Identifying Personalised Treatments and Clinical Trials for Precision Medicine using Semantic Search with Thalia

Piotr Przybyła, Axel J. Soto, and Sophia Ananiadou

National Centre for Text Mining, School of Computer Science,
The University of Manchester, UK

Abstract

This paper reports the main methods applied in our submission to TREC 2017 Precision Medicine Track. The goal of this challenge was to retrieve documents containing potential treatments and clinical trials for specific patient characteristics. Our main strategy involved using a semantic search engine called Thalia (Text mining for Highlighting, Aggregating and Linking Information in Articles), which allows the recognition of diseases and genes mentioned in text. The recognition of named entities and its linking to concepts in ontologies facilitates more accurate retrieval than just relying on plain textual search and matching. We also highlight the different strategies applied when querying Thalia in the context of this Precision Medicine challenge, which aimed to support different use cases (i.e. more focused or broader searches).

1 Introduction

Precision medicine is used to describe a recent trend that aims to provide treatment to patients based on their individual characteristics [22, 1]. This can help achieve better results than generic approaches, since patients' features, especially those present in their genome, greatly influence the effectiveness of possible therapies [6]. However, an important challenge needs to be addressed before the health professional can select an optimal course of treatment based on the patient's characteristics. A great deal of research on the relationship between treatments and genetic variances is available, but most of it remains scattered and hidden in the vast body of scholarly literature. As a result, this information is not readily available to clinicians. The lack of support tools, e.g. to suggest relevant research publications, is one of the major obstacles in precision medicine, which hinders the provision of relevant information inf context for a particular patient [16].

TREC 2017 Precision Medicine Track[1] aims to address this problem, focusing on the retrieval of treatment-related information for cancer patients from two sources: abstracts of articles (indexed in PubMed[2]) and conference proceedings, and clinical trial descriptions from the ClinicalTrials.gov database[3].

Our approach to address this challenge was based on the use of a biomedical semantic search engine called Thalia (Text mining for Highlighting, Aggregating and Linking Information in Articles), which has been developed at our research centre. The main purpose of Thalia is to enable semantic search in the context of biomedical literature by leveraging previous named entity (NE) annotation efforts. The key strategy to achieve a semantic behaviour is to normalise NEs, i.e. link entities to concepts in an openly available ontology, which effectively allows to map a concept with its multiple word forms. Thalia covers the entire PubMed, which at the point of this challenge contained about 27 million references.

The remaining of the paper describes related work in the area of precision medicine, the main methods applied to handle this problem and a discussion of the results. The methodology applied includes: annotating entities through text mining workflows, indexing of documents in a search engine, and the representation of the patient data and the different types of queries applied. Five different query strategies were provided for each sub task: this was done both to compare different matching strategies and to account for different user needs, i.e. more focused (higher precision) or broader (higher recall) searches.

2 Related work

Computational methods have been shown to be the cornerstone of precision medicine [13, 24, 10]. In particular, text mining has contributed to the advance of the field in different ways. For instance, several works have addressed the extraction of different types of associations from large bodies of research publications. Such relationships could be gene mutations and diseases [28], protein-variants and diseases [27], or disease-gene-variant triplets [29]. In the case of the work by Wei et al. [34], the authors go a step further aiming to normalise gene variants to unique identifiers.

A recent article by Breitenstein et al. [3] reported how a rule-based text mining approach can be used to infer breast cancer receptor status from electronic health records. This information can be complemented with structured information contained in cancer registries to provide the type of cancer treatment that is optimal for a given patient. Integration of knowledge extracted from unstructured data still requires addressing challenges that go from heterogenous standards for annotation to security and ethical factors [4].

Mining entities and relationships is a first step towards the goal of untapping hidden knowledge from vast amounts of biomedical literature. A consequent step

[1] `http://www.trec-cds.org/2017.html`
[2] `https://www.ncbi.nlm.nih.gov/pubmed/`
[3] `https://www.clinicaltrials.gov/`

is to provide search mechanisms that would allow medical doctors to retrieve articles or clinical trials that describe potential treatment for the particularities of patients. Several articles have proposed engines for searching entities and relationships, i.e. semantic search engines in PubMed [31, 21], clinical trials [5], or electronic health records [35, 12]. Recently, a semantic search engine has been made available, which specializes in personalised cancer therapies [17].

In summary, although there have been several attempts to use elements of text mining for precision medicine purposes, few works addressed the problem of providing support for therapeutic decisions in a clinical context. The TREC precision medicine track offers an excellent opportunity to invigorate research in this direction.

3 Task

The goal of the Precision Medicine Track at TREC 2017 is to support clinical decision by providing helpful information for a particular patient. The patient cases (referred to as *topics*) are described using the following attributes:

- target disease (a type of cancer),
- relevant genes and their variants present,
- demographic information: age and gender,
- other medical conditions.

Each of these attributes is provided as plain text, e.g. *12-year-old male* or *CDK4 Amplification*. The whole task consists of 30 topics.

There are two document collections to retrieve relevant information from: scientific abstracts and clinical trials. The first one is based on a January 2017 snapshot of PubMed (over 26 million documents) plus additional abstracts from proceedings of the American Association for Cancer Research[4] and American Society of Clinical Oncology[5] conferences (70,025 documents). The second collection contains a snapshot from April 2017 of the ClinicalTrials.gov[6] database (241,006 documents). According to the provided guidelines[7], a document is considered relevant for a topic if it:

- focuses on the same, more specific or more general disease than the one in the topic,
- takes into account the gene from the topic and either includes the same variant or does not specify any variant, and

[4] `http://www.aacr.org/`

[5] `https://www.asco.org/`

[6] `https://clinicaltrials.gov/`

[7] `http://trec-cds.appspot.com/2017.html` and `http://trec-cds.appspot.com/relevance_guidelines.pdf`

- does not exclude from the trial or study patients with the provided demographics and conditions.

In the case of the clinical trials task, relevance implies that the corresponding patient would have been eligible to participate in the trial. For a scientific abstract there is an additional requirement, which is to include information useful in the clinical context, e.g. about treatment or prevention of the disease prognosis.

The output for each topic consists of two ranked lists of the identifiers of up to 1000 relevant documents: one with scientific abstracts and the other one with clinical trials.

4 Methods

This section reports the main methodological steps of our work. We describe Thalia's building blocks (i.e. NE recognisers and document indexing), its customisation for the challenge, and how we converted the topics into search queries to retrieve the recommended documents for the patients.

4.1 Named entity recognition

In order to allow semantic search capabilities, Thalia makes use of NE recognisers developed as components of Argo [26], which is a web-based text mining system. Since Argo is based on the UIMA architecture [9], its workflows could be deployed in several ways. Given the heavy computational burden of annotating the whole PubMed, we created a standalone executable using the uimaFIT library[8], which was then ran on Computational Shared Facility (CSF), a cluster infrastructure available at the University of Manchester. Each task consists of reading a portion of PubMed, annotating NEs in text, and outputting them in the PubAnnotation format [15], which was extended to JSON-LD [18] by replacing object strings with links to ontologies. Thalia includes annotation of several types (Chemicals, Diseases, Drugs, Genes, Metabolites, Proteins, Species and Anatomical entities), but for the purpose of this task we used just two of them:

- **Genes** recognised by a conditional random field (CRF)-based model trained on the BioCreative II GM Track training corpus [25, 30], which was then followed by normalisation (linking) to the HGNC ontology [11];
- **Diseases** recognised by a CRF-based model trained on the NCBI disease training set [25, 8], which was then followed by normalisation to the UMLS Metathesaurus ontology [19].

To make this possible, the following UIMA components for biomedical language processing were pipelined:

[8] `https://uima.apache.org/uimafit.html`

- LingPipe sentence splitter[9],
- OSCAR 4 tokeniser [14],
- GENIA part of speech tagger [33],
- NERSuite CRF tagger[10],
- Acromine acronym recogniser [23],
- Concept normaliser [2],
- Binary UIMA CAS writer and reader from the DKPro component collection [7].

4.2 Search engine

Once the NER workflows were run on the whole corpus, the abstract text and its annotations were then indexed on Elasticsearch[11], along with any metadata. In order to adapt Thalia for this challenge, we also incorporated abstracts from AACR and ASCO. The index allows to search in a matter of seconds over the whole collection by providing a textual keyword or choosing a concept from an ontology. Standard tokenisation was applied on textual content, while for the annotations we indexed their word forms as well as their identifiers in the ontology.

For the clinical trials data, we needed to select fields that we identified as of interest for indexing, namely: *official title*, *brief summary*, *detailed description*, *age eligibility*, *gender eligibility* and *eligibility criteria*. The *eligibility criteria* field needed some preprocessing as it could contain inclusion as well as exclusion conditions. Upon examination of multiple cases, we manually generated rules to separate inclusion from exclusion criteria, which involved the identification of negation clauses as well as regular expressions to detect whether the text referred to inclusion or exclusion criteria.

There are two means of accessing the search index: one is via a web user interface, while the other one is via a RESTful API. Given the nature of the task, which required to experiment with different query strategies, the API was a more natural fit for this task. We developed our own query language for the API, which allows to search for textual strings or concepts, which can be also combined in a Boolean expression.

4.3 Topic representation

In the process of generating queries, each *topic*, i.e. a patient case, could be represented by the following data:

[9] `http://alias-i.com/lingpipe/`
[10] `http://nersuite.nlplab.org/`
[11] `https://www.elastic.co/`

1. **Target** disease to be treated, both as its name (*Liposarcoma*) and UMLS Metathesaurus concept identifier (C0023827);

2. List of **Genes** that have to be taken into account, also as a name (*KRAS*) and an HGNC identifier (6407), and optionally a corresponding **Variant** represented as a text string, including proper names (*V600E*) and modification types (*Amplification*);

3. List of **Other** conditions, represented in the same way as **Target**;

4. Demographic data: **Gender** as a Boolean (male of female) variable and **Age** as a number.

We use a script to automatically extract the above information from the topics XML file provided by the organisers of the challenge. To find the concept identifier for a given name of a disease or a gene we use a Thalia feature that returns the most commonly associated identifier with a given name based on the whole PubMed corpus.

4.4 Query construction

This subsection contains a description of the process of converting the topics represented as above into queries for Thalia. First, we present the techniques used as building blocks for queries (Subsections 4.4.1–4.4.4) and then we show the list of five queries that were prepared to search in abstracts (Subsection 4.4.6) and in the clinical trials (Subsection 4.4.7).

Each of the query strategies was used to retrieve lists of documents corresponding to particular topics, which were then submitted to the task organisers as *runs*. Since every run was created by using a corresponding query, we will use these terms interchangeably.

4.4.1 Matching modes

Since diseases and genes are semantically interpreted by Thalia, several modes of matching between a target NE and candidate occurrences in documents are possible:

- **T** (textual): a term is matched if it occurs in a document in the same textual form (allowing differences in number or letter case);

- **S** (semantic): a term is matched if the NE identifier is the same and regardless of whether a different name is used;

- **M** (mixed): a term is matched if either textual or semantic matching happens.

4.4.2 Fields of interest

Both scientific abstracts and clinical trials have several fields that can contain NEs, yet may not be relevant for this challenge. For the PubMed task, we concatenated the contents of title and abstract and treated them as a single document. In the case of the clinical trials the following fields were processed separately:

- official **Title**: one-sentence summary of a trial;
- brief **Summary**: one-paragraph description of the trial, including its main goals;
- detailed **Description**: several paragraphs of text, including details of the study, but also background, related work, etc;
- **Exclusions** in eligibility criteria: a list of conditions that disqualify a participant from the trial. Note that the exclusions were inferred from the eligibility criteria field.

4.4.3 Ontological expansion

The normalisation of NE against ontologies in Thalia allows us to manually explore the ontology structure and find concepts related to the target one, so that they can be used in the queries. In this challenge, we applied the following manual procedure:

1. For a given **Target** disease concept, we browsed the UMLS Metathesaurus to select:
   (a) the main concept corresponding to the disease,
   (b) the concept automatically recognised as most common for the given name by Thalia (see Section 4.3),
   (c) all the concepts immediately connected to the main concept with RN (narrower meaning) or RB (broader meaning) UMLS relations[12].
2. Replaced the original concept from the query with a list of all the concept identifiers found in the previous step.

This procedure, denoted as **Neighbours()**, allows us to broaden the search by accepting occurrences of related diseases in addition to the original one.

[12] If the main concept has only one relation, we use this single related concept to find RN and RB neighbours.

4.4.4 Precision medicine keywords

For the sake of suggesting articles that contain potential treatment for the patient, we can include an additional requirement: an abstract has to contain words that are likely to indicate that a medical treatment is mentioned. To account for this, we compiled a list of keywords based on manual inspection of the example documents, namely: *therapy, therapeutic, diagnosis, patient, target, profiling, prognosis, prognostic, predict.* A criterion denoted by **PMWords** is satisfied if at least one of these words is found in a document.

4.4.5 Demographic eligibility

For a clinical trial to be relevant for a topic, it needs to investigate the disease of interest, but also the patient has to be eligible for participation. This requires an extra matching criterion, **Demographic**, which is satisfied if both following conditions are met:

1. patient's age being between a minimum and maximum defined in a clinical trial;
2. the clinical trial accepting patients of the specified gender.

4.4.6 Queries for scientific abstracts

We combined the concepts explained above into the query strategies enumerated below. In each run, for every topic a maximum of 1000 documents are allowed as a result. Since some of the methods described below return less hits, we cascaded the queries—i.e. we run several sub-queries in sequence. This means that if the first query does not reach the maximum number of documents, the following query is used to add more (distinct) results. This is repeated until 1000 documents are retrieved or all the cascading queries are used. This cascading operation is denoted by '+'. The '−' operator has the opposite meaning: from the list of results of the first query, we remove those returned by the second query. This is helpful to exclude certain results, e.g. those clinical trials that a patient is not eligible for due to the conditions listed as **Other**.

1. **Textual**:
   `T(Disease) OR [ T(Gene) OR T(Variant) ]`[13]
   This is a baseline method with textual matching of crucial topic elements, connected through disjunction. It is the least restrictive query, hence it is aimed to achieve high recall.
2. **StrictTextPM**[14]:
   `T(Disease) AND [ T(Gene)] AND PMWords`
   This method uses text matching as well, but it is more strict, as it expects

[13] [...] denotes an alternative among multiple potential items in the brackets, e.g. in this case, different genes and variants.

[14] This method was not submitted on its own, but only as part of the **Broad** run.

both disease and genes to occur and also requires presence of the precision medicine words.

3. **Focused**:
   `S(Disease) AND [ S(Gene) AND T(Variant) ] AND PMWords`
   This is the strictest variant: it requires a disease, a gene and its variant to be matched semantically (gene variants are not annotated, so textual matching is used for them). It also expects the precision medicine words to be included. This is aimed to achieve the highest precision.

4. **Semantic**:
   `(S(Disease) AND [ S(Gene) AND T(Variant) ] AND PMWords) + ( S(Disease) AND [ S(Gene)] AND PMWords ) + ( S(Disease) AND [ S(Gene)] )`
   The main semantic matching method, which uses genes, variants and a target disease. It consists of three sub-queries in decreasing order of strictness.

5. **Broad**:
   `Semantic + StrictTextPM + Textual`
   A hybrid query that contains five cascading sub-queries ordered by decreasing strictness. This is expected to provide the most likely relevant results on top of the list, but also include others that might be relevant further down.

6. **Ontological**:
   `(S(Neighbours(Disease)) AND [ S(Gene) AND T(Variant) ] AND PMWords) + ( S(Neighbours(Disease)) AND [ S(Gene)] AND PMWords ) + ( S(Neighbours(Disease)) AND [ S(Gene)] )`
   This query is similar to **Semantic**, but in addition to a fixed concept for a disease it also uses all its (manually selected) narrower and broader neighbours. This should give some results that do not necessarily include the same disease, but remain relevant (see TREC PM relevance guidelines).

The five strategies above, i.e. **Textual**, **Focused**, **Semantic**, **Broad**, **Ontological** are used to obtain relevant document lists for each topic.

4.4.7 Queries for clinical trials

To find relevant clinical trials, just like for the scientific abstract task, we propose **Textual**, **Semantic**, **Broad**, **Focused** and **Ontological** queries, i.e:

1. **Textual**:
   `( T(Disease,Title) AND [ T(Gene,Title) ] AND Demographic ) + ( T(Disease,Summary) AND [ T(Gene,Summary) ] AND Demographic ) + ( T(Disease,Description) AND [ T(Gene,Description) ] AND Demographic ) - [ T(Other,Exclusions) ]`
   This is a text-based query, requiring that both the disease name and at

least one of the gene names are present and the demographic criteria are met. A cascading mechanism is used to prioritise the occurrences in title, followed by brief description and detailed description. This follows the assumption that trials mentioning the entities of interest in the title are more likely to be relevant than those including them only in the detailed description. The last argument indicates that if any of the patient's *other* conditions is mentioned in the exclusion list, the trial will not be retrieved.

2. **Semantic**:
   `( S(Disease,Title) AND [ S(Gene,Title) ] AND Demographic ) + ( S(Disease,Summary) AND [ S(Gene,Summary) ] AND Demographic ) + ( S(Disease,Description) AND [ S(Gene,Description) ] AND Demographic ) - [ T(Other,Exclusions) ]`
   The same procedure as above, but using a semantic matching instead of a textual one (except for **Exclusions**, which uses the textual one).

3. **Broad**:
   `( M(Disease,Title) AND [ M(Gene,Title) ] AND Demographic ) + ( M(Disease,Summary) AND [ M(Gene,Summary) ] AND Demographic ) + ( M(Disease,Description) AND [ M(Gene,Description) ] AND Demographic ) - [ T(Other,Exclusions) ]`
   This method uses mixed matching, which is intended to give the broadest list of results.

4. **Focused**:
   `( S(Disease,Title) AND [ S(Gene,Title) ] AND Demographic ) + ( S(Disease,Summary) AND [ S(Gene,Summary) ] AND Demographic ) - [ T(Other,Exclusions) ]`
   This is similar to the semantic query, but without taking into account documents that mention the target disease in the detailed description only. This seeks to retrieve less documents, but with higher relevance.

5. **Ontological**:
   `( S(Neighbours(Disease),Title) AND [ S(Gene,Title) ] AND Demographic ) + ( S(Neighbours(Disease),Summary) AND [ S(Gene,Summary) ] AND Demographic ) + ( S(Neighbours(Disease), Description) AND [ S(Gene,Description) ] AND Demographic ) - [ T(Other,Exclusions) ]`
   A variant of the semantic query that uses manually selected identifiers of related concepts in the ontology for matching the target disease.

5 Results

This section presents the results of our five query strategies applied on each of the tasks, i.e. the relevancy of documents in the corresponding runs. A comparison against the median performance by all the participants in the track is also included.

5.1 Scientific abstracts

In the case of the scientific abstracts, for each topic the set of documents that were labelled by the judges included all documents in top-10 ranks, i.e. depth-10 pools [32], plus a 30% random sample of the union of all documents retrieved at ranks 11–50. This union was taken over all runs submitted by all the participants on the same topic, and a document was considered for sampling if the best rank at which it was ever retrieved was in that range.

Three main evaluation measures were considered here: an estimation of the normalised discounted cumulative gain (infNDCG) [36], precision at 10 (P10), and R-precision (R-prec). A reference for the latter two is provided by Manning et al. [20].

The performance for the five runs averaged over all topics is found in Table 1. The same results are shown as a bar plot in Figure 1. Since 29 other teams participated in the track (with a total of 125 runs), for the sake of comparison we also added to the plot the median results of all the participating teams.

Table 1: Retrieval performance for the scientific abstracts task measured by infNDCG, P10 and R-prec on our five runs.

	Textual	Semantic	Ontological	Focused	Broad
infNDCG	0.3294	0.3561	0.1793	0.2536	**0.3800**
P10	0.4333	0.4600	0.4200	0.3633	**0.4667**
R-prec	0.2043	0.2078	0.1725	0.1252	**0.2287**

The results show some interesting aspects. The run using the **Semantic** query provided better results according to all evaluation measures compared to the **Textual** run. However, the **Broad** query, which relaxes the requirements to make it more inclusive by incorporating disjunctions as well as a cascading strategy, proved to perform the best. On the negative side, the **Focused** run missed too many relevant documents due to its restrictiveness. Despite having a manual component, the *Ontological* run contained the least relevant documents. We also note that the ranking of our runs were consistent across all measures.

For the best performing run, i.e. **Broad**, we also show the performance in terms of infNDCG across all topics in Figure 2. As a reference, we also include the best score by any team for each topic as well as the median performance. The figure shows how the **Broad** strategy outperformed the median scores consistently, and it is among the top performing approaches for topics #1 and #3.

5.2 Clinical trials

The ground truth labels for the Clinical Trials task were created taking the union of the top 15 results for all participants' runs (i.e. depth-15 pools). The organisers expected very few relevant trials, so only precision at 5, 10 and 15,

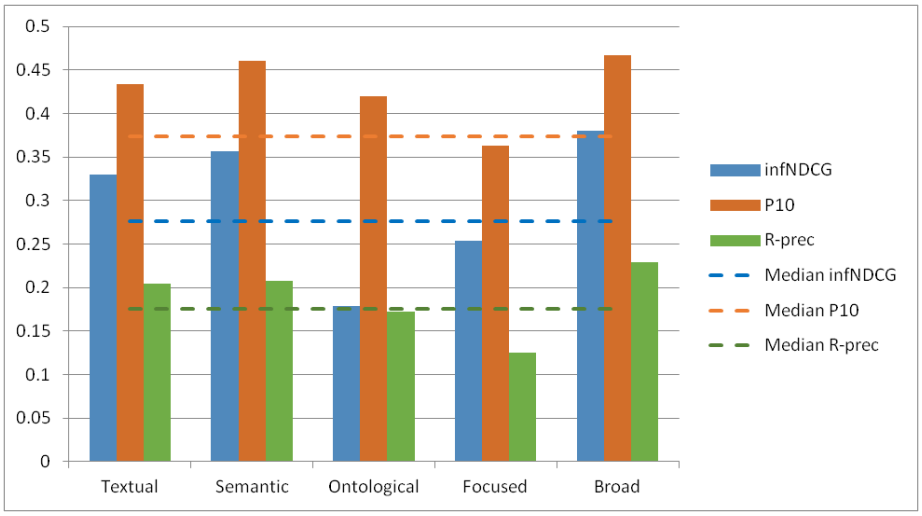

Figure 1: Retrieval performance for the scientific abstracts task measured by infNDCG, P10 and R-prec on our five runs. Dashed lines correspond to the median performance of all participants in the track.

i.e. P5, P10 and P15, are considered here. Also, topic #10 did not have any relevant documents, and as a result this topic was discarded by the organisers.

The performance for the five runs averaged over all topics is found in Table 2. The same results are shown as a bar plot in Figure 3, where the median results of all the participating teams is added for reference.

Table 2: Retrieval performance for the clinical trials task measured by P5, P10 and P15 on our five runs.

	Textual	Semantic	Ontological	Focused	Broad
P5	0.4207	0.3034	0.3655	0.2138	**0.4483**
P10	0.3276	0.2172	0.2862	0.1448	**0.3724**
P15	0.2759	0.1678	0.2253	0.1126	**0.3080**

To a certain degree, results are analogous to the ones obtained in the other task. The **Textual** run ranks fairly high compared to the median performance. However, limiting to semantic matching only seems to deteriorate the retrieval performance. The **Broad** run performed the best for all metrics, which shows that a mixed matching strategy results in a more robust retrieval. While the **Ontological** run performed comparably better than for the other task, it is still quite below the two best methods.

For the best performing run, i.e. **Broad**, we also show the performance in terms of P5 across all topics in Figure 4. As a reference, we also include the best score by any team for each topic as well as the median performance. The

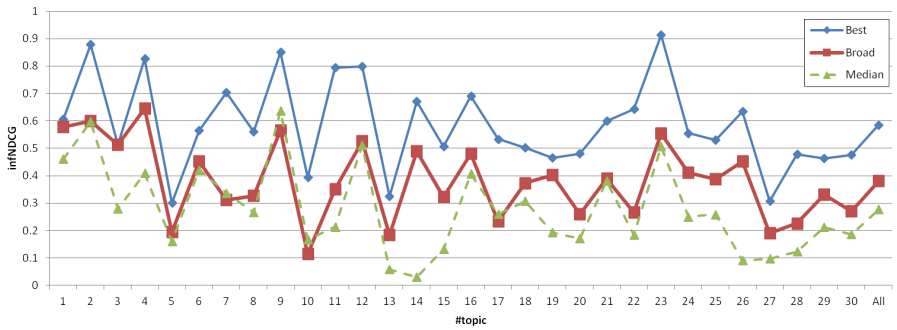

Figure 2: Retrieval performance for the Scientific Abstracts task in terms of infNDCG for each topic for the **Broad** run. Best and median performance for each topic is also shown for reference.

figure shows how the **Broad** strategy outperformed the median score for most of the topics, and it is among the top performing approaches for six different topics.

6 Discussion

Our main research question was testing whether a general-purpose semantic search engine for biomedical texts, such as Thalia, can be useful in a precision medicine context. The answer to this question is positive: the performance of most of our queries is well above the median computed across all the submitted solutions, the vast majority of which were designed just for this particular task.

We used the opportunity of submitting five runs per task to try out various approaches. The resulting differences in performance are substantial and consistent across measures, which allows to draw some conclusions on the underlying techniques.

The **Textual** baseline, treating entity names as mere text strings, is outperformed by only one or two methods and not every modification that was intended to improve it turned out to be beneficial. A fundamental change is introduced by the **Semantic** query. Since it involves matching conditions and genes mentioned in the patient description as concepts, not strings, this would enable proper handling of both synonymy (retrieving documents including other names of the same concept) and polysemy (excluding documents containing occurrences of the same name but referring to a different meaning). The effect on retrieval accuracy, however, appears to vary with respect to the document type: although we observe gains for scientific abstracts, it is not the case for clinical trials. We hypothesise that the reason for this mixed result is that the machine learning models included in NE recognition algorithms were trained on scholarly text and perform worse on clinical trial descriptions, which impedes the semantic matching. These problems are even more noticeable in the results

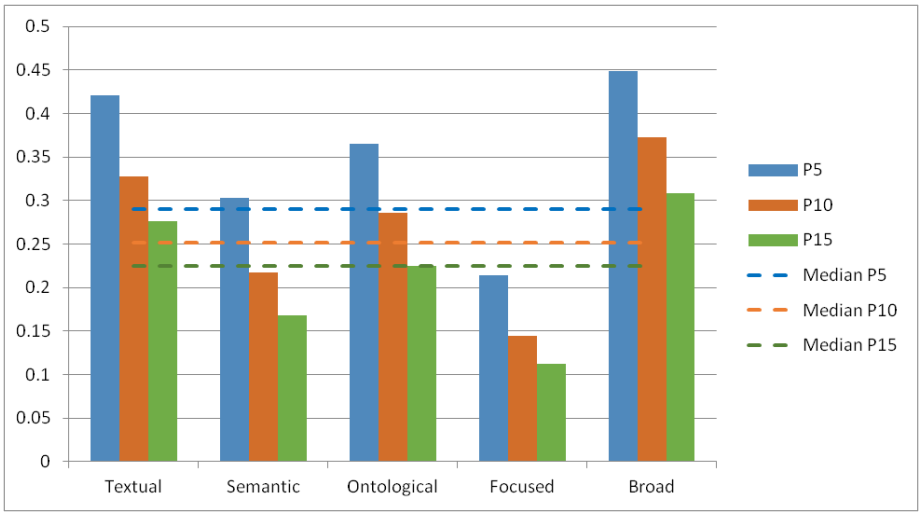

Figure 3: Retrieval performance for the clinical trials task measured by P5, P10 and P15 on our five runs. Dashed lines correspond to the median performance of all runs in the track.

of the **Focused** run, which requires both the genes and conditions to be semantically recognised, which does not happen in many relevant documents. This issue is mitigated by a hybrid approach, which involves combining the semantic and text-based techniques through mixed matching or cascading. The solution implemented as the **Broad** runs achieve better performance than any other of our proposed strategies, and in the case of several topics obtained the best results in the shared task.

The **Ontological** query, which extends the **Semantic** one by involving manual selection of related alternatives for conditions, also behaves differently with respect to the type of document. While this strategy allows to find more clinical trials, its performance in scientific abstracts suffers due to the presence of many highly-ranked irrelevant documents. Apart from the semantic matching problems mentioned above, the issue may lie in the discrepancy of relations of 'more general' and 'more specific' between UMLS Metathesaurus and evaluation judges. For example, it was not strictly defined how much more general a condition name could be to remain relevant for a particular case.

Clearly, many challenges need to be addressed before biomedical search engines, such as Thalia, could match users' queries to documents in a fully semantic way. Nevertheless, the presented results suggest that this is a direction worth pursuing and also show the benefits of searching beyond string-based keywords.

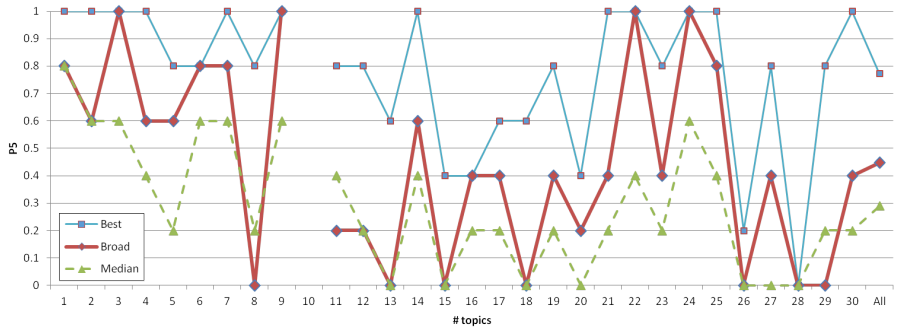

Figure 4: Retrieval performance in terms of P5 for each topic for the **Broad** run in the clinical trials task. Best and median performance for each topic is also shown for reference.

Acknowledgements

This work is jointly supported by the MRC project: Manchester Molecular Pathology Innovation Centre (MMPathIC), grant ID: MR/N00583X/1 and the BBSRC project: Enriching Metabolic PATHwaY models with evidence from the literature (EMPATHY), grant ID: BB/M006891/1.

References

[1] An, G., and Vodovotz, Y. What is Precision Medicine and can it work? `https://www.elsevier.com/connect/what-is-precision-medicine-and-can-it-work`, 2015.

[2] Batista-Navarro, R., Carter, J., and Ananiadou, S. Argo: enabling the development of bespoke workflows and services for disease annotation. *Database: The Journal of Biological Databases and Curation* (2016).

[3] Breitenstein, M. K., Liu, H., Maxwell, K. N., Pathak, J., and Zhang, R. Electronic health record phenotypes for precision medicine: Perspectives and caveats from treatment of breast cancer at a single institution. *Clinical and Translational Science 11*, 1 (2018), 85–92.

[4] Castaneda, C., Nalley, K., Mannion, C., Bhattacharyya, P., Blake, P., Pecora, A., Goy, A., and Suh, K. S. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of clinical bioinformatics 5*, 1 (2015), 4.

[5] Chen, X., Chen, H., Bi, X., Gu, P., Chen, J., and Wu, Z. Biotcm-se: a semantic search engine for the information retrieval of modern biology and traditional chinese medicine. *Computational and mathematical methods in medicine* (2014).

[6] Dancey, J., Bedard, P., Onetto, N., and Hudson, T. The Genetic Basis for Cancer Treatment Decisions. *Cell 148*, 3 (2012), 409–420.

[7] de Castilho, R., and Gurevych, I. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT* (Dublin, Ireland, 2014), Association for Computational Linguistics and Dublin City University, pp. 1–11.

[8] Doan, R. I., Leaman, R., and Lu, Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics 47* (2014), 1–10.

[9] Ferrucci, D., and Lally, A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering 10*, 3-4 (2004), 327–348.

[10] Gonzalez, G. H., Tahsin, T., Goodale, B. C., Greene, A. C., and Greene, C. S. Recent advances and emerging applications in text and data mining for biomedical discovery. *Briefings in Bioinformatics 17*, 1 (2016), 33–42.

[11] Gray, K. A., Yates, B., Seal, R. L., Wright, M. W., and Bruford, E. A. Genenames.org: The HGNC resources in 2015. *Nucleic Acids Research 43*, D1 (2015), D1079–D1085.

[12] Hanauer, D. A., Mei, Q., Law, J., Khanna, R., and Zheng, K. Supporting information retrieval from electronic health records: A report of university of michigans nine-year experience in developing and using the electronic medical record search engine (emerse). *Journal of biomedical informatics 55* (2015), 290–300.

[13] Holzinger, A., Dehmer, M., and Jurisica, I. Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC bioinformatics 15*, 6 (2014).

[14] Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L., and Murray-Rust, P. OSCAR4: A flexible architecture for chemical textmining. *Journal of Cheminformatics 3*, 10 (2011).

[15] Kim, J.-D., and Wang, Y. PubAnnotation: a persistent and sharable corpus and annotation repository. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing* (2012), Association for Computational Linguistics, pp. 202–205.

[16] Kohane, I. S. Ten things we have to do to achieve precision medicine. *Science 349*, 6243 (2015), 37–8.

[17] Kurnit, K. C., Bailey, A. M., Zeng, J., Johnson, A. M., Shufean, M. A., Brusco, L., Litzenburger, B. C., Sánchez, N. S., Khotskaya, Y. B., Holla, V., Simpson, A., Mills, G. B., Mendelsohn, J., Bernstam, E., Shaw, K., and Meric-Bernstam, F. Personalized cancer therapy: A publicly available precision oncology resource. *Cancer Research 77*, 21 (2017), e123–e126.

[18] Lanthaler, M., and Gütl, C. On using JSON-LD to create evolvable RESTful services. In *Proceedings of the Third International Workshop on RESTful Design* (2012), ACM, pp. 25–32.

[19] Lindberg, D. A., Humphreys, B. L., and McCray, A. T. The Unified Medical Language System. *Methods of information in medicine 32*, 4 (1993), 281–91.

[20] Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to information retrieval*, vol. 1. Cambridge university press, 2008.

[21] Müller, B., Poley, C., Pössel, J., Hagelstein, A., and Gübitz, T. Livivo–the vertical search engine for life sciences. *Datenbank-Spektrum 17*, 1 (2017), 29–34.

[22] National Research Council. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. National Academies Press, Washington, D.C., 2011.

[23] Okazaki, N., and Ananiadou, S. Building an abbreviation dictionary using a term recognition approach. *Bioinformatics 22*, 24 (2006), 3089–3095.

[24] Peterson, T. A., Doughty, E., and Kann, M. G. Towards precision medicine: advances in computational approaches for the analysis of human variants. *Journal of molecular biology 425*, 21 (2013), 4047–4063.

[25] Rak, R., Batista-Navarro, R. T., Carter, J., Rowley, A., and Ananiadou, S. Processing biological literature with customizable Web services supporting interoperable formats. *Database: The Journal of Biological Databases and Curation* (2014).

[26] Rak, R., Rowley, A., Black, W., and Ananiadou, S. Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database: The Journal of Biological Databases and Curation* (2012).

[27] Ravikumar, K. E., Wagholikar, K. B., Li, D., Kocher, J.-P., and Liu, H. Text mining facilitates database curation-extraction of mutation-disease associations from bio-medical literature. *BMC bioinformatics 16*, 1 (2015), 185.

[28] Singhal, A., Simmons, M., and Lu, Z. Text mining for precision medicine: automating disease-mutation relationship extraction from biomedical literature. *Journal of the American Medical Informatics Association 23*, 4 (2016), 766–772.

[29] Singhal, A., Simmons, M., and Lu, Z. Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine. *PLoS computational biology 12*, 11 (2016).

[30] Smith, L., Tanabe, L. K., Ando, R., Kuo, C.-J., Chung, I.-F., Hsu, C.-N., Lin, Y.-S., Klinger, R., Friedrich, C. M., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C. A., Povinelli, R. J., Vlachos, A., Baumgartner, W. A., Hunter, L., Carpenter, B., Tsai, R., Dai, H.-J., Liu, F., Chen, Y., Sun, C., Katrenko, S., Adriaans, P., Blaschke, C., Torres, R., Neves, M., Nakov, P., Divoli, A., Maña-López, M., Mata, J., and Wilbur, W. J. Overview of BioCreative II gene mention recognition. *Genome Biology 9*, Suppl 2 (2008).

[31] Thomas, P., Starlinger, J., Vowinkel, A., Arzt, S., and Leser, U. Geneview: a comprehensive semantic search engine for pubmed. *Nucleic Acids Research 40*, W1 (2012), W585–W591.

[32] Tonon, A., Demartini, G., and Cudré-Mauroux, P. Pooling-based continuous evaluation of information retrieval systems. *Information Retrieval Journal 18*, 5 (2015), 445–472.

[33] Tsuruoka, Y., Tateishi, Y., Kim, J. D., Ohta, T., McNaught, J., Ananiadou, S., and Tsujii, J. Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics* (2005), vol. 3746 LNCS, pp. 382–392.

[34] Wei, C.-H., Phan, L., Feltz, J., Maiti, R., Hefferon, T., and Lu, Z. tmvar 2.0: integrating genomic variant information from literature with dbsnp and clinvar for precision medicine. *Bioinformatics 34*, 1 (2018), 80–87.

[35] Wu, H., Toti, G., Morley, K. I., Ibrahim, Z. M., Folarin, A., Jackson, R., Kartoglu, I., Agrawal, A., Stringer, C., Gale, D., et al. Semehr: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *Journal of the American Medical Informatics Association* (2018).

[36] Yilmaz, E., Kanoulas, E., and Aslam, J. A. A simple and efficient sampling method for estimating ap and ndcg. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (2008), ACM, pp. 603–610.