SIB Text Mining at TREC 2018 Precision Medicine Track

Emilie Pasche[a,b], Paul van Rijen[a,b] , Julien Gobeill[a,b], Anaïs Mottaz[a,b,c], Luc Mottin[a,b], Patrick Ruch[a,b]

[a] HES-SO / HEG Geneva, Information Sciences, Geneva, Switzerland
[b] SIB Text Mining, Swiss Institute of Bioinformatics, Geneva, Switzerland
[c] Laboratory of Cognitive Neurorehabilitation, Faculty of Medicine, University of Geneva, Switzerland
contact: {emilie.pasche;paul.vanrijen}@hesge.ch

Abstract

The TREC 2018 Precision Medicine Track largely repeats the structure and evaluation of the 2017 track. The collection remains identical. Again, our team participated in the both tasks of the track: 1) retrieving scientific abstracts addressing relevant treatments for a given case and 2) retrieving clinical trials for which a patient is eligible. Regarding the retrieval of scientific abstracts, we queried all abstracts concerning one of the entities of the topic (i.e. the disease, the gene or the genetic variant) using various strategies (e.g. search in annotations of the collection, free text search using or not using synonyms, search in the MeSH terms, etc.). Then, for a given topic, the complete set of abstracts was based on the generation of different queries with decreasing levels of specificity. The idea was to start with a very specific query containing gene, disease and variant, from which less specific queries would be inferred. Abstracts were then re-ranked based on different strategies to favor abstracts that we considered more relevant to the given task. In 2017 we tested the use of drug densities to identify abstracts related to treatment. For this year we refined this strategy by giving more weight to drugs related to cancer treatment. Secondly, we used demographic information to favor abstracts concerning patients of the specified age-group and gender, and disfavoring abstracts targeting other age-group or gender patients. For the third strategy we utilized a word-level convolutional neural network to increase the rank of abstracts related to precision medicine. The fourth strategy consisted to expand the query to parent and children diseases. Finally, we tested an exact run which only retrieved abstracts respecting all information given in the topic. Results showed that all strategies but the last one resulted in some improvement of the retrieval power of the engine. As expected, our final run, focusing of precision, resulted in our best results regarding precision at rank 10, while other measures were negatively impacted. Regarding the retrieval of scientific abstracts, we boosted our last year's approach – which achieved competitive results – with supplementary strategies issued from other participants. Regarding the retrieval of clinical trials, we investigated filtering strategies for managing the condition (disease), and standard information retrieval for managing the gene and genetic variant. The results show that, despite the presence of a structured condition tag in the document, better performances are obtained when relaxing constraints: using synonyms and detecting the diseases in various fields, such as the summary.

Introduction

The SIB Text Mining group [1], at the Swiss Institute of Bioinformatics in Geneva, has a long history of participation in TREC campaigns, including TREC Genomics [2], TREC Medical Records [3], TREC Chemical IR [4], TREC Clinical Decision Support [5 ,6] tracks and the TREC 2017 Precision Medicine Track [7]. In parallel, the group is currently involved in several research projects, including the Swiss Variant Interpretation Platform for Oncology (SVIP-O), which aims at providing a centralized and curated database for clinical somatic variants providing from Swiss hospitals and related institutions.

The TREC 2018 Precision Medicine track focus on the identification of scientific articles and clinical trials, regarded as useful for clinicians treating cancer patients. The structure and evaluation for TREC Precision Medicine 2018 are largely repeated from the previous year. The collection remained identical. Similar to the 2017 track, the topics consisted of a disease, one or more mutated genes and some demographic information. Again, two tasks were proposed: 1) the retrieval of scientific abstracts and 2) the retrieval of clinical trials. Our team participated in both challenges. This year, training data from TREC 2017 Precision Medicine (i.e. 2017 topics and relevance judgments) were available to prepare the runs. The runs of TREC 2018 Precision Medicine were then evaluated by a pool of clinicians that judged the relevance of a subset of the submitted documents.

For producing the runs for the scientific abstracts task, similarly to 2017, we developed a core system, based on a set of queries, each differentially weighted. Assuming this strategy would enable us to retrieve a large subset of relevant abstracts, we then applied different strategies to work on the ranking of the retrieved abstracts. Successful strategies from 2017 were reused with some additional investigation, such as boosting cancer-related drugs for the re-ranking based on drug occurrences. Moreover, additional strategies were tested: i.e. use of demographic information, development of a precision medicine classifier and use of an exact match run.

For producing the runs for the clinical trials task, our strategies mainly focus on exploiting different fields of the documents – especially conditions and summary – for filtering or retrieving the query information. Diseases in both the collection and topics were normalized thanks to the NCI Thesaurus, then used for filtering unrelated trials. In the same way, demographic features were normalized then used for filtering. Then, a search engine was used for finding relevant trials according to the genes information. Compared to the 2017 campaign, we also performed boosting scores of retrieved trials according to phases, primary purposes, and study types. Boosting values were computed with last year gold file distribution.

1. Data

The Precision Medicine track provides two collections, one for each task: scientific abstracts and clinical trials. Both tasks share a common topics set.

1.1 Scientific abstracts

The scientific abstracts collection is composed of a snapshot of PubMed abstracts (January 2017) together with additional abstracts from AACR (American Association for Cancer Research) and ASCO (American Society of Clinical Ontology) proceedings. The XML version of the PubMed collection is used. It contains 26,670,000 abstracts, corresponding to 26,669,401 unique PMIDs. The latest version of a duplicated PMID is used. Title, abstract, publication date, publication types and MeSH terms are extracted for each abstract. AACR and ASCO abstracts are provided as TXT file. They contain respectively 33,018 and 37,007 abstracts. Only title, abstract and publication date are available for this subset.

1.2 Clinical Trials

The scientific abstracts collection is designed from a snapshot of ClinicalTrials.gov (April 2017). Approximately 240,000 clinical trials populate the collection. All trials are in XML format, and thus have – theoretically – a formal structure: information is stored in dedicated sections, such as the study phase, the sponsors, the design of the study, or the eligibility criteria. Some sections contain formatted fields (such as demographic conditions) while much contain free text.

1.3 Topics

The topics set consists of 50 semi-structured synthetic cases created by precision oncologists at the University of Texas MD Anderson Cancer Center. Each topic consists of the disease, genetic variants, and demographic information. While in topics of 2017, a “other” field was provided, mentioning other potential factors that may be relevant for the case, this field was retracted for 2018 topics.

Ontologies and resources

Several publicly available ontologies and resources have been used for developing our systems.

neXtProt [8] is a comprehensive human-centric discovery platform, developed by the Swiss Institute of Bioinformatics. With more than 20,000 proteins manually annotated, neXtProt provides high-quality synonyms for both protein and gene names. We used this resource for normalizing gene names.

The NCI Thesaurus (NCIt) [9] provided by the National Cancer Institute, is a reference terminology for biomedical coding, broadly used by both public and private care actors. This terminology covers clinical care, translational and basic research and public information and administrative activities. We used this resource for disease mapping, as it contains information for nearly 10,000 cancer and related diseases.

The Medical Subject Headings (MeSH) [10], provided by the US National Library of Medicine, is a controlled vocabulary used for indexing articles in MEDLINE. The MeSH is known for being less granular than specialized ontologies such as the NCIt, but also for being easily identified by Natural Language Processing, thanks to synonyms.

DrugBank [11] is a database containing biochemical and pharmacological information about drugs and drug targets. DrugBank includes more than 10,500 records. It also provides a high number of synonyms, as well as products names.

2. Strategies

In this section, we describe the strategies applied for each task.

2.1 Scientific abstracts retrieval

We participated in the scientific abstract task for the second time. We reused successful strategies tested last year and tried to improve them with some new ideas. Again, we have submitted five runs. The topics and relevance judgments of 2017 have been used as training data.

Again, we annotated diseases, genes and drugs within the whole collection, based on existing terminologies (i.e. NCIt for diseases, UniProtKB for genes and DrugBank for drugs). These annotations were then indexed together with the title, abstract, publication date, publication types and MeSH terms for each document. Solr Apache 7.3.1 is used for indexing and retrieval.

2.1.1 Baseline

We consider a topic to be constituted of three elements: a disease, one or several genes and one or several variants (e.g. an amino acid change). The gene or the variant can be missing (e.g. topic number 37 has no variant and topic number 20 has no gene). When several genes and/or variants are present, we treat each of them as a subtopic and merge the set of abstracts retrieved for each subtopic to define the final set of abstracts for the topic.

First, our system retrieves all the abstracts concerning one element of a topic (or subtopic). This search is based on a set of queries: the exact term is searched in the title of the abstract, in the core of the abstract, in the MeSH terms assigned to the abstract, in our annotations of the title of the abstract and in our annotations of the core of the abstract. Moreover, all queries are also performed using synonyms of the term. All the retrieved abstracts for a given element are then merged together, with different weights depending of the query providing the abstracts (e.g. the element was retrieved in the title using its main term; the element was retrieved in the abstract using a synonym). Tuning of the weights given to each query has been defined using the topics and relevance judgments of 2017.

Synonyms of genes are generated using UniProtKB terminology, while synonyms of diseases are retrieved using NCIt. Regarding variants, a synonym list has been manually created for copy number variants (e.g. amplification), while a SNVs synonym generator has been developed for single nucleotide variants. Given variant information (i.e. the gene name and the amino acid change), the SNVs generator produced the standard nomenclature format at the protein level as described by the Human Genome Variation Society (HGVS) [12]. When a corresponding dbSNP ID was found through neXtProt, the HGVS standard description was also generated for the transcript and genomic DNA description levels. Additionally, non-standard formats found in the literature [13] were generated for these different levels of description. It included at the protein level the use of single and three letters amino acid codes (e.g. Val600Glu) as well as hyphens and greater-than characters (e.g. V-600-E). At the DNA level, the use of hyphens along with greater-than characters was proposed (e.g. 1799T->A). When found, the dbSNP ID was also used as a synonym - although dbSNP is more likely to be impactful for germline variants than somatic variants.

Second, our system generates a set of different queries with decreasing levels of specificity. Indeed, our assumption is based on the fact that an abstract of interest may sometimes not mention the specified variant, but for instance another variant affecting the gene in a similar manner. Similarly, an abstract about the variant of interest for another cancer type may still be valuable from a clinical point of view. Additionally, such strategy might compensate the failure to collect abstracts about one specific element. Therefore, our approach is based on the generation of a set of four queries:

- Query 1: Disease + Gene + Variant
- Query 2: Disease + Gene
- Query 3: Gene + Variant
- Query 4: Disease + Variant

Thus, our system merges the abstracts common to the elements of the query. For instance, for the query number 2, the final set of abstracts is the union (abstractsSet_disease ∪ abstractsSet_variant) of the abstract retrieved for the disease and the abstract retrieved for the variant. Results for the four queries are then merged together through linear combination. Each set of results is differentially weighted.

This strategy aims at retrieving a maximum of relevant abstracts. We then apply additional strategies in order to re-rank the abstracts.

2.1.2 Drug density

Similarly to last year, our first run (*SIBTMlit1*) assumes that an abstract with a high frequency of drug names is probably more relevant to support our task, which consists to retrieve existing knowledge in the scientific literature regarding treatment of cancer. We thus use the pre-annotation of the abstracts with DrugBank to estimate the drug density of a publication (i.e. the number of occurrences of drug names in the abstract and title). This year, we tried to favor density of drugs related to cancer treatment. For this, a list of 384 DrugBank records has been defined based on different resources: cancer-related categories provided by DrugBank (e.g. *Antineoplastic agents*), the Cancer Drugs List provided by the National Cancer Institute [14], the List of Cancer Chemotherapy Drugs provided by the Navigating Care [15] and the Oral Chemotherapy Drugs List provided by CareFirst [16]. Results from the baseline run are re-ranked based on the number of occurrences of drug names per abstract, with a stronger weight for drugs from the cancer-related list. We also investigate attributing different weights whether the drug name is found in the title or in the core of the abstract.

2.1.3 Demographic density

While demographic information was not used last year, we investigate re-ranking abstracts based on gender and age-groups. MeSH terms are used to determine the demographic categories of the abstract. Run 2 (*SIBTMlit2*) is based on the re-ranking of the run 1 based on demographic information. For each abstract returned in our run 1 (*SIBTMlit1*), we attribute a score based on the sum of the age and gender scores as described in Table 1 depending if the abstract matches, does not match or does not discuss the topic's demographic information. We then defined the weight attributed to this re-ranking based on the tuning set.

Abstract	Age	Gender
Match	0.5	0.5
Does not match	-0.5	-0.5
Does not discuss	0.25	0.25

Table 1 Score attributed to an abstract

2.1.4 Precision medicine classifier

In results from 2017, we observed that a large set of incorrect results were abstracts that were not concerning precision medicine. Last year's strategy to determine if an abstract concerned precision medicine consisted of a manually-defined list of keywords considered as relevant or not relevant to distinguish between precision medicine and not precision medicine. This year, we ties leveraging the large training set available from last year to build a binary classifier. We developed a 2-layer convolutional neural network [17] on top of word embeddings developed specifically for Biomedical NLP [18]. For training and evaluation, we used the PM assessment from the TREC 2017 evaluation. Run 3 (*SIBTMlit3*) is based on the re-ranking of the run 2 (*SIBTMlit2)* based on the probability that an abstract concerns precision medicine. The weight attributed to this re-ranking is based on the tuning set.

2.1.5 Hierarchical query expansion

Last year, we assumed that an article targeting a more general (supertype) or more specific (subtype) cancer type may still be valuable from a clinical point of view. We used the simplified hierarchy provided by NCIt [19], which only includes concepts in the *Neoplasm by Site* and *Neoplasm by Morphology* categories. This year, we refined this strategy by attributing different scores to supertypes and subtypes, as well as the localization of the disease term (in the title or in the core). Run 4 (*SIBTMlit4*) is based on the linear combination of the run 3 (*SIBTMlit3)* and the run generated by expanding diseases. The weight attributed to each parameter of this re-ranking is based on the tuning set.

2.1.6 Exact run

While previous strategies attempt at maximizing the recall, this run (*SIBTMlit5*) aims at maximizing the precision. Indeed, we tries to retrieve abstracts fully respecting the topic: the disease is the same (or more specific), the gene and variant are the same and the demographic information are either respected or not discussed in the abstract. If this exact run returns less than 1000 results, abstracts returned by run 4 (*SIBTMlit4*) are pushed afterwards.

2.2 Clinical trials retrieval

For the retrieval of clinical trials, we also reused successful strategies investigated in 2017. These strategies mostly rely on a succession of Information Retrieval and filtering steps.

We submitted three runs.

Trials were first filtered according to condition (detected disease). For detecting diseases, we exploited the concepts, synonyms and hierarchy of the National Cancer

Institute (NCI) Thesaurus in order to match concepts in both the topics and the trials. Different sections of the trials were considered: conditions, mesh_conditions, and keywords. In the 2017 campaign, we assumed that a relevant trial should have the corresponding query disease in condition. The results showed that we were far from truth, as many relevant trials – or judged as relevant – did not contain the query disease in the condition; the condition was often stored in free text, or in inadequate sections. Thus, this year, we also detected conditions in title and summary for one run (run SIBTMct1), in order to relax constraints.

Trials were also filtered according to the demographic features. This information is perfectly encoded in trials.

Then, Information Retrieval – with the Terrier platform, and the Okapi BM25 weighting scheme – was used on filtered trials for finding documents related to the query genetic variants. The exclusion criteria were discarded, as we found examples of trials that excluded specific genes in this section. For one run (run SIBTMct2), we also used diseases and specific mutation keywords (such as "mutation"), in order to relax constraints.

3. Results & Discussion

In this section, we present the results for the scientific abstracts retrieval task and the clinical trials retrieval task.

3.1 Scientific abstracts retrieval

3.1.1 Tuning settings

The selection of the best settings for our system relies on the topics and relevance judgments from 2017.

The linear combination of the four different queries uses the following weights: results from Query 1 receives a weight of 0.7, results from Query 2 gets a weight of 0.9 while results from Query 3 and Query 4 are attributed a weight of 0.1. Regarding the drug density run, we observed that a boost of 5 times for the cancer-related drugs performed the best. We obtained the best results when a weight of 0.05 was given to the drugs in the both the title and the abstract. Regarding the demographic information, the best results were obtained when a weight of 0.1 was given to this parameter. The scoring function of the binary classifier had the most impact on our results when using a modest boosting coefficient of 0.15. Although relatively modest, such a result suggests that further works might significantly improve the search task. Regarding the expansion to more general and specific diseases, we obtained the best results when a weight of 0.05 was given to the expanded queries.

3.1.2 Final results

Results for the 50 topics are presented in Table 2. Metrics used for this task are infNDCG, P10 and R-Prec. The infNDCG (inferred non discounted cumulative gain) reflects the gain brought by a document based on its position in the ranked results. P10 (precision at rank 10) represents the proportion of relevant documents retrieved in the top ten results. It thus reflects the ability of the system to retrieve relevant results at high ranks. Finally, R-Prec (R-Precision) return the number of relevant documents returned in the top R document, where R corresponds to the number of relevant documents for the query.

	infNDCG	P10	R-Prec
SIBTMlit1	0.526	0.586	0.354
SIBTMlit2	0.537	0.614	0.356
SIBTMlit3	0.538	0.618	**0.357**
SIBTMlit4	**0.541**	0.626	0.357
SIBTMlit5	0.528	**0.632**	0.340

Table 2 Final results for the 50 topics for the scientific abstracts task

Our first strategy resulted in an infNDCG of 0.526, a P10 of 0.586 and a R-Prec of 0.354. When using in addition the demographic information, our results are improved regarding all measures, respectively of +2.1% for the infNDCG (0.537), +4.8% for the P10 (0.614) and +0.6% for the R-Prec (0.356). Using our precision medicine classifier also resulted in an improvement regarding all measures: +0.2% for the infNDCG (0.538), +0.7% for the P10 (0.618) and +0.3% for the R-Prec (0.357) which was our best results regarding R-Prec. The use of language models computed from word embeddings is promising but yet inconclusive. Indeed, within the top-10 runs, we were one of the rare teams who evaluated the impact of such methods. Further, the hierarchical query expansion had a positive impact regarding the infNDCG (+0.6%) and P10 (+1.3%). This run was our best result regarding infNDCG. Finally, the exact run resulted in mixed results: both the infNDCG and the R-Prec decreased (respectively -2.4% and -4.8%), while the P10 reached its best performance (+1%, 0.632).

3.2 Clinical trials retrieval

Results for the 50 topics are presented in Table 2 3. Metrics used for this task are infNDCG, P10 and R-Prec. As stated in the Methods section, the run 3 can be considered as the baseline. The run 1 investigated the use of title and summary for detecting, then filtering conditions. At last, the run 2 investigated the use of diseases and specific keywords for the Information Retrieval step based on gene and variant information.

	infNDCG	P10	R-Prec
SIBTMct1	**0.430**	**0.404**	**0.318**
SIBTMct2	0.328	0.330	0.250
SIBTMct3	0.335	0.400	0.287

Table 3 Final results for the 50 topics for the clinical trials task

The best run is run 1 for all metrics. Compared to the baseline (run 3), improvement is +28% for infNDCG and +11% for R-Prec. This means that detecting the condition not only in the dedicated section, but also in title and abstract, leads to better performances. P10 is also slightly improved, while title and summary are likely to contain false positives. Comparing run 2 and baseline, we observe that our strategy of query expansion for Information Retrieval was unsuccessful.

Conclusion

While information regarding disease, gene and variant is usually retrieved in full text articles, scientific abstracts reporting on treatments do not always mention all this information. Therefore, the system we developed here for the scientific abstracts task is based on a constraint relaxing strategy, aiming to retrieve a maximum number of potentially relevant abstracts. Further strategies focus on the proper ranking of the retrieved abstracts. Results showed that the fourth first strategies all beneficed to our ranking, with slight improvements among runs. As expected, our final run, favoring publications fully compliant with the topic, resulted in our best results regarding precision at rank 10, while other measures were negatively impacted.

For clinical trials, we investigated strategies for filtering unrelated clinical trials according to the condition (disease), and for retrieving trials relevant for the gene and variant using a search engine. For condition, our assumption since TREC 2017 was that the dedicated disease is included in the trial document in a structured way: the condition tag. All our experiments show that this is not true, and that relaxing constraints leads to better results. Results improve when using synonyms and hierarchy in the NCBI thesaurus, and when detecting condition in different parts of the trial (keywords, title, summary). For genes and variants, none of our query expansion strategies improved our baseline provided by the indexation of the collection by a search engine and the querying with genes and variants names.

Acknowledgments

The study reported in this paper has been partially supported by an SPHN (Swiss Personnalized Health Network) funding [SVIP-O project]. This work also benefited from discussions with Daniel Stekhoven, Franziska Singer and Nora Toussaint.

References

[1] “BiTeM.” [Online]. Available: http://bitem.hesge.ch/. [Accessed: 31-Oct-2018].

[2] J Gobeill, F Ehrler, I Tbahriti, and P Ruch. Vocabulary-driven Passage Retrieval for Question-Answering in Genomics. In TREC. 2007.

[3] J Gobeill, A Gaudinat, E Pasche, D Teodoro, D Vishnyakova, and P Ruch. BiTeM Group Report for TREC Medical Records Track 2011. In TREC. 2011.

[4] J Gobeill, A Gaudinat, E Pasche, D Teodoro, D Vishnyakova, and P Ruch. BiTeM group report for TREC Chemical IR Track 2011. In TREC. 2011.

[5] J Gobeill, A Gaudinat, E Pasche, and P Ruch. Full-texts representation with Medical Subject Headings and co-citations network reranking strategies for TREC 2014 Clinical Decision Support Track. In TREC. 2014.

[6] J Gobeill, A Gaudinat, and P Ruch. Exploiting incoming and outgoing citations for improving Information Retrieval in the TREC 2015 Clinical Decision Support Track. In TREC. 2015.

[7] E Pasche, J Gobeill, L Mottin, A Mottaz, D Teodoro, P Van Rijen, and P Ruch. Customizing a Variant Annotation-Support Tool: an Inquiry into Probability Ranking Principles for TREC Precision Medicine. In TREC. 2017.

[8] P Gaudet, PA Michel, M Zahn-Zabal, A Britan, I Cusin, M Domagalski, PD Duek, A Gateau, A Gleizes, V Hinard, V Rech de Laval, F Nikitin, M Schaeffer, D Teixeira, L Lane, A Bairoch. The neXtProt knowledgebase on human proteins: 2017 update. Nucl. Acids Res. 2016.

[9] N Sioutos, S de Coronado, HW Haber, FW Hartel, WL Shaiu, LW Wright. NCI Thesaurus: a semantic model integrating cancer related clinical and molecular information. J Biomed Inform. 2007:40(1):30-43.

[10] F Minguet, L Van Den Boogerd, TM Salgado, C Correr, and F Fernandez-Llimos. Characterization of the Medical Subject Headings thesaurus for pharmacy. Am. J. Health. Syst. Pharm. 2014:71:1965-72.

[11] V Law, C Knox, Y Djoumbou, T Jewison, AC Guo, Y Liu, A Maciejewski, D Arndt, M Wilson, V Neveu, A Tang, G Gabriel, C Ly, S Adamjee, ZT Dame, B Han, Y Zhou, and DS Wishart. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 2014:42(Database issue):D1091-D1097

[12] JT den Dunnen et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum. Mutat. 2016:37(6):564-569.

[13] YL Yip, N Lachenal, V Pillet, AL Veuthey. Retrieving mutation-specific information for humain proteins in UniProt/Swiss-Prot Knowledgebase. J Bioinform Comput Biol. 2007:5(6):1215-31.

[14] “A to Z List of Cancer Drugs”, National Cancer Institute”. [Online] Available: https://www.cancer.gov/about-cancer/treatment/drugs. [Accessed: 01-Feb-2019].

[15] “List of Cancer Chemotherapy Drugs – Navigating Care”. [Online]. Available: https://www.navigatingcare.com/library/all/chemotherapy_drugs. [Accessed: 01-Feb-2019].

[16] “CareFirst. Oral Chemotherapy Drugs”. [Online]. Available: https://member.carefirst.com/carefirst-resources/pdf/oral-chemotherapy-drug-list-sum2714.pdf. [Accessed: 01-Feb-2019].

[17] A Rios, and R. Kavuluru. Convolutional Neural Networks for Biomedical Text Classification: Application in Indexing Biomedical Articles. In Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics. ACM. 2015:258-267.

[18] B Chiu, G Crichton, A Korhonen, et al. How to train good word embeddings for biomedical NLP. In Proceedings of the 15th workshop on biomedical natural language processing. 2016:166-174.

[19] “NCIt Neoplasm Core Hierarchy By Site and Morphology”. [Online]. Available: https://evs.nci.nih.gov/ftp1/NCI_Thesaurus/Neoplasm/Neoplasm_Core_Hierarchy.html [Accessed: 01-Feb-2019].