IBI at TREC 2018: Precision Medicine Track
Notebook Paper

Francesco Ronzano, Emilio Centeno, Judith Pérez-Granado and Laura Furlong
Integrative Biomedical Informatics Group
Research Programme on Biomedical Informatics (GRIB)
Hospital del Mar Medical Research Institute & Pompeu Fabra University
Barcelona, Spain
{francesco.ronzano,laura.furlong}@upf.edu
{ecenteno,jperez2}@imim.es

Abstract

Nowadays, the growing amount of biomedical scientific literature that can be accessed online represents a valuable source of information useful to tailor medical decisions to a specific clinical case. With this respect, Information Retrieval tools play an essential role in enabling physicians to automatically analyze huge amounts of publications so as to retrieve relevant recent information related to the treatment, prevention or prognosis of specific clinical conditions and traits.

In this paper, we present and discuss the biomedical scientific Information Retrieval strategy we developed in the context of our participation to the Precision Medicine Track of the Text Retrieval Conference 2018. Given the description of a clinical case, we describe how we retrieve from PubMed a ranked list of scientific abstracts that discuss medical care that may be relevant for the patient under consideration. To this purpose we rely on the query formulation capabilities provided by Elasticsearch, a full-text search engine, complemented by data processing steps useful both to properly build search queries and to refine the ranking of search results proposed by Elasticsearch.

1 Introduction

Nowadays, many areas of medical care are more and more characterized by a wide range of possibilities to tailor medical decisions (i.e. treatments, procedures, etc.) to each individual. Indeed, currently we have access to an increasingly large number of clinical and biological datasets like Electronic Health Records, collections of human genomic sequences, and proteomic and metabolomic databases. This data availability can be coupled with powerful computational tools to integrate and mine information, thus giving physicians the opportunity to adapt their decisions to a specific medical scenario by exploiting clinical, genetic, molecular or cellular traits of the patient. This approach towards the customization of clinical decisions to a specific patient (or to a subgroup of the patients' population) is usually referred to as *precision medicine* (PM) (Jameson and Longo, 2015).

In this scenario, the increasing amount of biomedical scientific literature that is currently available online represents a valuable source of information to support PM. However, due to the huge and rapidly growing number of biomedical scientific publications that can be accessed on the web, it is impossible for clinicians to manually explore their content so as to be up-to-date on new, relevant evidence-based treatments. For instance, PubMed, the main search engine for biomedical literature, currently includes more than 27 million papers and is growing at a rate of about 1,500 new publications indexed per day. As a consequence, the availability of Information Retrieval (IR) tools that, given a clinical case, allow to effectively retrieve relevant recent information related to treatment, prevention or prognosis is essential.

In this regard, in the context of the Text Retrieval Conference 2018 (TREC), the Precision Medicine Track (PM Track) has been organized. Participants to the PM Track have been proposed a set of 50 descriptions of oncological clinical cases. Each case is characterized by the type of cancer suffered by the patient, the relevant genetic variants and other demographic information. Given this input data, two IR subtasks have been defined: for each clinical case, participants were required to generate (i) a ranked list of *scientific abstracts*, mainly from PubMed, describing treatments that may be relevant for the patient and (ii) a ranked list of *clinical trials* from ClinicalTrials.gov in which the patient could be enrolled.

Here we present and discuss our participation (team: IBI_PM) to the PM Track subtask related to the ranking of *PubMed abstracts*. In particular, the rest of this papers is organized in 4 sections. Section 2 provides a brief overview of the PM Track dataset, considering both the clinical case descriptions and the collection of scientific abstracts. Section 3 describes our IR approach by providing: the introduction to the TREC_ResMarkerDB Graph (subsection 3.1), a knowledge resource tailored to support a better ranking of Pubmed abstracts; the presentation of our data indexing approach (subsection 3.2); the explanation of the term expansion applied to diseases, genes and variants (subsection 3.3); the description of the our query building approach (subsection 3.4); and finally the definition of our customized refinement of search result ranking (subsection 3.5). Ultimately, section 4 presents the results of our participation to the PM Track and section 5 exposes our conclusions and future venues of research.

2 Dataset

A set of 50 clinical case descriptions has been proposed to the participants of the PM Track at TREC 2018. Eeach clinical case has been created by oncologists and describes the disease, genetic variants and demographic information of a patient. Table 1 shows an example of a clinical case description.

For each clinical case, participants have been required to deal with the following two subtasks: retrieve a ranked list of relevant *scientific abstracts* and a ranked list of *clinical trials* of interest. In our participation to the PM Track we faced the first subtask. Thus, we focused our efforts on the retrieval of *scientific abstracts* that provide medical care information useful to deal with each specific clinical case. With respect to the collection of scientific publications to examine, PM Track organizers provided a January 2017 snapshot of PubMed complemented with a set of abstracts from the proceedings of the American Association of Cancer Research (AACR) and the American Society of Clinical Oncology (ASCO).

Disease	melanoma
Gene	APC loss of function
Demographic	47-year-old male

Table 1: Example of clinical case description of the PM Track 2018.

3 Approach

This Section describes our approach to the *scientific abstracts* retrieval subtask of the PM Track. We relied on Elasticsearch[1] (Gormley and Tong, 2015), an open source Lucene-based full-text search engine, to index and query the collection of about 25 million scientific abstracts contemplated in the PM subtask. In particular, we complemented Elasticsearch powerful capabilities to formulate complex queries to semi-structured textual documents with: (i) an ad-hoc strategy for term expansion, useful to automatically expand the set of terms that will be exploited in Elasticsearch queries to refer to diseases, genes and gene variants; (ii) a customized approach to refine the ranking of query results generated by Elasticsearch. This approach relies on both biomedical information extracted from each scientific abstract and the contents of the TREC_ResMarkerDB Graph, a knowledge resource properly built for this specific IR scenario.

In Figure 1 we provide an overview of the different steps of the *scientific abstracts* retrieval approach we propose.

In the rest of this Section, we describe in detail how we indexed data and the different steps of our *scientific abstracts* retrieval strategy.

3.1 The TREC_ResMarkerDB Graph

In order to better refine and rank *scientific abstract* search results we build the TREC_ResMarkerDB, a knowledge resource that includes structured information on biomarkers of drug response in cancer. Currently this information can be found spread across different databases and in scientific literature. Thus, TREC_ResMarkerDB was created as a centralized repository to gather and provide structured, uniform access to knowledge of biomarkers of drug response in cancer. It was built based on the pipeline used for ResCur[2]: homogenization, standardization, and relation constitution. Although three major changes were in-

[1] https://www.elastic.co/products/elasticsearch

[2] http://resmarkerdb.org

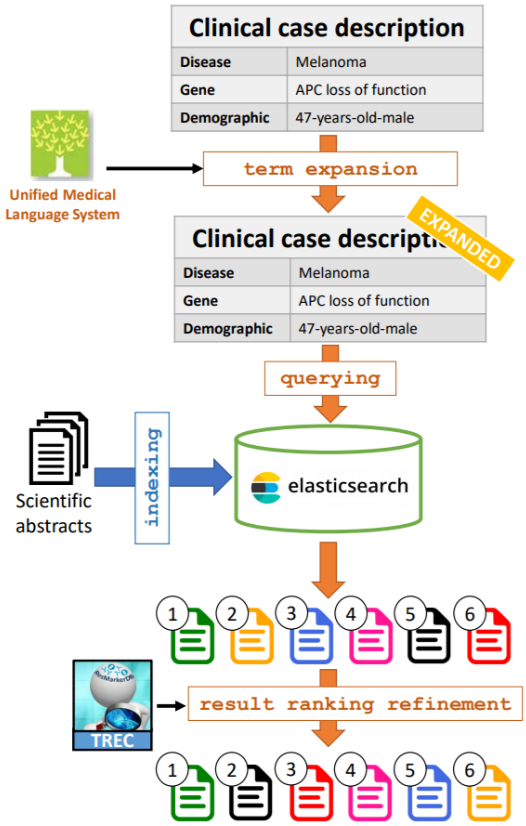

Figure 1: Overview of PM Track approach.

troduced. Firstly, the source databases selected were: Cancer Genome Interpreter (Tamborero et al., 2018) (v.2018/01/17), Clinical Interpretations of Variants in Cancer (Griffith et al., 2017) (v.2018/05/01), ResCur (v.2018/07/20) and COSMIC (Forbes et al., 2016) (v.2018/05/30). Secondly, diseases and chemicals were annotated and standardized using UMLS (Bodenreider, 2004), as it is a unified language system of controlled vocabularies. Finally, the information was all centralized in the association between a gene or biomarker, a disease and a drug; additionally, there is information about the evidence level and exact associated response. Table 2 shows an example of some records of TREC_ResMarkerDB.

3.2 Data indexing

We relied on the full-text search engine Elasticsearch (version 6.2.2) in order to support indexing and querying on the *scientific abstracts* that constitute the IR document base of the PM Track 2018 (see Section 2 for more details). To this purpose, we set up a three-node Elasticsearch cluster. For each abstract of a scientific publication, from PubMed, AACR or ASCO, we indexed the following fields:

Disease	Biomarker or gene	Drug	Evid. level
COREAD	AKT1 E17K	Cetuximab	CL
LUAD	RET fusion	Sunitinib	ET
OS	NF1 deletion	Trametinib	CS
THCA	RET M918T	Vandetanib	GD

Table 2: Snapshot of the TREC_ResMarkerDB. COREAD: Colorectal carcinoma, LUAD: Lung adenocarcinoma, OS: Osteosarcoma, THCA: Thyroid carcinoma. The last column specifies the Evidence level and has the following values: CL: clinical, ET: early trials, CS: case study, GD: guidelines.

- *id*: the ID of the article (considering the PubMed ID for PubMed articles and the file name without extension for AACR or ASCO articles);
- *title*: the title of the article;
- *abstract*: the abstract of the article;
- *publication_date*: the publication date of the article. We considered: (i) for PubMed articles, the most recent date that characterizes the article among the values of the DateCreated, the DateCompleted and the DateRevised XML elements as defined in the PubMed DTD[3]; (ii) for AACR or ASCO articles, the year of publication.

3.3 Term expansion: diseases, gene and variants

As shown in Figure 1, given a clinical case description, the first step of our *scientific abstracts* IR strategy is term expansion. It involves the collection of alternative expressions to describe both the disease(s) suffered from the patients and their genetic variant(s). To this purpose we relied on the Unified Medical Language System (UMLS) (Bodenreider, 2004) (see Subsection 3.3.1) to expand gene and disease names. We also exploted the NCI Thesaurus OBO Edition (NCIT) (Musen et al., 2011) and manually created sets of synonyms to deal with gene variants (see Subsection 3.3.2).

[3] Access https://www.nlm.nih.gov/bsd/licensee/ elements-descriptions.html for a complete description of the date-related information associated to each

3.3.1 UMLS-based disease and gene term expansion

The objective of the disease and gene term expansion is to support the formulation of queries with a higher recall by gathering alternative expressions useful to refer to the diseases and genes mentioned in the clinical case descriptions. With this respect, the 2017AB version of the UMLS has been exploited as follows: given a disease or gene term, both synonyms and terms referring to more specific concepts were retrieved. We implemented a term expansion procedure consisting of the following steps:

1. **retrieval of synonym CUIs**: given the gene or disease label L, we queried the UMLS Metathesaurus and Semantic Network for all Concept Unique Identifiers (CUIs) that have an English label equal to L. Then, diseases and genes were treated distinctly as follows:
   - if L is a disease it was checked whether the CUI belongs to one of the following UMLS semantic types: Congenital Abnormality (T019), Acquired Abnormality (T020), Finding (T033), Injury or Poisoning (T037), Pathologic Function (T046), Disease or Syndrome (T047), Mental or Behavioral Dysfunction (T048), Cell or Molecular Dysfunction (T049), Experimental Model of Disease (T050), Sign or Symptom (T184), Anatomical Abnormality (T190) or Neoplastic Process (T191);
   - if L is a gene it was examined if the CUI belongs to the UMLS semantic type Gene or genome (T028) and is included in one of the following sources: the Medical Subject Headings MeSH (MSH), the National Cancer Institute (NCI) Thesaurus (NCI) or the HUGO Gene Nomenclature (HGNC);

   The employed set of UMLS sources and semantic types exploited were identified by manual exploration and iterative refinement of term expansion results.

2. **retrieval of hyponym CUIs**: by relying on the UMLS Semantic Network, for each synonym CUI retrieved by the previous step, we gathered all the CUIs of child concepts up to a depth level of 3 for diseases, and 1 for genes[4]. We considered the parent-child relations ”has_child (CHD)” and ”has_narrower” (NR) formalized in the UMLS Semantic Network. We filtered all the hyponym CUIs by means of the same constraints over semantic type and source already exploited in the previous step;

3. **CUI lists enrichment**: we enriched each list of CUIs previously created (synonym CUIs and hyponym CUIs) by including all the other UMLS CUIs related by the source asserted synonymy (SY) relation to at least one CUI of the list.

As final result of the term expansion process, we retrieved all the English labels associated to each CUI belonging to these two lists. Thus we created for each initial disease and gene term two lists of terms: the list of synonyms and the list of hyponym terms. An overview of the number of CUIs and terms collected as a result of the UMLS-based disease and gene term expansion is shown in Table 3 and Table 4 respectively. In particular, the disease term expansions generated up to 85 synonyms for each disease term (’acute myeloid leukemia’), while the gene term expansion generated up to 54 synonyms for each gene term (’ERBB2’). When we considered the hyponym concepts up to a depth of three, for diseases the expansion process generates up to 2,706 more specific terms (’sarcoma’), while for genes the hyponym concepts up to a depth of one level managed to gather up to 27 hyponym terms (’CDKN2A’).

3.3.2 Gene variants expansion

According to their nature, variants were expanded following different patterns. In particular, we identified two ways to express variants: (i) non-linguistic variants and (ii) linguistic variants. While non-linguistic variants are identified by means of a specific code-based convention (e.g. V600R, V600K, etc.), linguistic variants are described by means of a linguistic expression like: ’amplification’, ’loss of function’ or ’deletion’.

Non-linguistic variants were mapped to their RSID, when possible, using Biomart Ensembl tool (Ensembl Genes and Variation, version

[4] The final depth levels chosen to gather hyponyms of gene and disease concepts have been identified by comparing term expansion results with depth level of 1, 3, 5 and 7 for both diseases and genes.

DISEASE TERM	Num. CUIs in syn. list	Num. syn. terms	Num. CUIs in hypo. list	Num. hypo. terms
acute myeloid leukemia	3	85	128	949
adenoid cystic carcinoma	1	16	34	208
anaplastic large cell lymphoma	1	29	41	127
basal cell carcinoma	2	27	44	179
breast cancer	2	51	162	835
cholangiocarcinoma	2	26	67	249
colorectal cancer	4	51	134	587
esophageal cancer	2	41	91	479
gastric cancer	2	38	82	450
glioblastoma	2	22	41	201
glioma	2	26	252	1207
head and neck squamous cell carcinoma	1	11	112	1006
leukemia	1	24	277	2153
lung cancer	3	60	150	974
medullary thyroid carcinoma	1	28	12	61
melanoma	1	14	146	613
neuroblastoma	1	6	58	189
non-small cell carcinoma	1	4	65	472
papillary thyroid carcinoma	1	28	25	103
prostate cancer	2	31	61	307
sarcoma	1	27	621	2706
thyroid cancer	1	21	41	270

Table 3: Result of UMLS-based disease term expansion (syn.: synonym, hypo.: hyponym, num.: number).

GENE TERM	Num. CUIs in syn. list	Num. syn. terms	Num. CUIs in hypo. list	Num. hypo. terms
ABL1	1	12	1	8
ALK	2	11	1	6
APC	2	20	1	12
BRAF	1	10	1	12
CDK6	1	6	1	7
CDKN2A	1	51	1	27
EGFR	1	12	1	15
ERBB2	2	54	1	10
FGFR1	2	32	1	22
FLT3	2	13	1	8
IDH1	1	11	1	10
KIT	3	21	1	8
MDM2	1	9	1	10
MET	3	25	1	10
NF1	1	23	1	7
NRAS	2	17	1	11
NTRK1	2	20	1	10
PTCH1	1	8	1	13
PTEN	1	9	1	14
RET	3	25	1	11
ROS1	1	10	1	8
TP53	2	32	2	16

Table 4: Result of UMLS-based gene term expansion (syn.: synonym, hypo.: hyponym, num.: number).

91) (Kersey et al., 2009). Regarding linguistic variants, their description was expanded by means of a set of synonyms provided by different ontologies like NCI Thesaurus, OBO Edition (NCIT). For instance, the expressions 'copy number alteration' and 'CNA' were added as synonyms of the linguistic variant 'amplification'.

3.4 Query building

By properly assembling the set of terms generated by the term expansion (see Subsection 3.3), we could proceed with the query building. The Elasticsearch query had to be useful to retrieve the ranked list of the top-10,000 most relevant *scientific abstracts*. In this Section we describe how we formulated such query. And, in the next Subsection (3.5) we will specify the approach we followed to refine the ranking of the set of 10,000 query results returned by Elasticsearch: in this way we generate the final PM Track 2018 results that include the top-1,000 most relevant *scientific abstracts* for each clinical case description.

When building queries we relied on Elasticsearch support for **query term boosting**. When a term-based query is formulated in Elasticsearch, a term boost score can be specified for each search term contemplated by the query. In this way it is possible to tune the relative relevance of that term in ranking query results (in our case *scientific abstracts*). In general, the relevance of a document is defined by the contributions of all the search terms that appear in that document. By default each search term is given a boost score equal to 1. But, by changing the boost score, the computation of the relevance score of a document changes. So, if we assign a boost score of 2 to a term, it will bring twice the contribution of a term with boost score of 1.

First we analyzed the Elasticsearch query results of several term boosting configuration. Then, we consequently defined the following boosting rules:

- if a query term appears in the title it is assigned a boost score of 8;
- if a query term appears in the abstract text it is assigned a boost score of 3;
- if a query term is a disease / gene mentioned in the clinical case description or one of its synonyms it is assigned a boost score of 5;
- if a query term is a hyponym of a disease / gene mentioned in the clinical case description it is assigned a boost score of 2;
- if a term is a non-linguistic gene variant mentioned in the clinical case description it is assigned a boost score of 35;
- if a term is a linguistic gene variant mentioned in the clinical case description, or one of its synonyms, it is assigned a boost score of 25.

We also created a list of 21 terms and expressions related to precision medicine (e.g. 'Precision Medicine', 'Personalized medicine', 'PM','customized medicine', 'Tailored treatment','Patient-specific treatment', 'Molecular diagnostics', etc.) and assigned to each one of these terms a boost score equal to 20.

Hereafter we present an example considering all the term boosting scores defined before. If we find in the abstract of a paper an occurrence of a disease / gene mentioned in the clinical case description or one of its synonyms derived by term expansion, such matching term will have a final boost score equal to 15. That is 5 multiplied by 3: 5 because the term is a disease / gene mentioned in the clinical case description or one of its synonyms and 3 because it occurs in the abstract of the paper.

As far as it concerns the match of multi-word linguistic gene variants (e.g. 'loss of function'), we performed the search for these expressions by specifying a slop value equal to 4. This strategy was followed by taking into account both the main linguistic variant and its synonyms (e.g. 'copy number alteration' and its synonym 'CNA'). Thus, we considered as valid matches the cases in which the words of the multi-word expression are separated by at most 4 other words (e.g. 'loss of specific function' represent a match for the search term 'loss of function').

In order for a *scientific abstract* to be considered as a candidate search result of an Elasticsearch query, the abstract should: (i) match at least one term among the set of gene and disease terms derived by term expansion (by considering the term mentioned in the clinical case description together with its synonyms and hyponyms), (ii) possibly match also a gene variant (non-linguistic or linguistic one) and (iii) possibly match one or more expressions belonging to the list of 21 precision medicine related terms. Moreover, we considered as candidate results for our queries only papers with non empty abstract texts.

3.5 Refinement of query result ranking

By performing the Elasticsearch query described in Subsection 3.4, we retrieved the 10,000 top-ranked *scientific abstracts* for each clinical case description. For each search result (i.e. abstract) Elasticsearch computes a relevance score. This score is used to rank the result with respect to the other ones; in this way it is possible to quantify the relative relevance of each abstract in answering the considered query. The starting point of this subsection is the ranked list of 10,000 *scientific abstracts* retrieved by Elasticsearch for each clinical case description. Then, here we explain how, in addition to the Elasticsearch relevance score, we computed and combined a set of other custom document relevance scores so as to determine the final relevance score ($finalRelevanceScore$) of each abstract. Such score is exploited to generate, for each clinical case, the final ranking of *scientific abstracts*, thus allowing us to select the top-1,000 most relevant ones that constitute the results of the PM Track 2018.

To support the computation of the $finalRelevanceScore$, as a pre-processing step, we retrieved a list of chemicals from UMLS according to their semantic type. To this purpose the semantic types considered were inferred from the semantic types of known chemicals used in treatments for breast cancer and colorectal cancer. These treatments were extracted from ResMarkerDB. We then identified all the mentions of a chemical matching one of the selected semantic types in both the title and the abstract of each paper.

We computed the $finalRelevanceScore$ of the 10,000 top-ranked *scientific abstracts* retrieved by Elasticsearch for each clinical case by means of the following formula, by relying on a custom combination of a varied set of document relevance scores.

$$\boxed{\begin{array}{l} finalRelevanceScore = \\ 2.0 * ESscoreNoerm + \\ 1.0 * ResMarkDBscore + \\ 0.75 * numDisMatchAll\_EXACT + \\ 0.25 * numDisMatchAll\_HYPO + \\ 0.90 * numGenMatchAll\_EXACT + \\ 0.5 * numChemMentions \end{array}}$$

In particular, the computation of the $finalRelevanceScore$ is based on the following contributions:

- **Normalized Elasticsearch relevance score** ($ESscoreNorm$): the relevance score given by Elasticsearch to each document retrieved by the query (normalized to the interval $[0, 1]$);

- **'TREC ResMarkerDB' score** ($ResMarkDBscore$): after identifying mentions of chemicals in the title and abstract of scientific articles (as explained before in this Section), we considered all the abstracts that include mentions of at least one gene, one chemical and one disease. Each (gene, chemical, disease) combination occurring in these abstracts was evaluated using the TREC_ResMarkerDB (see Subsection 3.1). This centralized repository assessed the current knowledge regarding the response (e.g. sensitive or resistant) and the evidence level (e.g. preclinical level, clinical level or guidelines) of the reported association. To finally generate a score in the interval $[0, 1]$ documents without the occurrence of any (gene, chemical, disease) combination were scored 0. If there was more than one (gene, chemical, disease) combination matching in TREC_ResMarkerDB, the highest score was kept.

- $numDisMatchAll\_EXACT$ / $numGenMatchAll\_EXACT$ is the number of occurrences of the disease / gene term mentioned in the clinical case description, or one of its synonyms (derived by term expansion), considering the title and abstract of a retrieved paper;

- $numDisMatchAll\_HYPO$ is the number of occurrences of a disease term that is a hyponym of the disease mentioned in the clinical case description (derived by term expansion), considering the title and abstract of a retrieved paper;

- $numChemMentions$ is the number of different chemicals mentioned considering the title or abstract of a retrieved paper.

The values of $numDisMatchAll\_EXACT$, $numGenMatchAll\_EXACT$,

$numDisMatchAll\_HYPO$ and $numChemMentions$ were all normalized to the interval $[0, 1]$ by considering the range of values occurring in the 10,000 top-ranked *scientific abstracts* retrieved by Elasticsearch. The weight of each member of the previous formula has been defined by a trial and error approach, by evaluating the $finalRelevanceScore$ obtained by different weights' combinations.

After performing the refinement of query results, we could compute the $finalRelevanceScore$ of each one of the 10,000 top-ranked *scientific abstracts* retrieved by Elasticsearch for each clinical case. We submitted three runs to PM Track subtask related to the ranking of *scientific abstracts*. Each run is characterized by a specific approach to select the ranked list of 1,000 abstracts to consider as the final results of the PM Track 2018, starting from the set of 10,000 abstracts retrieved by Elasticsearch and ranked by their $finalRelevanceScore$. Hereafter we describe the approach we followed in each run.

Run 1: we started off with the set of 10,000 abstracts retrieved by Elasticsearch. We then selected the final list of top-1,000 abstracts to consider as the 'Run 1' results of the PM Track 2018. These consisted on documents from the following subgroups (understanding all the documents of the first subgroup as the top-ranked set, then all the ones from the second subgroup, etc.):

- abstracts with at least one mention of a disease (synonym or hyponym term), a gene (synonym or hyponym term) and a variant in the title;
- abstracts with at least one mention of a disease (synonym or hyponym term) and a gene (synonym or hyponym term) in the title and no variant mentions;
- abstracts with at least one mention of a disease (synonym or hyponym term), a gene (synonym or hyponym term) and a variant considering both title and abstract;
- abstracts with at least one mention of a disease (synonym or hyponym term) and a gene (synonym or hyponym term) considering both title and abstract and no variant mentions;
- abstracts with one mention of a disease (synonym or hyponym term) or (XOR) one mention of a gene (synonym or hyponym term) together with one or more variant mentions, considering both title and abstract;
- abstracts with one mention of a disease (synonym or hyponym term) or (XOR) one mention of a gene (synonym or hyponym term) considering both title and abstract and no variant mentions;

In each subgroup, documents are ordered by their $finalRelevanceScore$ and, in case of equal $finalRelevanceScore$ by their publication date.

Run 2: with respect to 'Run 1', in 'Run 2' we do not gave precedence to term matches in the title with respect to term matches in the abstract text. As a consequence, again, we started off with the set of 10,000 abstracts retrieved by Elasticsearch. We then selected the final list of top-1,000 abstracts to consider as result of the PM Track 2018. The documents were selected by considering those from the following subgroups (again, understanding all the documents of the first subgroup as the top-ranked set, then all the ones from the second subgroup, etc.):

- abstracts with at least one mention of a disease (synonym or hyponym term), a gene (synonym or hyponym term) and a variant, considering both title and abstract
- abstracts with at least one mention of a disease (synonym or hyponym term) and a gene (synonym or hyponym term), considering both title and abstract (no variant mentions)
- abstracts with one mention of a disease (synonym or hyponym term) or (XOR) one mention of a gene (synonym or hyponym term) eventually with one or more variant mentions, considering both title and abstract

As for other runs, in each subgroup, documents are ordered by their $finalRelevanceScore$ and in case of equal $finalRelevanceScore$ by their publication date.

Run 3: in 'Run 3' we ranked documents by matching the age of the patient specified by each clinical case description with the set of age-ranges mentioned in the abstract, if any. In particular, we extracted age-ranges of patients from the 10,000 top-ranked *scientific abstracts* retrieved by Elasticsearch (e.g. "between X and Y years old", "aged X - Y years"), where available. To this purpose we exploited a set of linguistic rules built by means of the JAPE tool[5]. Thus, given the age of the patient specified by the clinical case description, we were able to identify all the abstracts in which such age is included in one or more age-ranges occurring in the same document. As a consequence, we selected the final list of top-1,000 abstracts to consider as result of the PM Track 2018 by considering those from the following two subgroups (understanding all the documents of the first sub-group as the top-ranked set, then all the ones from the second sub-group):

- abstracts with one or more age-ranges including the age of the patient;
- abstract with age-ranges not including the age of the patient or without any age-range specified.

Inside each one of these two subgrous the abstracts were ordered by the same approach defined for 'Run 1' (all the documents of the first subgroup defined for 'Run 1' and with one or more age-range matches as the top-ranked set, then all the ones from the second subgroup defined for 'Run 1' and with one or more age-range matches, etc.). But, first we considered the abstracts matching the age-ranges and then the remaining abstracts up to reaching a maximum of 1,000.

4 System performance

Among the three run submitted (see Subsection 3.5) the approach followed in the 'Run 1' is the one that obtained the best overall evaluation scores in the PM Track 2018. In Figure 2, Figure 3 and Figure 4, we show the evaluation results of our 'Run 1' by means of different parameters. In particular we consider the following system evaluation metrics: the Normalized Discounted Cumulative Gain in Figure 2, the Precision@10 in Figure 3 and the R-precision in Figure 4. Those evaluation results are shown for each one of the 50 clinical case descriptions and, together with the median and best result obtained by all the participating teams of the PM Track 2018.

5 Conclusions and future work

In this paper we presented and discussed our participation to the Precision Medicine Track organized in the context of Text Retrieval Conference 2018 (TREC). In particular, we described in detail our approach to retrieve from PubMed and other repositories of biomedical scientific publications, papers that propose medical cares relevant to a specific clinical case. As future venues of research, we would like to perform an in depth evaluation of the different steps of the Information Retrieval strategy we developed in order to improve both its effectiveness and efficiency. We would also like to set up and make available online a Precision Medicine search engine that implements the scientific abstract search approach proposed.

References

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research* 32(suppl_1):D267–D270.

Simon A Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, et al. 2016. Cosmic: somatic cancer genetics at high-resolution. *Nucleic acids research* 45(D1):D777–D783.

Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine.* " O'Reilly Media, Inc.".

Malachi Griffith, Nicholas C Spies, Kilannin Krysiak, Joshua F McMichael, Adam C Coffman, Arpad M Danos, Benjamin J Ainscough, Cody A Ramirez, Damian T Rieke, Lynzey Kujan, et al. 2017. Civic is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature genetics* 49(2):170.

J Larry Jameson and Dan L Longo. 2015. Precision medicinepersonalized, problematic, and promising. *Obstetrical & Gynecological Survey* 70(10):612–614.

Paul J Kersey, Daniel Lawson, Ewan Birney, Paul S Derwent, M Haimel, Javier Herrero, Stephen Keenan, Arnaud Kerhornou, Gautier Koscielny, Andreas Kähäri, et al. 2009. Ensembl genomes: extending ensembl across the taxonomic space. *Nucleic acids research* 38(suppl_1):D563–D569.

Mark A Musen, Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Christopher G Chute, Margaret-Anne Story, Barry Smith, and NCBO team. 2011.

[5] https://gate.ac.uk/sale/tao/splitch8.html

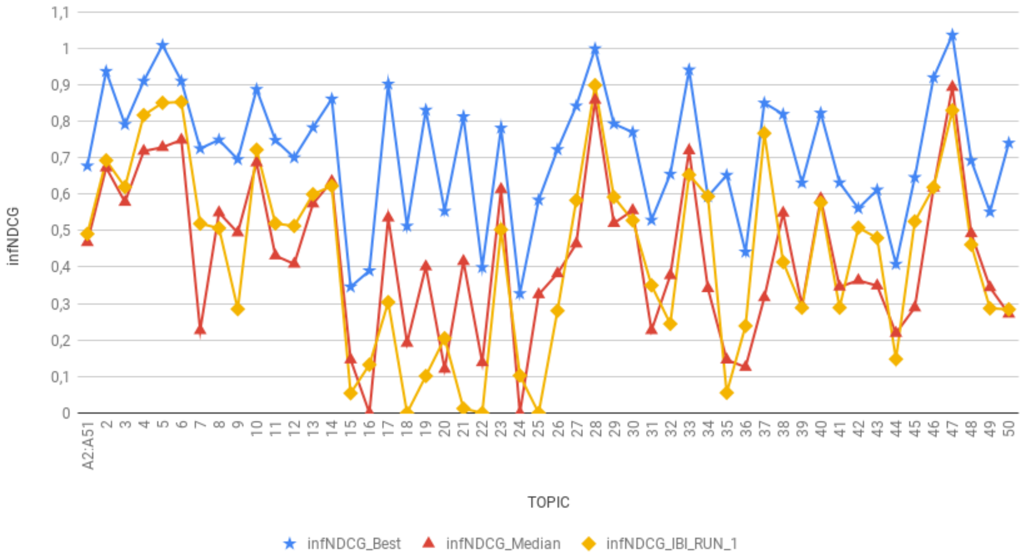

Figure 2: Inferred Normalized Discounted Cumulative Gain (infNDCG) of best run (RUN 1) per topic (together with median and best infNDCG among all PM Track runs)

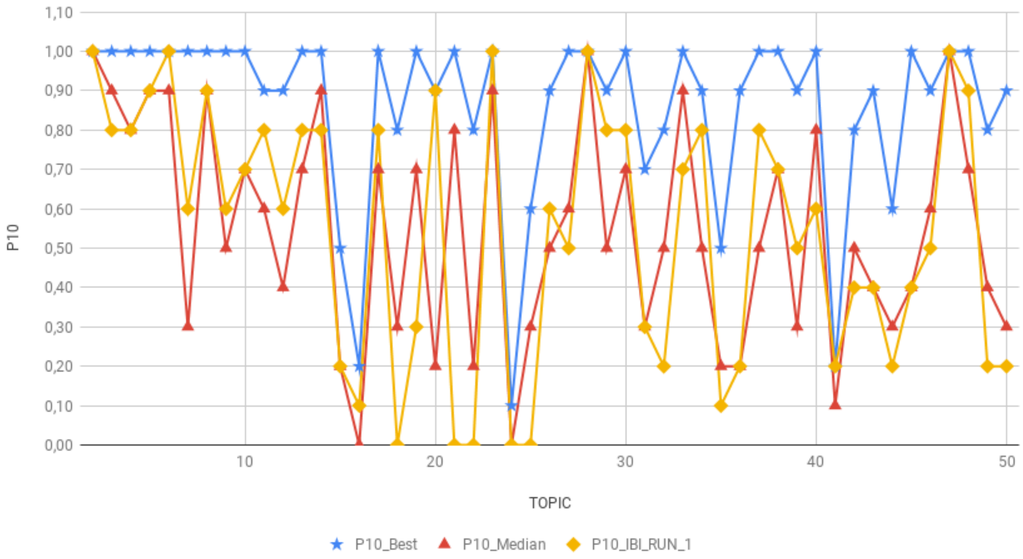

Figure 3: Precision@10 (P10) of best run (RUN 1) per topic (together with median and best P10 among all PM Track runs)

The national center for biomedical ontology. *Journal of the American Medical Informatics Association* 19(2):190–195.

David Tamborero, Carlota Rubio-Perez, Jordi Deu-Pons, Michael P Schroeder, Ana Vivancos, Ana Rovira, Ignasi Tusquets, Joan Albanell, Jordi Rodon, Josep Tabernero, et al. 2018. Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome medicine* 10(1):25.

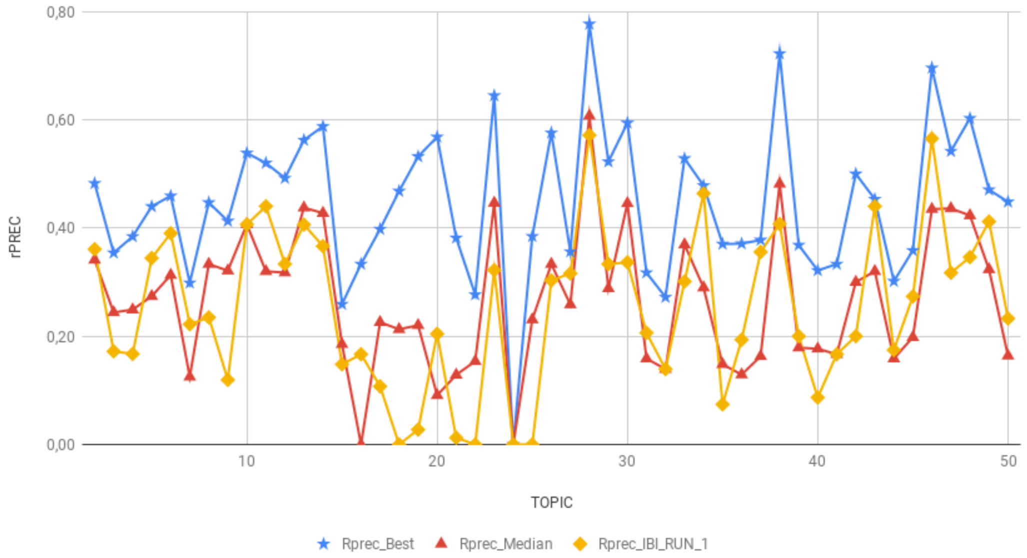

Figure 4: R-precision (rPREC) of best run (RUN 1) per topic (together with median and best rPREC among all PM Track runs)