LIMSI @ 2015 Clinical Decision Support Track

Eva D'hondt, Brigitte Grau and Pierre Zweigenbaum

LIMSI, CNRS
Université Paris-Saclay
91400 Orsay, France

firstname.lastname@limsi.fr

Abstract

In this paper we present our participation to the 2015 TREC Clinical Decision Support (CDS) Track. The goal of this track is to find relevant medical literature for a given medical case report which should help address a specific clinical aspect of the case (known as the Clinical Question Type). We investigated the relative merit of concept- versus word-based representations of the information contained in the case reports, and experimented with different approaches to model Clinical Question Types. We submitted six runs in total, three for subtask A (CDS search without information on the patient's diagnosis) and subtask B (CDS search with information on the patient's diagnosis for topics from the *Test* and *Treatment* Clinical Question Types). In both subtasks our best runs were MeSH-based, and they achieved a P@10 and infNDCG of 0.37 and 0.25, and 0.47 and 0.35, for subtask A and B respectively.

Keywords

Document Retrieval, UMLS, MeSH, Clinical Decision Support

1 Introduction

Similar to the 2014 track, the goal of the 2015 Clinical Decision Support Track was to retrieve relevant scientific articles in the PubMed Open Access Central set[1] for a given medical case report. A medical case report is a short and well-formed description of a patient's medical record which summarizes the pertinent information for that case. In the CDS track, case reports are provided in two levels of detail: Case *summaries* are short descriptions that contain one or two sentences at most, for example,

> Summary: 26-year-old obese woman with bipolar disorder, on zolpidem and lithium, with recent difficulty sleeping, agitation, suicidal ideation, and irritability.

[1] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

Case *descriptions*, on the other hand, are more detailed and are written in less terse, more natural style. They contain additional contextual information on the patient's condition and history, for example,

> A 26-year-old obese woman with a history of bipolar disorder complains that her recent struggles with her weight and eating have caused her to feel depressed. She states that she has recently had difficulty sleeping and feels excessively anxious and agitated. She also states that she has had thoughts of suicide. She often finds herself fidgety and unable to sit still for extended periods of time. Her family tells her that she is increasingly irritable. Her current medications include lithium carbonate and zolpidem.

All 30 topics in the track were manually created by medical experts, who took care to ensure that the most pertinent information is present in both versions of the case reports [9]. Participants were only allowed to use one of the two representations for a submitted run[2].

The CDS track is especially interesting because of the complexity of the search problem. Not only do the issues of ambiguity and scope inherent in natural language severely complicate the identification and extraction of relevant information from the case reports, but the definition of 'relevancy' in this task differs from the traditional notion of 'aboutness': A document is judged relevant only if it is both topically relevant *and* helps answer a clinical question for the given case, such as "What is the (correct) medical diagnosis for this case?" or "Which treatment should be prescribed to this patient?". This additional dimension of relevance is called Clinical Question Type (CQT). In the CDS track 3 different CQTs are defined: *Diagnosis* ("What is the (correct) medical diagnosis for this case?"), *Test* ("Which is the most efficient (diagnostic) test to be used in this case?"), *Treatment* ("Which treatment should be prescribed to this patient?"). The question of how to (best) represent the CQTs and the interaction between topicality and CQT have been topics of great interest since last year's CDS but are still open areas of research.

The set-up of this year's instalment was very similar to that of the 2014 CDS track: A total of 30 manually created topics were released by the organizers, which comes to 10 topics per CQT, and retrieval experiments were carried out on the 21/1/2014 snapshot of PubMed's Open Access subset[3]. A novel addition to the track was the introduction of a second subtask: After a the initial retrieval based solely on the text in the case reports (subtask A) was finished, information on the correct diagnoses of topics 11 to 30 were released which could be used in second retrieval task (subtask B).

In our contribution to this year's track we focused on three points of interest: (a) Similar to last year, we explored the relative usefulness of using MeSH representations compared to word-based representations; (b) We experimented with multiple models of the Clinical Dimension Type in the retrieval process; and (c) We investigated to what extent the information in the case reports is complementary when the correct diagnosis is known.

The paper is organized as follows: In Section 2, we describe the individual components of the different systems that were built in the course of this track. In Section 3 we describe the six submitted runs . Results are presented in Section 4.

---

[2]It has been found, however, that combining both representation leads to improved retrieval results since the most pertinent information is indirectly given a higher weight [7]

[3]Can be found at `http://trec-cds.appspot.com/2015.html`

2 System Components

2.1 Extending MeSH coverage in the corpus

During our participation last year we found that using MeSH concepts to represent the content of a document leads to the most precise retrieval results but we were severely hindered by the low coverage[4] of MeSH terms in the Open Access set[2]. This year we opted to resolve this problem by extending the MeSH coverage of the articles ourselves: We used the MTI indexer[5, 4], available at the NIH[5], to assign MeSH Main Heading and Subheading labels to the unlabelled documents in the OA set (300K documents). We obtained MeSH label suggestions for 201,213 documents, which boosted coverage to 87% of the corpus. We have, however, no information on the quality of the assigned labels[6]. Analysis of the documents that were not assigned any MeSH labels by the MTI indexer shows that these are mostly short documents, such as 'letters to the editor' from PubMed Central which are unlikely to be of importance in medical search.

Since the high-precision nature of concepts has proven problematic for search [8] we decided to increase recall by mimicking the MeSH term explosion [1] that is implemented in PubMed. During indexation phase we automatically extended the set of MeSH terms associated with a document to include all ancestors of the initial MeSH terms. We kept track of the hierarchical arrangement of the terms by assigning logarithmically decreasing weights to concepts that appear higher up in the hierarchy. We combined several weighting schemes in the index: For example, MeSH terms (and their ancestral trail) of major topics were indexed with the double the weight of non-major headings. To limit the number of MeSH features, the antecedent trail of Main Heading and Subheading combinations were indexed separately. Figure 1 illustrates the term explosion procedure for a document with three assigned MeSH terms.

This indexing strategy favours the more specific terms but allows for a back-off to more generic terms if not direct match can be found between query and document. For example, a document that has been assigned a highly specific MeSH term, e.g. "Umbilical Arteries" (D004373), can still be found by a more general query, e.g. " Umbilical Cord" (D014470), because parts of their ancestral chains of MeSH terms coincide. The more generic MeSH terms have a small impact on the final retrieval score due to their high document frequencies and the low weights with which they are indexed. The index of MeSH terms was used in runs 2 and 5.

2.2 Disease hypotheses generation

In subtask B (runs 4–6) participants were provided with the correct diagnoses for topics 11 up to 30 (CQT *T*est and *Treatment*) but for the first ten topics (of the *Diagnosis* CQT) this information was not available. In order to be able to run the same system pipelines over all 30 topics we therefore generated disease hypotheses for the first ten topics based on the text in the case reports. We opted to use the same strategy as last year[2], but with different resources. Last year's disease hypotheses retrieval system covered a small set of very common diseases (taken from the Disease-Symptom Knowledge Database[7]) and a large set of extremely rare (genetic) disease[10]. This year we opted for a more balanced coverage and we therefore selected Wikipedia and MedlinePlus as sources for disease hypothesis generation. We extracted a corpus of 2496 articles from Wikipedia, and another

[4] Only 59% of articles in the set had been assigned MeSH terms.
[5] `ii.nlm.nih.gov/Interactive/mti.shtml`
[6] We did configure the MTI to apply strict filtering in order to obtain high-precision labelling.
[7] Freely accessible at `http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html`

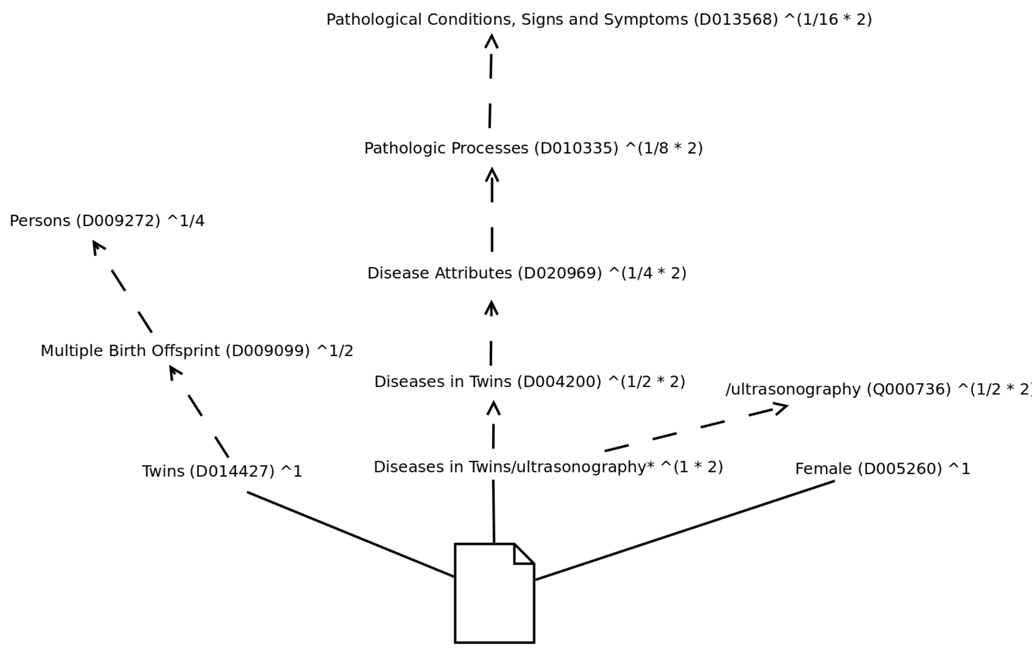

Figure 1: Example of MeSH explosion for a document with three initial MeSH terms (indicated by full lines). The dashed lines indicate automatically retrieved ancestor MeSH terms. Weights are indicated by the ˆsymbol. The figure shows the weight calculations, e.g., $\hat{}(1/8 * 2)$ indicates that the term was indexed with a weight of .25.

corpus of 1796 web pages from MedlinePlus[8] of common to moderately rare diseases. Each document describes exactly one disease or condition, and the two corpora overlapped, i.e. all diseases in the MedlinePlus corpus has a corresponding page in the Wikipedia corpus. These two corpora were chosen with an eye on their coverage and language use: The webpages in MedlinePlus are written in a very accessible style and contain a lot of descriptions of signs and symptoms. Wikipedia articles are more varied in style and are more likely to contain medical terms. The corpus was indexed using Solr[9]. To generate a disease hypothesis for a given topic, we used the text that topics to query the index. The title of the web page with the highest retrieval score was extracted and processed with MetaMap to output the CUI of the most likely disease. Since the diagnoses for topics 1–10 are not known, we cannot evaluate the impact of this step on retrieval performance. However, an evaluation on the diagnoses of topics 11–30 show a moderate performance.

2.3 Reranking

2.3.1 Time-based reranking

The aim of a CDS search is to provide a user with relevant articles that enable him or her to make a well-informed decision for a given medical case. If the system returns outdated information,

[8] http://www.nlm.nih.gov/medlineplus/
[9] http://lucene.apache.org/solr/

this may negatively impact the decision making process. In the survey on last year's track [7] it was noted that the assessors who judged the systems output immediately rejected literature that they deemed too old. We therefore decided to incorporate a time-based reranking step in our runs in which the more recent documents were boosted and older documents were demoted. The implementation of this re-ranking step was fairly basic: For each document in the OA set, we extracted the publication data and took the difference of the year of publication with the number 2014 to indicate the rank of the documents. Because of this low granularity, many documents have the same rank. We then combined the time-based rankings with either the retrieval score rankings or both the retrieval score and CQT-based rankings (see below), using the weighted Borda Fuse algorithm[10]. Borda Fuse is a simple rank aggregation algorithm in which ranks are counted as the number of votes in an election. Time-based re-ranking was used in all submitted runs.

2.3.2 Reranking based on CQT

As mentioned in Section 1, CQT is a difficult concept to model in retrieval. One of our strategies (next to UMLS-based feature selection (runs 3 and 6) and MeSH-based query expansion (run5)) was an additional re-ranking step in which retrieved documents were re-ranked according to pre-trained CQT classification models. These were created as follows: Using a set of manually crafted (boolean) MeSH queries[11] (see Table 1), we selected four different subsets of documents from the MeSH-annotated part of the OA snapshot. These four subsets correspond to the three CQTs (*Diagnosis*,*Test* and *Treatment*) and a selection of negative training material (*Other*). The 'diagnosis' and 'treatment' MeSH queries were meta-terms (pre-made MeSH queries) that were manually constructed for the CiSMeF project[6]; The 'test' query was manually constructed by a medical specialist who was involved in the CiSMeF project. Table 1 also shows the sizes of the four subsets that were selected.

We can see that the 'test' query is too specific and retrieved fairly few documents. In order to build well-balanced models, a random subset of max. 12,210 documents was selected from all subsets. This led to a classification subcorpus of 31,735 documents in total.

The documents for all four categories were indexed as uni-, bi- and trigrams, so as to catch those terms and phrases that are descriptive for and frequently used in the related clinical question type domain. The full text of the document (i.e. Title, Abstract and Body of the articles, as well as information on publication-type) was used. We then used the Winnow algorithm in the Linguistic Classification System (LCS)[3] to construct classification models for each category in an one-vs-all training. This algorithm has been used successfully for patent and literature classification. One of its main advantage is that the resulting models are human-readable and easy to transform into features for other implementations.

Table 2 shows the highest-ranking terms, i.e. terms that have been found to be most distinctive for that particular category compared to the rest of the corpus, for the three different clinical question types.

In the reranking step, we scored the retrieved documents that were retrieved for a topic with the model of the CQT that was associated with that model. The documents were then reranked according to their classification score with the highest scores, i.e. most relevant features, at the highest ranks. We then combined the CQT ranking with the retrieval score ranking using the weighted Borda Fuse algorithm. CQT reranking was used in runs 1, 2 and 4.

[10]We implemented our own version based on the `borda 0.1` Python package.

[11]These queries were also used in our participation to the CDS track last year.

category	MeSH query	# of documents in the subset
Diagnosis	"diagnosis"[MeSH Terms] OR "diagnosis, oral"[MeSH Terms] OR "diagnostic equipment"[MeSH Terms] OR "diagnostic services" [MeSH Terms] OR "nursing diagnosis"[MeSH Terms] OR "reagent kits, diagnostic"[MeSH Terms] OR "diagnosis"[Subheading] OR "diagnostic use"[Subheading]	11,201
Test	"psychiatric somatic therapies"[MeSH Terms] OR "psychotherapy" [MeSH Terms] OR "root canal therapy"[MeSH Terms] OR "therapeutics"[MeSH Terms] OR "treatment outcome"[MeSH Terms] OR "therapeutic use"[Subheading] OR "therapy"[Subheading]	1,790
Treatment	"diagnostic techniques and procedures" [MeSH Terms] OR "psychological tests" [MeSH Terms] OR " toxicity tests" [MeSH Terms] OR "dental caries activity test" [MeSH Terms] OR "dental pulp test" [MeSH Terms] OR "genetic complementation test" [MeSH Terms] OR "maternal serum screening tests" [MeSH Terms] OR "mutagenicity tests" [MeSH Terms] OR "radioimmunosorbent test" [MeSH Terms] OR "mandatory testing" [MeSH Terms]	12,210
Other	-	10,000

Table 1: MeSH queries to select 4 corpus subsets and their document distribution

rank	Diagnosis	Test	Treatment
1	research-article	complementation	immunotherapy
2	diagnosis	mutagenicity	therapy
3	clinical	complemented	patients with
4	background	genotoxicity	therapeutic
5	diagnostic	8217 s	research-article
6	case report	toxicity	treated
7	cancer screening	mutagenic	treatment of
8	figure	toxicological	chemotherapy
9	test results	processed	treated patients
10	tomography	toxicology	embolization

Table 2: Top 10 highest-ranking features in models for diagnosis, test and treatment categories

2.4 Retrieval

All reported retrieval experiments were carried out using Solr. We applied key-word stemming and stopword removal, and indexed the following fields:

- `title`: Text extracted from `article-title` tag
- `abstract`: Text extracted from `abstract` tag
- `text`: Text extracted from `body` tag after removal of tables
- `cat`: Text extracted from `article-type` tag
- `id`: String extracted from `pmid` tag
- `pmc`: Text extracted from `pmc` tag (if available)
- `MeSHterms`: MeSH concept identifiers extracted from PubMed using the `pmc` identifier, or assigned to the document by the MTI indexer (see Section 2.1)

Pseudo-Relevance Feedback was carried out using Solr's "More Like This" (MLT) algorithm.

3 Runs

3.1 LIMSIrun1BoW

Our first run was a baseline bag-of-words run in which all words from the topic summary.

1. **Initial retrieval step:** Topic summary (words) is used to query the index (Title, Abstract and Text fields).
2. **Pseudo-Relevance Feedback:** Terms from Title and Abstract fields of the top 40 retrieved documents are added to the original query for a second retrieval step.
3. **Reranking:** Retrieved documents are re-ranked for time and CQT; Rankings are combined in weighted Borda Fuse (.6 weight for retrieval ranking; .2 weight for temporal ranking; .2 weight for CQT ranking)

3.2 LIMSIrun2MSH

This run combines words and MeSH terms in the initial query.

1. **Initial retrieval step:** Topic description is processed with the MTI indexer. Resulting MeSH suggestions are added to the topic description to query the index (Title, Abstract, Text and MeSHterms fields).
2. **Pseudo-Relevance Feedback:** Terms from the Title and Abstract fields of the top 40 retrieved documents are added to the original query for a second retrieval step.
3. **Reranking:** Retrieved documents are re-ranked for time and CQT; Rankings are combined in weighted Borda Fuse (.6 weight for retrieval ranking; .2 weight for temporal ranking; .2 weight for CQT ranking)

3.3 LIMSIrun3SmF

In this run we replaced the CQT reranking with UMLS-based filtering of terms based on Semantic Types.

1. **Initial retrieval step:** Topic summary (words) is used to query the index (Title, Abstract and Text fields).
2. **Pseudo-Relevance Feedback:** Terms from the Title and Abstract fields of the top 40 retrieved documents are added to the original query for a second retrieval step.
3. **Term filtering:** Text from the Title and Abstract fields of the top 40 retrieved documents is used for a second PRF step. The newly selected additional terms are processed with MetaMap with restrictions on accepted Semantic Types[12]. MetaMap's output is then added to the query (topic summary + terms extracted from first PRF).
4. **Reranking:** Retrieved documents are re-ranked for time; Rankings are combined in weighted Borda Fuse (.7 weight for retrieval ranking; .3 weight for temporal ranking)

3.4 LIMSIrun4Syn

A baseline bag-of-words run using all words from the topic summary as well as the UMLS synonyms for the given diagnosis[13].

1. **Representation of diagnosis:** The diagnosis is processed by MetaMap to find a CUI; We then extracted all term variants from the UMLS for that CUI.
2. **Initial retrieval step:** Topic summary (words) and diagnosis synonyms are used to query the index (Title, Abstract and Text fields for topic; Title, Abstract fields for diagnosis).
3. **Pseudo-Relevance Feedback:** Terms from Title and Abstract fields of top 40 retrieved documents are added to original query for a second retrieval step.
4. **Reranking:** Retrieved documents are re-ranked for time and CQT; Rankings are combined in weighted Borda Fuse (.6 weight for retrieval ranking; .2 weight for temporal ranking; .2 weight for CQT ranking)

3.5 LIMSIrun5MPF

Run 5 is mostly MeSH-based with an added PRF step to improve recall by adding highly relevant words.

1. **Representation of diagnosis:** Using the MTI indexer the diagnosis is transformed into MeSH terms.

[12]For *Diagnosis* only the terms of the following Semantic Types were accepted: `irda, lbtr, fndg, sosy, diap, inpo, patf, dsyn, mobd, neop, comd, emod`; For *Test* only terms from the medd, resd, irda, lbtr, lbpr, diap categories. For *Treatment*, the `medd, drdd, resd, clnd, phsu, antb, bodm` categories.

[13]For topic 1-10 we generated our own diagnosis hypothesis, see Section 2.2

2. **Initial retrieval step:** The diagnosis MeSH terms are combined with the CQT MeSH query associated with that topic (see Table 1). Extra weight is given to MeSH terms in the query that consist of the diagnosis Main Heading and a relevant subheading.

3. **Pseudo-Relevance Feedback:** Terms from Title and Abstract fields of top 40 retrieved documents are added to original query for a second retrieval step.

4. **Reranking:** Retrieved documents are re-ranked for time; Rankings are combined in weighted Borda Fuse (.7 weight for retrieval ranking; .3 weight for temporal ranking)

3.6 LIMSIrun6Wik

Run 6 is an experimental run in which we tried to extract disease-specific terms that capture the relevant CQT from the Wikipedia page of the diagnosis.

1. **Representation of diagnosis:** The diagnosis is processed by MetaMap to find a CUI; We then extracted all term variants from the UMLS for that CUI.

2. **Disease-specific CQT-related terms:** We extracted a section of the Wikipedia page of the diagnosis, according to the CQT (The 'Signs and Symptoms' section for *Diagnosis*; The 'Diagnosis' section for *Test*; and the 'Treatment' or 'Managing' sections for *Treatment*)[14]. These sections were then processed using MetaMap, with severe restrictions on the semantic types of the concepts that should be found per Clinical Question Type. For each recognized concept, we looked up the term variants in the UMLS.

3. **Initial retrieval step:** The diagnosis synonyms and extracted CQT-related terms are combined to form an initial query. Additional weight is given to terms recognized in the Title and/or Abstract fields. Constructed query is then used to query the index (Title, Abstract and Text fields).

4. **Pseudo-Relevance Feedback:** There is no additional PRF step

5. **Reranking:** Retrieved documents are re-ranked for time; Rankings are combined in weighted Borda Fuse (.7 weight for retrieval ranking; .3 weight for temporal ranking)

Table 3 gives an overview of the strategies used in the different runs.

[14]For two diagnoses, we did not find a relevant paragraph in the Wikipedia article. As a back-off we used the introductory paragraph.

runName	query and document representation	strategy for CQT	representation of diagnosis
LIMSIrun1BoW	words	reranking	-
LIMSIrun2MSH	MeSH + words	reranking	-
LIMSIrun3SmF	words	filtering on Semantic Types (MetaMap)	-
LIMSIrun4Syn	words	reranking	term variants (UMLS)
LIMSIrun5MPF	MeSH + words	MeSH query	MeSH
LIMSIrun6Wik	words	filtering on Semantic Types (MetaMap)	term variants (UMLS)

Table 3: Comparison of strategies for different runs

4 Results

Table 4 summarizes the results of our officially submitted runs, as well as one additional run (run5bisMPFMSH) which was performed after the competition ended. We observe (similar to our findings last year) that the best runs for both subtask A and B are those which utilize MeSH terms to capture relevant information (run2 and run5, respectively).

runName	infAP	infNDG	R-prec	P@10
LIMSIrun1BoW	0.0362	0.1642	0.1450	0.2667
LIMSIrun2MSH	**0.0602**	**0.2507**	**0.1881**	**0.3767**
LIMSIrun3SmF	0.0197	0.1199	0.0694	0.2067
LIMSIrun4Syn	0.0827	0.3303	0.1993	**0.5000**
LIMSIrun5MPF	0.0982	**0.3507**	0.2284	0.4767
run5bisMPFMSH	***0.1014***	*0.3437*	***0.2321***	*0.4600*
LIMSIrun6Wik	0.0098	0.0783	0.0371	0.1300

Table 4: Retrieval scores for submitted runs and additional run

The difference between run1 and run4 shows that, unsurprisingly, adding the (correct) diagnosis information increases retrieval scores substantially. This raises the question to what extent the information in the case report is useful if a diagnosis is known. To examine this, we performed an additional run (run5bisMPFMSH) which combines the strategies from run2 and run5. In this run the MTI indexer is used to provide MeSH label suggestions for both the text in the case description, as well as for the diagnosis that is provided for a topic. In the initial retrieval step, the case report MeSH terms and diagnosis MeSH terms are combined with the CQT MeSH query according to the protocol described above for run5. We find that adding information from the case report (in the form of MeSH terms) improves recall but lowers precision at the top of the rankings. Likely the MeSH terms extracted from the case reports lead to query drift.

References

[1] Darlene Chapman. Advanced search features of pubmed. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 18(1):58, 2009.

[2] Eva D'hondt, Brigitte Grau, Stéfan Darmoni, Aurélie Névéol, Matthieu Schuers, and Pierre Zweigenbaum. Limsi@ 2014 clinical decision support track. In *Text Retrieval Conference*.

[3] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2, April 1988.

[4] James G Mork, Dina Demner-Fushman, S Schmidt, and Alan R Aronson. Recent enhancements to the nlm medical text indexer. In *Working Notes for CLEF 2014 Conference, Sheffield, UK*, pages 1328–1336, 2014.

[5] James G Mork, Antonio J. Jimeno Yepes, and Alan R Aronson. The nlm medical text indexer system for indexing biomedical literature.

[6] Aurélie Névéol, Lina F Soualmia, Magaly Douyère, Alexandrina Rogozan, Benoît Thirion, and Stéfan J Darmoni. Using CISMeF MeSH "Encapsulated" terminology and a categorization algorithm for health resources. *International journal of medical informatics*, 73(1):57–64, 2004.

[7] Kirk Roberts, Matthew Simpson, Dina Demner-Fushman, Ellen Voorhees, and William Hersh. State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the trec 2014 cds track. *Information Retrieval Journal*, pages 1–36, 2015.

[8] Wei Shen and Jian-Yun Nie. Is concept mapping useful for biomedical information retrieval? In Josiane Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, GarethJ.F. Jones, Eric SanJuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 9283 of *Lecture Notes in Computer Science*, pages 281–286. Springer International Publishing, 2015.

[9] Matthew S Simpson, E Voorhees, and William Hersh. Overview of the trec 2014 clinical decision support track. In *Proceedings of the 23rd Text Retrieval Conference (TREC 2014)*. National Institute of Standards and Technology (NIST), 2014.

[10] Steffanie S Weinreich, R Mangon, JJ Sikkens, ME Teeuw, and MC Cornel. Orphanet: a European database for rare diseases. *Nederlands tijdschrift voor geneeskunde*, 152(9):518–519, 2008.