descriptive Statistics	Document MAP	Aspect MAP	Passage MAP	Passage2 MAP
Minimum	0.0329	0.0197	0.0029	0.0008
Median	0.1897	0.1311	0.0565	0.0377
Mean	0.1862	0.1326	0.0560	0.0398
Maximum	0.3286	0.2631	0.0976	0.1148
IR-n	0.2351	0.1976	0.0486	0.0606

Table 2: Results obtained by our system at Genomics Track 2007

5 Conclusion and future work

In our first participation in TREC Genomics, we have obtained results above the median and mean values. For this, we have adapted IR-n (an information retrieval system based on passages) to our propouse. We want to highlight that its results are above expectation because we did not use any Medical Ontology to improve the accuracy of the results.

Despite of using a generalist system, the results are quite encouraging. For the future, we hope to obtain a better system capable of retrieving information from the task in a more precise way.

Therefore, we would like to continue the task next year and measure improvements that we introduced in our system in comparison to the state-of-art of the moment.

6 Acknowledgements

This work has been partially supported by the framework of the project QALL-ME (FP6-IST-033860), which is a 6th Framenwork Research Programme of the European Union (EU), and the Spanish Government, project TEXT-MESS (TIN-2006-15265-C06-01).

References

[1] Wen-Juan Hou Kevin Hsin-Yih Lin and Hsin-Hsi Chen. Report on the TREC 2006 Experiment: Genomics Track. In E. M. Voorhees and Lori P. Buckland, editors, *NIST Special Publication 500-272: The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*.

[2] Fernando Llopis. *IR-n: Un Sistema de Recuperación de Información Basado en Pasajes*. PhD thesis, University of Alicante, 2003.

[3] P.Ruch, A. Jimeno Yepes, F. Ehrler, J. Gobeill, and I. Tbahriti. Report on the TREC 2006 Experiment: Genomics Track. In E. M. Voorhees and Lori P. Buckland, editors, *NIST Special Publication 500-272: The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*.

[4] Miguel E Ruiz. UB at TREC Genomics 2006: Using Passage Retrieval and Pre-Retrieval Query Expansion for Genomics IR. In E. M. Voorhees and Lori P. Buckland, editors, *NIST Special Publication 500-272: The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*.

[5] Jacques Savoy Samir Abdou. Report on the TREC 2006 Experiment: Genomics Track. In E. M. Voorhees and Lori P. Buckland, editors, *NIST Special Publication 500-272: The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*.

[6] TREC. *http://trec.nist.gov/pubs/trec15/t15_proceedings.html*, 2006.

[7] Phoebe Roberts William Hersh, Aaron M. Cohen and Hari Krishna Rekapalli. TREC 2006 Genomics Track Overview.