MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190529

Unsupervised Phrase-Level Query Rewriting for Assisting Search in Clinical Free Text

Hans Moen[a], Laura Peltonen[b], Henry Suhonen[b], Mikko Koivumäki[b], Tapio Salakoski[a], Sanna Salanterä[b]

[a] *Department of Future Technologies, University of Turku, Turku, Finland,*
[b] *Department of Nursing Science, University of Turku, Turku, Finland*

Abstract

We report on the pilot evaluation of an experimental query-based search functionality that enables phrase-level query rewriting in an unsupervised way. It is intended for supporting search in clinical text. Qualitative evaluation is done by three clinicans using a prototype search tool. They report that they find the tested search functionality to be beneficial for making query-based searching in clinical text more efficient.

Keywords:

Natural Language Processing, Information Storage and Retrieval, Electronic Health Records.

Introduction

Query-based searching for information in clinical free text, such as information about a specific health-related phenomena or concept, can be challenging as the same concept can be described in many different ways. An important job for a search engine is to try to bridge the gap between user queries and how associated phrases of similar meaning (semantics) are documented in the targeted text. The aim of the presented work is to test an experimental search functionality that aims to find, suggest and highlight phrases which have similar meaning and length as queries provided by the user. Intended users of a search engine with this functionality are clinicans who are looking for previously documented information in a patient's electronic health record (EHR). One example could be a clinician who is interested in reading what, if any, information concerning *confusion symptoms* has been documented by searching with queries similar to how he/she would typically document such symptoms.

Unlike traditional information retrieval that focuses on retrieving relevant documents, paragraphs or sentences, we are focusing on phrase-level matching and highlighting. Labeled training data, e.g. in the form of search history logs (see e.g. [1]) for enabling supervised query rewriting and expansion techniques, is typically not readily available from EHR systems. Thus, our approach relies primarily on un-labeled free text for training. We are not aware of related works that have proposed suitable, unsupervised approaches/methods for enabling the search functionality that we are aiming for.

The two main challenges in the identified scenario are: 1) Individual words may have several synonyms, near synonyms, and/or closely related words which refers to the same or similar underlying concept (e.g., "oxygen" vs. "SaO2" (oxygen saturation)); 2) When using multiple words, more complex concepts may be expressed and there are typically a greater number of ways to describe a single concept with variations in the choice of words, compositionality and length (e.g., "DM II" vs. "type 2 diabetes mellitus," see also Table 1).

Our approach/method is based on unsupervised machine learning and uses primarily three components: A statistical language model (KenLM toolkit [2]); an off-the-shelf search engine (Apache Solr search platform); and a semantic model of word n-gram vectors (or embeddings) trained with the Word2Vec toolkit [3] – which learns semantics from word co-occurrence statistics in a large text corpora in an unsupervised way. Such models of distributional semantics have been shown to capture word synonymy and relatedness. Similar to Zhao et al. [4], we combine co-occurrence statistics from word n-grams of two different sizes: unigrams (single words) and bigrams (word pairs) when training a single semantic model. One motivation for doing this is that we want to be able to map from unigrams to semantically similar bigrams and vice versa. Another motivation is that Zhao et al. [4] found this to produce improved word representations compared to only using unigram co-occurrence statistics.

We hypothesize that the proposed search functionality (described in the Methods section) can be beneficial to clinicians for saving time and effort when seeking for information of interest in clinical free text. Evaluation of the search functionality is enabled and tested through a prototype query-based search tool/interface. We report on a pilot evaluation performed by three domain experts. The evaluation provides qualitative feedback on how the search functionality performs, and highlights strengths and weaknesses as seen by the evaluators.

Methods

The data set we use is a relatively large corpus of clinical text consisting of physician notes and nursing shift notes from patients admitted to a Finnish hospital. It consists of 136 million tokens (1.5 million unique tokens). For training the semantic model, we first preprocess a version of the corpus with lowercasing, tokenization and stemming (Snowball stemmer for Finnish). We decided to use stemming primarily to reduce the vocabulary size. Next, the text is converted into uni- and bi-grams. As an example, the sentence "a nice flower" becomes: "a a_nice nice nice_flower flower". We train the semantic model with Word2Vec using the SkipGram architecture and hierarchical softmax. We use a dimensionality of 300 and a window size of 6. To train the statistical langauge, we use a version of the corpus only containing stemmed and lowercased unigrams.

Briefly explained, the way we generate rewritten candidate suggestions for a given query is as follows. First, we create two query vectors by splitting the query into uni- and bi-grams before summing the associated vectors (normalized) from the semantic model. If, let us say, the query is: "this is a query", we create two query vectors as follows:

$$\text{qvec}_{\text{unigram}} = \overrightarrow{this} + \overrightarrow{is} + \vec{a} + \overrightarrow{query}$$
$$\text{qvec}_{\text{bigram}} = \overrightarrow{this\_is} + \overrightarrow{is\_a} + \overrightarrow{a\_query}$$

Next, these are both used to extract the semantically most similar unigrams and bigrams using the semantic model and cosine similarty (sim) as vector distance measure. In addition to forming individual candidate suggestions, we use them to generate multi-word phrase candidates. For bigrams, this is done by iteratively combining bigrams that have one overlapping word in order to create longer phrases (e.g., "is_a" and "a_solution" is combined to "is_a_solution"). To reduce the number of nonsensical phrases being generated, we use the statistical language model to iteratively assess whether or not a potential phrase candidate is likely to exist in the corpus. If no additional n-grams can be added to any of the phrase candidates, the process stops. Another stopping criteria we introduce is a max length relative to the length of the query. Next, we again use the semantic model, this time to create phrase vectors ($p_i$vec) for each phrase candidate in the same way as we did with the query (i.e., $p_i\text{vec}_{\text{unigram}}$ and $p_i\text{vec}_{\text{bigram}}$). Then, we calculate a similarity score between the query and each of the phrases before sorting them according to their similarity score. As similarity function between a query (q) and a phrase candidate (p_i) we use:

$$\text{sim}(q, p_i) = \text{avg}(\text{sim}(\text{qvec}_{\text{unigram}}, p_i\text{vec}_{\text{unigram}}), \text{sim}(\text{qvec}_{\text{unigram}}, p_i\text{vec}_{\text{bigram}}), \text{sim}(\text{qvec}_{\text{bigram}}, p_i\text{vec}_{\text{unigram}}), \text{sim}(\text{qvec}_{\text{unigram}}, p_i\text{vec}_{\text{bigram}}))$$

As a final step, we use the search engine (Apache Solr search platform) where we have indexed the clinical notes in the care episodes to search in. We use a filter that enables matching stemmed queries with their inflected forms in the indexed text. Finally, the search engine is used to find and highlight the top *n* phrase candidates that actually occur in the targeted care episode. Candidates not found are discarded.

Prototype Search tool and Experimental Setup

We asked three domain experts with a background as a hospital nurse to test the search functionality through a prototype query-based search tool/interface. The purpose was to have them evaluate how well the (rewrite) suggestions by the tool helps them to better find the information that they are searching for (if it exist at all), compared to only the exact matches of the query. When searching with a query, the tool highlights in the clinical notes/documents exact matches (if any) and additional rewritten phrase suggestions found with unique colors. Each unique match/suggestion are also listed in a separate table showing their similarity score to the query and occurrence count in the targeted care episode. The evaluators were given a set of patient phenomena to search for, including *state of mind*, *smoking status*, *secretion* and *activity level*. They were also encouraged to search for other phenomena of interest. We also encouraged them to use multi-word queries when searching. 20 different care episodes (i.e. all physician and nursing notes from 20 patients' hospital stays) were indexed and searchable (one at a time). The evaluators were given a set of questions to consider and comment on while they were doing the testing. These were questions about whether or not the tool made it easier and faster to find the information they were looking for (compared to only relying on the exact matches), weaknesses, strengths, problems and suggestions for improvements. Finally, their answers and comments formed the basis of a joint feedback meeting.

Results

All three evaluators reported that the suggestions provided by the tool makes it easier and faster to search for and find information in clinical text. Since the way clinicans document various phenomena can vary greatly, the rewrite suggestions help to identify potential words and phrases that are semantically related to the search query but written with different words. Even though some of the suggestions by the tool were not relevant, they reported that it was easy to simply ignore these. As problems, they reported that it can be difficult to know that something is not present in a care episode even if it is not found by the tool (false negatives). Also, they noticed that the use of stemming was not optimal as different inflections of the same word were not always connected and found. Examples of rewrite suggestions provided by the tool for a few queries can be seen in Table 1.

Table 1– Example of Query Rewrite Suggestions Found by the Tool. Translated from Finnish to English.

Query	Rewrite suggestions
"runs daily"	"exercise a lot"
"speaks funny things"	"speaks nonsense" "speaks delirious" "speaks by himself"
"patient sees small green men"	"sees a lot of things" "sees things that are not there" "sees illusions"

Conclusions

We present our ongoing work toward enabling unsupervised phrase-level query rewriting to support searching in clinical text. The evaluators report that they find the prototype search tool, with the underlying query rewrite functionality, to be beneficial for making searching and finding information in clinical text faster and easier compared to only exact query matching. As future work we plan to perform a quantitative evaluation. We also aim to test using this search functionality to supporting manual annotation of clinical text. Additional plans include looking into alternatives to stemming, the inclusion of trigrams in the semantic model(s), other similarity measures, query segmentation and algorithm optimization.

Acknowledgements

The research was supported by the Academy of Finland.

References

[1] Y. He, J. Tang, H. Ouyang, C. Kang, D. Yin, and Y. Chang. Learning to rewrite queries. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016, 1443–1452.

[2] K. Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2011, 187–197.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, 2013, 3111–3119.

[4] Z. Zhao, T. Liu, S. Li, B. Li, and X. Du. Ngram2vec: Learning improved word representations from ngram co-occurrence statistics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, 244–253.

Address for Correspondence

Corresponding author: Hans Moen. E-mail: hans.moen@utu.fi.