Identifying Breast Cancer Concepts in SNOMED-CT Using Large Text Corpus

Zharko Aleksovski and Merlijn Sevenster

Philips Research Europe,
High Tech Campus 37 (room 2.044),
5656 AE Eindhoven, The Netherlands
zharko.aleksovski@philips.com

Abstract. Large medical ontologies can be of great help in building a specialized clinical information system. First step in their use is to identify the subset of concepts which are relevant to the specialty. In this paper we present a method to automatically identify the breast cancer concepts from the SNOMED-CT ontology using large text corpus as source of knowledge. In addition to finding them, the concepts are also assigned relevance values.

In our experiments the method produced results of an overall high quality. The precision was high, and the recall was relatively low, but the concepts which were not found are complex and arguably ambiguous, which limits their applicability in practice. This research was application driven, and the breast cancer concepts found have been applied in a real oncology information system.

Keywords: ontology, SNOMED-CT, breast cancer, term frequency.

1 Introduction

Large medical ontologies such as SNOMED-CT contain hundreds of thousands of clinical concepts usually organized in a hierarchy and interconnected by domain specific relations, together representing the explicit semantic knowledge describing a medical field. For a given application it is often desirable to restrict oneself to a smaller subontology. But the relevant concepts are rarely found under one sub-branch of the large ontology, instead they are usually scattered over multiple high-level categories, e.g. clinical findings, procedures, body locations, etc.

In this paper we describe a study on the identification of breast cancer concepts in SNOMED-CT. We use a large text corpus of medical documents, a portion of which is dedicated to breast cancer, and by analyzing how frequently SNOMED-CT concepts occur in different parts of the corpus we measure how relevant they are to breast cancer.

Our experiments show that large text corpora of medical literature can be successfully used to identify the concepts relevant to a clinical setting, in our case that is breast cancer. The concepts are also assigned relevance score, such that concepts that are key to breast cancer receive highest score. Evaluating the

M. Szomszor and P. Kostkova (Eds.): E-Health 2010, LNICST 69, pp. 27–34, 2011.

ranked list of concepts revealed that our method exhibits high accuracy: all the top concepts are relevant to breast cancer and the relevancy decreases as we move down the list.

In addition to identifying the breast cancer concepts, we examine how complete is the resulting list. First, we find that there are breast cancer concepts in SNOMED-CT that we did not find, because most of them could not be extracted from the text due to their complex names. This characteristic makes these breast cancer concepts somewhat difficult to use in practice, and it is therefore not a serious drawback of our method. Second, we test if SNOMED-CT is rich enough to cover the breast cancer field. Using a different experimental setup we show that SNOMED-CT does not seem to lack any key breast cancer concepts.

Having identified these breast cancer concepts enables variety of applications. The very basic ones are highlighting important parts in medical documents, providing more relevant autocompletion suggestions, and building fulltext search engine specialized for oncology related documents.

Related Work. Identifying disease-centric subdomains in medical ontologies has been studied in [2,6]. The first study uses a predefined set of queries as a source of knowledge to find the concepts, and the second uses formalized clinical guidelines. They expand these initially found sets using the ontology structure, and also the UMLS[1] meta-ontology. Importantly, these studies define a subdomain as a set of concepts which excludes the possibility that concepts can be relevant to varying degree. The fact that we allow gradual relevance measure is major advantage of this work.

The *TFIDF* measuring scheme is well established in the field of information retrieval. Its basic use is to assign representative keywords to individual documents in a collection, but it has been modified and extended in various ways [1]. Perhaps the most similar application of *TFIDF* to our work is reported in [5] where it is used to find topic-relevant keywords.

Finally, the problem of identifying topic-relevant concepts in ontology can be compared to ontology modularization [3,4]. Ontology modules are autonomous parts of ontologies intended for reuse, and identifying these modules to some extent resembles our problem.

The paper is structured as follows: in the next section we describe our method in detail together with the experimental results. Sections 3 and 4 report on experiments to test the completeness of the results: in Section 3 we try to find concepts in SNOMED-CT that we might have missed, and in Section 4 we test if SNOMED-CT is complete enough and if it contains all the breast cancer important concepts. Section 5 presents the concluding remarks.

2 Methodology: Finding Breast Cancer Concepts in SNOMED-CT

Concepts that are frequently used in the context of breast cancer are relevant to the topic: the more frequently used - the more relevant the concept is. But breast

[1] `http://www.nlm.nih.gov/research/umls/`

cancer concepts should also have "exclusivity". When they are used primarily to talk about breast cancer and no other things, that makes them more relevant. Terms like Patient or Disease are frequently used in breast cancer, but due to their general meaning they carry little information in medical context, so they are not among the most relevant breast cancer concepts. Our method uses a ranking scheme that makes trade-off between these two requirements. Before presenting the details of the method, we first describe the data that was used in the experiments.

2.1 Experimental Data

We use large medical text corpus as source of knowledge to identify the breast cancer concepts in SNOMED-CT. Because part of the corpus were documents about breast cancer and the rest about other medical topics, we were able to observe how often individual SNOMED-CT concepts occur in breast cancer documents, and how often in other medical documents.

SNOMED-CT(*Systematized Nomenclature of Medicine - Clinical Terms*), is a systematically organized computer-processable collection of medical terminology covering most areas of medicine such as diseases, findings, procedures, etc. [9]. The version used in this study is from July 2009 and consists of 307,754 concepts interconnected with over 100 relation types. The concepts are organized in 19 mutually exclusive hierarchies called *SNOMED categories*. Some of these categories are: *Body structure*, *Clinical finding*, etc. SNOMED-CT has been under active development for more than 35 years, and is constantly evolving.

Text Corpus of Medical Documents. The text corpus comprises 103 medical digital documents, mostly books. We divided the documents in two: *the breast cancer corpus* of 11 documents about breast cancer, and *the general medical corpus* of the remaining 92 documents about other medical topics. The breast cancer corpus consisted of 10 million characters, and the general medicine corpus consisted of 512 million characters. The texts were extracted from electronic documents in formats like PDF and PS.

2.2 The Method

The method proceeds in two steps: *annotation* and *TFIDF ranking*. In the annotation step we extracted SNOMED-CT concepts from the text corpus. Single occurrence of SNOMED-CT concept in a document we call an *annotation*. After an exhaustive annotation process we counted how many times each concept was annotated. Then we fed these numbers into an adapted *TFIDF* ranking scheme, which produces the resulting ranked list of breast cancer concepts.

Annotation. In the annotation step we lexically searched for SNOMED-CT concepts in the text corpus. We considered that a concept is annotated when its label or one of the synonyms was found in the text.

The search itself allowed for some flexibility. The case of the words was ignored, stopwords were ignored, the order of the words was ignored - as long as they appeared in consecutive sequence, and before comparing the words were stemmed using the Porter stemmer algorithm [7].

Results of the Annotation. The breast cancer corpus established 1,259,844 annotations to 12,647 different concepts. That is 99.6 annotations per annotated concept on average. These annotations were distributed very unequally over the concepts. After ranking the concepts by the number of annotations on average the top 5 had more than 10,000 annotations each, the top 100 had more than 2,100 annotations each, and the top 4,560, which is 36% of the annotated concepts, had 10 or more annotations. The general medicine corpus established 64,248,152 annotations to 30,092 different concepts, and again the number of annotations were very unevenly distributed per concept comparable to the situation with the breast cancer corpus.

TFIDF. (Term frequency - inverse document frequency) ranking measure [8] is used in information retrieval to estimate the importance of terms to particular document in a collection of documents. It is calculated as

$$TFIDF(t,d) = TF(t,d) \times IDF(t)$$

where $TF(t,d)$ is the relative term frequency of the term t within the document d, which is the number of times t occurs in d divided with the number of all term occurrences in d. $IDF(t)$ is a measure of how general the meaning of the term t is. It is obtained by dividing the number of documents in the collection by the number of documents containing the term, and then taking the logarithm of that quotient. If D is the collection of documents, then IDF is calculated as

$$IDF(t) = \log \frac{|D|}{|\{d : t \in d\}|}$$

The intuition behind the $TFIDF$ measure is that a term is descriptive to a document if it occurs frequently in the document, and is infrequent in the other documents in the collection. These properties perfectly fit the requirements about the relevance ranking scheme that we discussed in the beginning of this section. We used the scheme differently than it was originally intended as we counted occurrences of concepts and not terms. For instance, when breast cancer or malignant tumor of breast were found in the text they were counted as two occurrences of the same concept, even though they are different terms.

Being interested in how important individual SNOMED-CT concepts are to breast cancer we concatenated the breast cancer documents into one single document. We also cut the general medical documents into smaller chunks of predefined size of maximum 50,000 characters and considered each of them as separate document. This was needed because the general medical documents were mainly large books with average of 5 million characters, and even though written on other medical topics breast cancer specific concepts occurred in most of these documents which was not a desired property for our ranking scheme[2]. After conducting several experiments we choose size of 50,000 characters per chunk (and

[2] Good solution to this problem would be to extract the separate sections of these books, but because of the overwhelming manual effort required we turned to this less appealing automatic solution.

also include the leftover with smaller size as separate document), though, finding the optimal size can be a subject of further investigation. Finally, we restricted our focus to concepts that are annotated at least 10 times in the corpus, discarding the ones with fewer annotations. They can neither score high $TFIDF$ value, nor can they change the ranking order of the other concepts.

2.3 Results

The method produced an ordered list of 4,560 concepts. The top 10 concepts in the list, i.e. the concepts with highest $TFIDF$ score are shown in Table 1. Table 2 shows other parts of the list, the first 3 concepts starting from the 100, 200 and 500^{th} position respectively.

Table 1. Top 10 most relevant breast cancer concepts found by the method

Concept's label	SNOMED-CT code	TFIDF
1. Breast cancer	254837009	0.0002434223
2. Mamma	181131000	0.0002404850
3. Breast	76752008	0.0002389841
4. Malignancy	363346000	0.0001399625
5. Cancer	86049000	0.0000966287
6. Mammogram	71651007	0.0000854837
7. DCIS	86616005	0.0000617422
8. Mastectomy	172043006	0.0000486403
9. Excision of breast tissue	69031006	0.0000462173
10. Tamoxifen	373345002	0.0000365901

Table 2. Selected concepts with their ranking in the breast cancer list of concepts

Concept's label	SNOMED-CT code	TFIDF
100. PET - Positron emission tomography	82918005	0.0000091759
101. FH - Family history	57177007	0.0000091479
102. Development of the breasts	364375002	0.0000089859
200. Atypical hyperplasia	32416003	0.0000062063
201. Specimen	123038009	0.0000061673
202. Has specimen	116686009	0.0000061673
500. Dense	255596001	0.0000032862
501. Phenotype finding	8116006	0.0000032774
502. Interested	225469004	0.0000032748

Precision. The first 10 concepts in the list are clearly key breast cancer concepts. As we go down the list it becomes harder to evaluate how the concepts are ranked. We manually inspected the first 100 concepts, and another set of 100 randomly drawn concepts from the whole list. Of course, the ranking of each individual concept can be debated, but there was no concept in the evaluation sets for which we could say that it is wrongly ranked. This evaluation suggested that the results are very precise.

When we rank the concepts by the TF component only, interestingly, still most of the top concepts are key breast cancer concepts like Breast cancer or

Breast, but when we look at the first 30 concepts then we also find generic concepts like Patient, Study and Clinical. On the other hand, when we rank by IDF only, the core concepts like Breast cancer or Breast are down in the ranking, and the list is topped by concepts that are very specific to breast cancer like Nipple preserving subcutaneous mastectomy or Mammographic breast density.

3 Evaluation Experiment I: Completeness of the Results

In the previous section we briefly discussed the precision of our results, and now we look into the recall, that is, we test if breast cancer concepts in SNOMED-CT were missed. To make sure that we found all the breast cancer concepts, we would have to check for each SNOMED-CT concept. This is unrealistic due to the size of SNOMED-CT - over 300,000 concepts, so we choose for alternative. The so-called "seed queries" method reported in [2] also extracts a subdomain from an ontology but in a very different way. It is reported to have high precision. Comparing with the seed queries can give an indication of the recall of our method. For this comparison we used simplified version of the seed queries method.

According to this method we searched for concepts in SNOMED-CT using six queries: *breast cancer*, *breast carcinoma*, *breast neoplasm*, *breast tumor*, *ductal carcinoma* and *mastectomy*. Each concept that contains all the words from one of the queries is found in this search. The search found 355 concepts.

Comparing with the Seed Queries. Of the 355 concepts found by the seed query method, our method finds 24 concepts, which estimates the recall of our method as compared to seed queries, to only 7%, which is very low. We analyzed the seed queries results to find the reasons for this low recall. Majority of the missed concepts have very precise meanings which is reflected in their linguistically complex labels. Below are some representative examples:

94964004	Neoplasm of uncertain behavior of nipple of female breast
373182002	pT2: Tumor $>$ 2 cm but $\leq$ 5 cm in greatest dimension (breast)
94182000	Metastatic malignant neoplasm to axillary tail of female breast

One possibility is that our annotation technique was not good enough to annotate these concepts to the text corpus. For this reason we used a state-of-the-art tool called *MetaMap*, which is specialized for concept annotation in free text[3]. We run the MetaMap tool on the breast cancer corpus, and it managed to annotate 7 concepts of these found by the seed queries, which is even less than what our annotation found[4]. Hence, not finding these concepts in the free text is not necessarily a weakness of our annotation method.

Since annotation tools fail to find these concepts in free text, they cannot be used in applications that require their automated discovery in text, such as highlighting important parts in medical record. So, having missed these concepts is not a serious drawback of our method.

[3] MetaMap is developed as part of the UMLS project: `http://mmtx.nlm.nih.gov/`

[4] This comparison does not reflect the quality of the MetaMap tool because it is general-purpose annotation tool not tailored to the requirements of our study.

4 Evaluation Experiment II: SNOMED-CT Coverage of Breast Cancer

In the previous section we assessed the completeness of our method, i.e. if we have found all the breast cancer concepts that are in SNOMED-CT. In this section we assess the completeness of SNOMED-CT, i.e. whether it contains all the important breast cancer concepts. Now we analyze the text corpus alone, and construct a list of terms from the text that are important to breast cancer. If an important breast cancer term is not present in SNOMED-CT we hope to find it in this list.

The method of Section 2.2 finds the breast cancer concepts by looking at which concepts are representative for the breast cancer part of the text corpus. Now, we look at which terms are representative for the breast cancer part of the text corpus. We calculate the *TFIDF* importance of every term that occurs at least 10 times in the breast cancer corpus. The same setup as in Section 2.2 was used: the breast cancer documents are put together in a single document, the general medical documents are chopped into chunks of max. 50,000 characters, and when comparing if two terms are equal we were flexible as in Section 2.2: ignore stopwords, ignore word case, ignore word order and use stemming.

Results. This experiment reported 21,519 terms. Due to the large size, we restrict the evaluation to the top 1,000 terms in the list. For each of these terms we searched in SNOMED-CT if a concept has it as a label, and for those not found we investigated as to why it was not found. If it was an important breast cancer term not given a SNOMED-CT concept we would expect to find it here.

In the first 1,000 terms we did not identify any meaningful term that did not have appropriate concept in SNOMED-CT. This means that SNOMED-CT is very complete in describing the most important terms used to communicate about breast cancer. Most of the terms not found in SNOMED-CT were not valid noun terms, for example breast and ovarian or early breast. Also there were other artifacts as well, like 2005, riskof or Clin Oncol which might have occurred due to imperfections in the preparation of the test data, or simply because these artifacts are only being used in an informal communication, and hence are not given appropriate concepts in SNOMED-CT.

5 Conclusions

We presented a novel method to automatically extract topic-based concepts from an ontology using large text corpus. The method was applied to extract the breast cancer concepts from the SNOMED-CT ontology. It produced a ranked list of concepts with good enough quality to be applied in practice.

We conducted three rounds of evaluation on the quality of the results. First, testing the precision showed that the top of the list are all related to breast cancer. Second, the evaluation experiment 1 showed that the recall of the method is low, but we concluded that this is not a serious drawback for the method. The missed concepts are complex and have limited usefulness in practice because

even state-of-the-art tools cannot automatically find them in free text. Third, the evaluation experiment 2 showed that SNOMED-CT is complete in covering the terminology used in breast cancer, and can be used in real clinical information systems designed to support the breast cancer care cycle.

References

1. Aizawa, A.: An information-theoretic perspective of tf-idf measures. Information Processing & Management 39(1), 45–65 (2003)
2. Aleksovski, Z., Vdovjak, R.: Overlap of selected ontologies in the context of the breast cancer domain. In: Proceedings of SIIM Annual Meeting (2009)
3. Clark, K., Parsia, B.: Modularity and owl. Literature survey (2008)
4. Cuenca Grau, B., Horrocks, I., Kazakov, Y., Sattler, U.: Just the right amount: extracting modules from ontologies. In: Proceedings of WWW, pp. 717–726 (2007)
5. Lawrie, D., Croft, W.B., Rosenberg, A.: Finding topic words for hierarchical summarization. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 349–357. ACM, New York (2001)
6. Milian, K., Aleksovski, Z., Vdovjak, R., ten Teije, A., van Harmelen, F.: Identifying disease-centric subdomains in very large medical ontologies: A case-study on breast cancer concepts in snomed ct. or: Finding 2500 out of 300.000. In: Riaño, D., ten Teije, A., Miksch, S., Peleg, M. (eds.) KR4HC 2009. LNCS, vol. 5943, pp. 50–63. Springer, Heidelberg (2010)
7. Porter, M.F.: An algorithm for suffix stripping, pp. 313–316. Morgan Kaufmann Publishers Inc., San Francisco (1997)
8. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)
9. Stearns, M.Q., Price, C., Spackman, K.A., Wang, A.Y.: Snomed clinical terms: overview of the development process and project status. In: Proceedings of the AMIA Symposium, p. 662. American Medical Informatics Association (2001)