I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

Bio-medical Question & Answering (QA)

We have developed a novel framework aimed at answering a user’s bio-medical question posed in natural language.

PubMed is a premier literature repository in the biomedical domain. It represents a precious daily tool for most researchers. At present, PubMed is reaching the target of 25 million indexed documents (including: articles, abstracts, letters and other documents) with a constant increase per year that exceeded the threshold of one million items in 2014. The standard keyword based searching and ranking has shown to be inadequate for this large amount of data in this restricted domain. In fact, even specific words (i.e. gene names, accession numbers, pathology names) can return tens of thousands results. In response to this problem, ad-hoc text mining solutions or sophisticated advanced interfaces have been developed to limit the searching range and thus the number of results. However, these solutions are unable to model the real question the user wants to answer and often lack in usability. In this work we propose a different approach that exploits the question-answering (Q&A) paradigm in natural language processing as a tool
for both searching and ranking.

**Phase I**: The clinical diagnosis investigates techniques for linking medical cases to information relevant for patient care. We have developed and test a new graph-based proof-of-concept paradigm for processing the knowledge base and the user questions expressed in natural language. The user question is mapped as a subgraph matching problem onto the internal graph representation, and thus can handle efficiently also partial matches.

**Phase II**: An adaptive user interface has been designed and implemented for enabling a simple interaction and assuring a high level of accessibility and usability also for person with different abilities. Specifically the need of blind persons has been addressed to provide easy access to visual interface exploiting different sensorial channels for delivering the same information. Starting from user requirement collection applying participatory design techniques, the development moves to fast prototyping and early test with different category of users, for better refine system functions and interfaces. Both subjective and objective assessment of the system have been carried out for evaluating its effectiveness and efficiency as well as the user experience.

**Results**. We carried out a user study where we have evaluated the system effectiveness for 96 well-formulated questions submitted from human experts. We requested the user to inspect the first ten results of each question and judge whether the answers are appropriate. Judgment is a score in the range (1; 5) where 3 is the minimum score to consider the answer relevant. We left to the user the option of not evaluating some answers. Our evaluation strategy allowed us to reduce the effort of the experts at the cost of the impossibility to measure the system's recall. Table 1 reports the result of the evaluation with the three measures for different values of $k$.

<table>
<thead>
<tr>
<th>Measure</th>
<th>$k=1$</th>
<th>$k=5$</th>
<th>$k=10$</th>
<th>$k&gt;10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>c@n</td>
<td>0.72</td>
<td>0.76</td>
<td>0.83</td>
<td>0.94</td>
</tr>
<tr>
<td>accuracy@n</td>
<td>0.56</td>
<td>0.67</td>
<td>0.78</td>
<td>0.91</td>
</tr>
<tr>
<td>MRR</td>
<td></td>
<td></td>
<td>0.66</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Evaluation results

The MRR value 0.66 indicates that, on average, a good scoring answer is found either in the first or the second position in the returned list of results. The c@k value for $k=1$ and $k=5$ indicates that about 85% of the times an high quality answer is found within the top 5 answers. These measures attest the good quality of the ranking function. The c@k value for $k>10$ indicates that a good scoring answer is found for about 95% of the queries. This indicates the quality of the subgraph matching strategy.

A deeper analysis of the questions not correctly answered by our system showed that the MeSH ontology covers only partially the biomedical domain. As future development of our system we plan to extend our knowledge base with other medical ontologies. We believe that our QA approach is general and could be successfully applied to other text mining tasks in the biomedical field.
Tweet Contextualization

Twitter is increasingly used for on-line client and audience fishing; this motivated the tweet contextualization task. The objective is to help a user to understand a tweet by providing him with a short summary (500 words). This summary should be built automatically using local resources like the Wikipedia and generated by extracting relevant passages and aggregating them into a coherent summary.

A hybrid Tweet Contextualization task of document retrieval and multi-document summarization has been carried out. The document retrieval is based on Nutch\(^1\) architecture and the multi-document summarization task is based graph, cluster, sentence compression & fusion and sentence ordering. The system was tested on the data set of three years of INEX QA track from 2011 to 2013.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

1. From Literature to Knowledge: exploiting PubMed to Answer Biomedical Questions in Natural Language, Pinaki Bhaskar, Marina Buzzi, Filippo Geraci, and Marco Pellegrini, In the proceedings of 6\(^{th}\) International Conference on Information Technology in Bio- and Medical Informatics - ITBAM 2015, September 3 - 4, 2015, Valencia, Spain. (Accepted)

2. PubMed searching made easy and effective using natural language processing techniques, Pinaki Bhaskar, Marina Buzzi, Filippo Geraci, and Marco Pellegrini, In the proceedings of the Bioinformatics Italian Society (BITS) annual meeting 2015, University of Milan "Bicocca", Italy. (Accepted)

3. Tweet Contextualization using Multi-document Summarization, Pinaki Bhaskar, In the proceedings of First Workshop on Language Technology for Indian Social Media Text (Social India) with the 11th International Conference on Natural Language Processing (ICON) 2014, pp. 49-58, December 2014, Goa University.

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

Conference:
1. ICON-2014: 11\(^{th}\) International Conference on Natural Language Processing, organized by the International Institute of Information Technology, Hyderabad (IIIT-H), at Goa University, Goa, India, December 18-21, 2014

Workshop:
2. Language Technologies for Indian Social Media Text (Social India), chaired by Amitava Das, Björn Gambäck, Dipankar Das, ICON 2014, Goa University, Goa, India, 21\(^{st}\) December, 2014

\(^{1}\) http://nutch.apache.org/
Tutorial:

3. Question Answering Techniques for Structured, Semi-Structured and Unstructured Information Sources, given by Manoj Kumar Chinnakotla, Manish Shrivastava, Radhika Mamidi, ICON 2014, Goa University, Goa, India, 18th December, 2014


Seminar:

5. ERCIM – ABCDE Seminar IV, CNR, Pisa, Italy, 23rd – 24th October, 2014

IV – RESEARCH EXCHANGE PROGRAMME (REP)

Universitat Politècnica de València (UPV), Valencia, Spain; Prof. Paolo Rosso, prosso@dsie.upv.es, 10-16 February, 2015.

The research works have been presented to the research group of Dr. Rosso. Some of his students also presented their research works. Possibility of joint research work with Dr. Rosso’s team has been discussed too. We have exchanged lot of research work and ideas of future work. The experience was very good.