Published in proceedings of the Twenthieth text retrieval conference, 2011, which should be cited to refer to this work

BiTeM group report for TREC Medical Records Track 2011

J. Gobeill [a], A. Gaudinat [a], E. Pasche[b], D. Teodoro[b], D. Vishnyakova[b], P. Ruch[a]

[a] *BiTeM group, University of Applied Sciences, Information Studies, Geneva*
[b] *BiTeM group, University and Hospitals of Geneva, Geneva*

contact: {julien.gobeill;patrick.ruch}@hesge.ch

Abstract

The BiTeM group participated in the first TREC Medical Records Track in 2011 relying on a strong background in medical records processing and medical terminologies. For this campaign, we submitted a baseline run, computed with a simple free-text index in the Terrier platform, which achieved fair results (0.468 for P10). We also performed automatic text categorization on medical records and built additional inter-lingua representations in MeSH and SNOMED-CT concepts. Combined with the text index, these terminological representations led to a slight improvement of the top precision (+5 % for Mean Reciprocal Rank). But the most interesting is analysing the contribution of each representation in the coverage of the correct answer. The text representation and the additional terminological representations bring different, and finally complementary, views of the problem: if 40% of the official relevant visits were retrieved by our text index, an additional 15% part was retrieved only with the terminological representations, leading to 55% (more than half) of the relevant visits retrieved by all representations. Finally, an innovative re-ranking strategy was designed capitalizing on MeSH disorders concepts mapped on queries and their UMLS-equivalent ICD9 codes: visits that shared this ICD9 discharge code were boosted. This strategy led to another 10% improvement for top precision. Unfortunately, any deeper conclusion based on the official results is impossible to draw due the massive use of Lucene and the evaluation methods (pool): in our baseline text run, only 52% of our top 50 retrieved documents were judged, against 77% for another participant's baseline text run who used Lucene. Official metrics focused on precision are thus difficult to interpret.

Introduction

The goal of the new TREC Medical Records Track is to foster research on Information Retrieval in free-text fields of electronic medical records. Such an effort to provide a large, public and reliable benchmark is beneficial, because there was a particular lack in this field. These clinical records are very sensitive data, that needs multiple agreements and de-identifying processes before to be provided by institutions for research purposes. This is probably why the literature is not rich in this field. Nevertheless, the BiTeM group [1] already worked with medical records obtained from the Hospitals and University of Geneva, in the framework of researches in medical coding [2] or ad hoc retrieval [3].

The first TREC Medical Records Track focused on an ad hoc search task of finding a population (i.e. medical records) over which comparative effectiveness studies can be done [4]. Generally, the topic consisted in a dozen of words, and specified a particular disease and a particular treatment or intervention. Participant's system had to return a list of visits ranked by decreasing relevance, among a collection of more than 100'000 records. These runs were then evaluated by a pool of physicians that judged the relevance of submitted documents.

The task provided no training data, and each group was allowed to submit only up to four runs. For producing its runs, the BiTeM group then relied on its skills and intuitions, and on strategies largely investigated in past medical Information Retrieval tasks and projects [5,6,7,8]. Such skills and strategies were :

– Choice of the best document representation and pre-processing steps for the free text fields indexing and retrieval.

– Automatic document and query annotation with medical controlled vocabularies (MeSH and SNOMED CT). Use of these annotations for building complementary inter-lingua indexes.

– Use of UMLS correspondence between annotated MeSH descriptors and ICD discharge codes for automatic re-ranking purposes.

Data

The collection was a set of de-identified medical records made available for research purposes through the University of Pittsburgh [4]. The query set was developed by physicians and contained 34 topics.

1) Collection

The collection contained 101'711 documents (i.e. medical records). Most records were associated with a "visit" identifier, visit being seen as an episode of care. Organizers chose to use the visit as the response unit. Therefore, 5'283 documents had no visit id and were simply discarded, and there was a total of 17'267 answerable visit ids. The number of reports per visit varied between 1 and 415, with a median of 3.

Each record contained a set of different fields :

* "checksum", that was the unique report id.
* "type", a local and very general descriptor that seemed to be chosen among 9 values. For example, *RAD* (for radiology) was the type of 46% of the records contained in the collection.
* "subtype", a more precise description, that had 317 different values in the collection. Tab 1 shows the 20 most frequent subtypes in the collection.

subtype	#
CHEST	19773
CT	10312
vide	8883
EVAL	6918
CONSULT	6285
PALLIATIVE CONSULT	3333
ER	2926
CCM ATTEND	2740
DISCHARGE	2680
US	2582
IP-PROGRESS	2052
PALLIATIVE PROGRESS	1776
ABD	1749
MR	1696
GIM ATTEND	1582
ADMISSION	1548
OPERATION	1241
SPINE	1183
HISTORY	933
NM	880

Table 1. The 20 most frequent subtypes in the collection.

* "chief_complaint", a short and free description of the admission's primary reason. This field had 7676 different values in the collection, and its format seemed to be very free, as *CONGESTIVE HEART FAILURE* was as frequent as its abbreviation *CHF* (1016 vs 1026 records).
* "admit_diagnosis" that contained one ICD9 CM code, and "discharge_diagnosis" that contained an average of 10 codes per record.
* "report_text" that was the core of the medical record. It contained free text, with an average of 406 words per record.

The correspondence table between report ids and visit ids was given by the organizers in a separate file.

2) Query sets

Topics were developed by physicians, which had been instructed that they had to exploit information from the free text fields. Four sample topics were provided, in order to illustrate the syntactic format of the test topics rather than really constitute training data. 35 topics were additionally given as test set. Participants had to freeze their system once they discovered the topics. Topics were numbered from 101 to 135. Topic 130 was finally discarded.

Strategies

For this first TREC Medical Records Track, there was no really training data, and participants were allowed to post only four runs. Therefore, we were not allowed to exhaustively explore all strategies we could, and rather used strategies that showed their efficiency in past works. The first strategy was obviously to build a basic

index with free-text fields, as we considered it should remain the core of the system. Then, we built additional indexes in inter-lingua, MeSH and SNOMED-CT. MeSH and SNOMED-CT terms describing each record were obtained with automatic annotation processes. Finally, we applied a boosting strategy based on diseases mapped in the topics and documents' discharge codes.

1) Document representation for the basic index

The first step was to choose the indexing unit. We foremost considered creating a virtual file for each of the 17'267 visits, by concatenating all fields contained in the related medical records. But we finally chose to treat and index all the 96'428 medical records independently. When several medical records belonging to the same visit were retrieved for the same topic, we simply kept only the first one, with its original score.

The second step was to choose what fields to keep in order to build the index. We discarded the type, as we considered that the nine different values contained not enough information. We also decided to discard the subtype, as we found too much abbreviations and acronyms in this field (as showed in Tab. 1). We considered that subtypes would bring more noise than signal. We finally chose to keep the chief complaint, and the report text.

Several pre-processing steps were performed before the indexing :

- We chose to apply Porter stemming [10], and to use a standard list of general stop words.
- We added 22 specific stop words by manually screening the 500 most frequent words. Words that were considered not to have discriminative power were discarded, such as *patient*, *room*, *diagnosis* or *medical*. For instance, the word *medical* was present in 31'000 records.
- We built a specific acronyms thesaurus by manually screening the most frequent chief-complaint (Tab. 2).
- We used more representative document frequencies computed with a sample of 8M of MEDLINE abstracts.

MI	Myocardial Infarction
CHF	Congestive Heart Failure
ABD	Abdomen
SOB	Shortness Of Breath
CVA	Cerebral Vascular Accident
COPD	Chronic Obstructive Pulmonary Disease
FIB	fibrillation
MVA	Motor Vehicle Accident
UTI	Urinary Tract Infection
CAF	Congestive Heart Failure
N V	Nausea and Vomiting
TIA	Transient Ischemic Attack
CAD	Coronary Artery Disease
CP	Chest Pain
MVC	Motor Vehicle Accident
TX	Transplant
SCC	Squamous Cell Carcinoma
IPF	Idiopathic Pulmonary Fibrosis
LV	Liver
C H F	Congestive Heart Failure
ALOC	Altered Level Of Consciousness
FX	Fracture
DKA	Diabetic ketoacidosis

Table 2: the manually designed thesaurus used.

Obviously, the same pre-processing was applied to topics.

2) Information Retrieval Platform

Indexing and Retrieval were computed with the Terrier platform [9], which is designed for large collections and which we already used in past similar competitions. We chose settings which proved to be efficient in past works: Okapi-BM25 as weighting scheme with default parameters. A short tuning was performed with the sample set.

A first run was then computed, with only free-text: BiTeMbase.

3) Documents medical annotation and inter-lingua indexes with MeSH and SNOMED-CT

The BiTeM group has a strong experience in the use of medical controlled vocabularies and ontologies for annotation and indexing. In a past similar IR competition [11], the goal was to find medical images based on their description; we found that injecting MeSH descriptors in the model led to great improvements [6]. Thus, we decided to automatically annotate all the medical records with a medical controlled vocabulary, and then to use these annotations in order to build complementary and inter-lingua indexes.

We chose two different controlled vocabularies, both being a part of UMLS: the Medical Subject Headings (MeSH) and the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT). Working with UMLS Semantic Types, we actually restricted both vocabularies to four general semantic groups [12]

that we found relevant for annotation: Disorders, Anatomy, Procedures and Chemicals & Drugs. We chose to annotate medical records by performing naïve word-matching [13], as this technology needs no learning data, is not time-consuming, shows a good precision, and do not need a threshold [14].

When we processed the data, the MeSH contained 26'142 concepts; restricted to our four semantic groups, it represented 17'481 concepts. The same values for SNOMED-WT were 291'205 and 226'137 (more than ten times bigger). For both vocabularies, at least one concept was mapped for more than 99% of the medical records. Tab.3 shows the average number of concepts matched in medical records, for each vocabulary and for each semantic group. Excepted from Chemicals & Drugs, more concepts were annotated using SNOMED-CT (36 concepts per medical record) than MeSH (24.4). 5'400 different MeSH concepts were mapped in the collection. The same value for SNOMED-CT is 16'276.

Concept/record	MeSH	SNOMED-CT
Anatomy	7.9	9.2
Chem & Drugs	3.8	3.7
Disorders	9.5	17.4
Procedures	3.2	5.7
TOTAL	**24.4**	**36**

Table 3: average number of concepts matched in medical records, for MeSH and SNOMED-CT and for each semantic group.

Using these annotations, we chose to build complementary inter-lingua indexes. For both MeSH and SNOMED-CT, we indexed medical records using the annotated concepts. We chose to use the unique identifiers rather than the preferred form: it means that when a record was annotated with the MeSH concept *D009765 Obesity*, we chose to keep *D009765* rather than *Obesity* : concepts were normalized by their identifier.

We thus obtained two complementary indexes: one with MeSH ids, the other with SNOMED-CT ids. Then, we applied the same annotation to the topics, and were able to retrieve documents based on annotated concepts. Tab.4 shows the average number of concepts matched in topics, for each vocabulary and for each semantic group.

Concept/topics	MeSH	SNOMED-CT
Anatomy	0.2	0.3
Chem & Drugs	0.1	0.1
Disorders	1.2	1.3
Procedures	0.5	0.6
TOTAL	**2.0**	**2.3**

Table 4: average number of concepts matched in topics, for MeSH and SNOMED-CT and for each semantic group.

For instance, for the topic 112 "Female patients with breast cancer with mastectomies during admission", these MeSH concepts were mapped: *D008408: mastectomy*, *D009369: neoplasms*, *D001940: breast* and *D001943: breast neoplasms*. D008408 was annotated in 363 records, equally when the record contained the word *mastectomy* than when it contains its synonyms *mammectomy*. Records that contained *mammectomy* could not being retrieved by the basic index.

3 topics (108, 118 and 120) received no annotations.

We thus obtained two supplementary runs: one with the MeSH annotations index, the other with the SNOMED annotations index. Then, we simply combine both runs with the basic run by summing scores for each retrieved visit.

Two other runs were then computed: BiTeMmEsh and BiTeMsnomed.

4) Boosting based on ICD discharge codes

The last strategy we applied was based on discharge codes. Once we were able to annotate topics with disorders belonging to MeSH, we were able for some of them, thanks to UMLS unique identifiers, to find a corresponding ICD9 code. 2'189 ICD9 codes are thus linked to a MeSH concept. For instance, in the third sample topic "Patients with atrial fibrillation treated with ablation", the MeSH concept *D001281: atrial fibrillation* was annotated. Thanks to UMLS, we could link this MeSH concept to the ICD9 code 427.31, which was present as discharge code in 15'822 medical records. We thus could benefit from the information contained in the medical coding.

Such post-processing strategies were already experimented in TREC-CHEM 2010 Track [8] with International Patent Classification codes. Once a first run is computed, the alternatives are either to filter documents that do not have the code, or to boost

documents that have the code. Studying the fourth sample topic was useful: "Elderly patients with ventilator-associated pneumonia". The MeSH concept *D053717: Ventilator associated pneumonia* was found, and was relied to ICD9 code 997.31. Yet, it appeared that no document was coded with 997.31, as probably medical coders do not use this code. Actually, it seems that medical coding may depends of the habits in an institution, sometimes with no lexical reasons. Filtering could discard all records when the corresponding ICD9 code was not used by coders. Hence we chose to boost, by doubling, the scores of medical records that contained an ICD9 code corresponding to a MeSH disorder annotated in the topic.

We then started from the BiTeMmEsh run and obtained a fourth and last run, BiTeMmhICD. Finally, out of 15'167 visits retrieved in the run2, 930 contained an ICD9 code annotated in the topic and were boosted by this strategy.

Results & Discussion

147 runs were officially submitted by all TREC-MED participants. 47 were judged. Two runs out of the four we submitted were judged: the baseline run and the last run (which was supposed to be the most achieved). Table 5 shows some metrics.

Run	Judged	MAP	MRR	Bpref			R-prec			P10		
				Best	Median	Ours	Best	Median	Ours	Best	Median	Ours
Baseline	yes	0.192	0.668	0.761	0.412	0.269	0.609	0.309	**0.244**	0.876	0.476	**0.468**
MeSH + ICD	yes	**0.200**	**0.703**			0.307			0.234			0.429
MeSH	no	0.196	0.642	0.758	0.434	**0.309**	0.598	0.305	0.238	0.859	0.444	0.441
SNOMED	no	0.189	0.678			0.308			0.242			0.421

Table 5: official results for the BiTeM runs. MAP stands for Mean Average Precision and MRR for Mean Reciprocal Rank. Bpref, R-prec and P10 were the official metrics. Best values are on bold.

1) Limits to the evaluation

It is very difficult to compare our strategies and to draw conclusions from this first TREC-Medical Records Track, because the evaluation was strongly biased for our runs.

Indeed, due to understandable technical limits, organizers only retained 47 (37 %) of the submitted runs in the pool. Moreover, for these runs, only the top 10 retrieved records were systematically evaluated, the rest being evaluated depending on pool sampling. Then, if a majority of participants used the same Information Retrieval engine, they had greater chances to share retrieved documents with other participants, and to have their retrieved documents judged.

We think that a majority of participants used Lucene. For instance, a participant [15] that used Lucene shared a matrix in order to visualize how many of his top 50 retrieved documents were evaluated in his baseline run, which was not judged. 77% of the top 50 retrieved documents were judged for his run – he claimed 77% is not enough. The same value for our baseline run, which was judged, is 52%. It means that, compared to a Lucene baseline, 25% of our top 50 retrieved documents were considered as false positives for the computation of the official metrics.

Hence, it is difficult to compare our runs to others for bpref and R-prec values. The only useful metric is P10 for judged runs, as all documents were judged in order to compute this value. Unfortunately, P10 is biased for our SNOMED run too, as 16% of the top 10 retrieved documents were not judged.

Nevertheless, we were able to judge our baseline run in terms of Mean Reciprocal Rank and P10. Then, for our terminological representations run (MeSH and SNOMED), we were able to use the gold file in order to tune the combination, and to analyze the coverage of the different lingua for this evaluation. Finally, we were able to conclude that our re-ranking strategy based on ICD9 codes improved top precision.

2) Baseline run

Our baseline run achieved good performances for MRR: 0.668. It achieved 0.468 for P10, which is slightly inferior to the median. Absolute deviation for P10 is 0.3, which means that results were various across queries.

3) Terminological representations

From the terminological representation (MeSH and SNOMED), we built two additional indexes and computed two additional rankings. These rankings were then combined with the text run in order to obtain the two distinct terminological runs. Because we didn't have training data, we were not able to tune the combination. For the competition, we simply decided to sum scores because we much trusted in terminologies. Now that we have the gold file provided by organizers, it seems that the default values we chose for both were not optimal, as terminological representations performed slightly less well than text. More performing combinations were obtained by reducing the contribution of the inter-lingua runs, as shown in Tables 6 (for MeSH) and 7 (for SNOMED).

Run	MRR	MAP
Baseline*	0.685	0.186
MeSH	0.611	0.133
Base + 1 MeSH*	0.675	0.19
Base + 0.3 MeSH	0.687	**0.201**
Base + 0.8 MeSH	**0.71**	0.196

Table 6: results for different tunings of the linear combination between the text index and the MeSH index. * marks official runs.

Run	MRR	MAP
Baseline*	0.685	0.186
SNOMED	0.579	0.133
Base + 1 SNOMED *	0.69	0.184
Base + 0.5 SNOMED	**0.712**	**0.198**

Table 7: results for different tunings of the linear combination between the text index and the SNOMED-CT index. * marks official runs.

For both terminologies, the combination we used for the competition brought no improvement, while efficient tuning now does. For MeSH: +4% for MRR with α=0.8, +8% for MAP with α=0.3. For SNOMED: +4% for MRR and + 7% for MAP with α=0.5.

More interesting is a deeper analysis of the coverage for the three different representations: text, MeSH and SNOMED-CT. Figure 1 shows how different and complementary are the three representations.

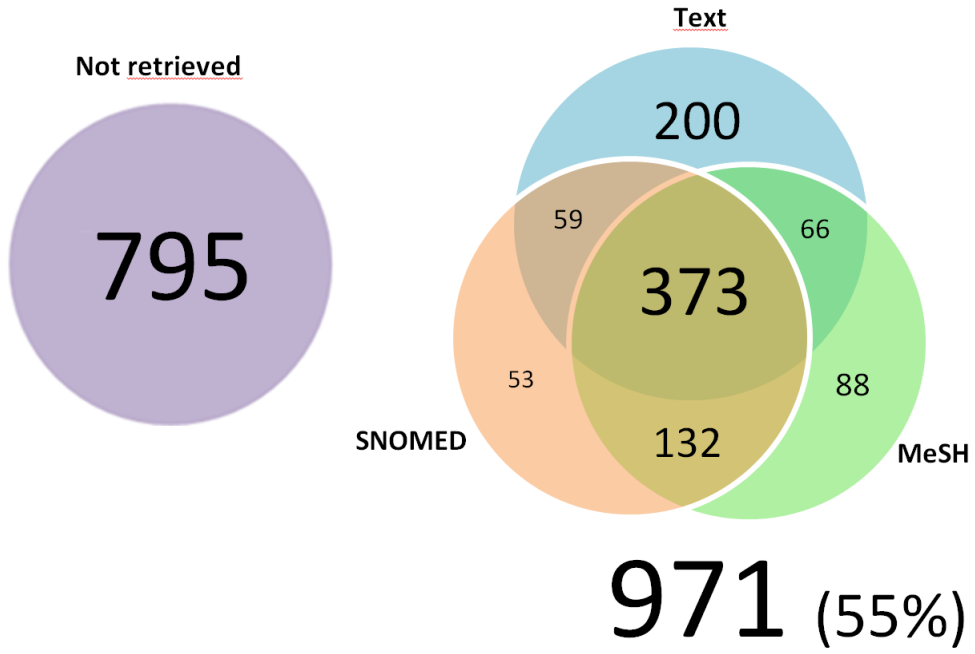

Figure 1: Coverage of the three different representations over the 1766 official relevant records.

28% of the relevant records (373+59+66) provided by the official gold file were both retrieved by the text and a terminological representation. 11% (200) were retrieved only by text. And an additional 15% part

(53+132+88) was retrieved using only the terminological representations MeSH or SNOMED. It shows that, even if the combination of the different representations only slightly improves the global system performances, terminological representations allow the system to retrieve visits that text similarity is not able to retrieve. This is probably the effect of both thesaurus – the terminologies provide a lot of synonyms – and normalization.

4) Boosting based on ICD9 discharge codes

Finally, the boosting based on discharge codes led to a 10% improvement for MRR from the official text + MeSH representation run. Further experiments with the gold file showed that doubling the score was the good tuning.

Deeper analysis reveals that the power of this strategy heavily varies depending on the query. For instance, for the query 102 "patients with complicated GERD who receive endoscopy", the mapped MeSH term *D005764 Reflux, Gastro-Esophageal* was linked to ICD9 code 530.81 thanks to UMLS; 64 out of the 89 official relevant visits (72%) contained 530.81 in their discharge codes. For the query 104, 7 out of the 8 official relevant visits (87%) contained the discharge code 185 *prostate cancer*. On the other hand, for the query 101, only 37% of the official relevant visits contain the ICD9 discharge for "hearing loss" 389.9. Worst, for the query 112 that deals with "breast cancer", none of the 66 official relevant visits contained the ICD9 code for breast cancer 174.

Conclusion

For this first TREC Medical Records Track, we tried to evaluate different representations based on free-text, but also medical terminological concepts. It's difficult to draw conclusions on the text representation because of the pool evaluation. However, the complementarity between text and terminological representations is established, and is particularly interesting for coverage. The terminological representation needs to be improved and refined, and we already have a few hints about how to proceed. We could use different medical terminologies and different mapping strategies for the different aspects we pointed out: Disorders, Anatomy, Procedures and Chemicals & Drugs. We also need a more efficient combination. At last, we need to better understand the discharge coding in order to take benefit from it.

Acknowledgments

The study reported in this paper has been partially supported by the European Commission Seventh Framework Program (DebugIT project grant no. FP7-ICT 217139).

References

[1] http://eagl.unige.ch/bitem/

[2] P Ruch, J Gobeill, I Tbahriti, P Tahintzi, P Lovis, A Geissbühler and F. Borst, "From clinical narratives to ICD-Codes: automatic text categorization for medico-economic encoding", SSIM 2007.

[3] P Ruch and R Baud, "Evaluating and Reducing the Effect of Data Corruption when Applying Bag of Words Approaches to Medical Records", Int J Med Inf, 67 (1-3):75-83, 2002.

[4] TREC Medical Records 2011 Guidelines

[5] Julien Gobeill, Imad Tbahriti, Frédéric Ehrler, Patrick Ruch. Vocabulary-Driven Passage Retrieval for Question-Answering in Genomics. In Proceedings of TREC'2007.

[6] J Gobeill, D Teodoro, E Pasche and P Ruch, ""Taking Benefit of Query and Document Expansion using MeSH descriptors in medical imageclef 2009". In Proceedings of CLEF 2009.

[7] D Teodoro, J Gobeill, E Pasche, D Vishnyakova, P Ruch and C Lovis. "Automatic Prior Art Searching and Patent Encoding at CLEF-IP '10" in CLEF 2010.

[8] J Gobeill, A Gaudinat, E Pasche, D Teodoro, D Vyshnyakova and P Ruch, "BiTeM site report for TREC Chemistry 2010: Impact of Citations Feedback for Patent Prior Art Search and Chemical Compounds Expansion for Ad Hoc Retrieval" in TREC 2010.

[9] I Ounis, C Lioma, C Macdonald and V Plachouras, "Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web", Novatica/UPGRADE Special Issue on Next Generation Web Search, vol 8, pp 49-56, 2007

[10] M Porter, "An algorithm for suffix stripping", Program, vol 14, pp 130-137, 1980

[11] H Müller, J Kalpathy–Cramer, I Eggel, S Bedrick, S Radhouani, B Bakke, C Kahn Jr. and W Hersh, "Overview of the CLEF 2009 medical image retrieval track", in CLEF 2009

[12] O Bodenreider and AT McCray, "Exploring semantic groups through visual approaches", Journal of Biomedical Informatics, 36(6):414-432, 2003.

[13] Y Yang and JO Pedersen, “A comparative study on feature selection in text categorization”, in Proc. 14th International Conference on Machine Learning, 412--420, 1997.

[14] J Gobeill, “Modèles de Question / Réponse pour la Biomédecine“, PHD Thesis, University of Geneva.

[15] Martijn Schuemie, TREC-MED Mailing List, 2011