Towards Retrieving and Ranking Clinical Recommendations with Cross-Lingual Random Indexing

Hans Moen, Erwin Marsi

Norwegian University of Science and Technology,
Department of Computer and Information Science,
7491 Trondheim, Norway
hans.moen@idi.ntnu.no, emarsi@idi.ntnu.no

Abstract. Clinicians have to deal with large amounts of textual data, searching for and navigating information that satisfies their informational needs. *Clinical notes* and *Clinical Practice Guidelines* (CPG) are textual resources that usually contain *free text*. As language use in the medical domain is rather specialized, generic information retrieval tools are suboptimal for such data. This calls for specialized information retrieval systems that incorporate domain-specific knowledge. Furthermore, information resources for relatively small languages are limited, so physicians speaking those languages often have to resort to information sources in English or one of the other major languages spoken. In our setting, given a patient record written in Norwegian, a physician is looking for recommendations in CPGs written in English. The recommendations are to be retrieved and ranked according to their relevance to (specified parts of) the patient's medical history and possibly a user-supplied query. This means that the query language and the resource language differ, requiring some form of C*ross-Lingual Information Retrieval*. Most approaches rely on translating the input query to the target language using bilingual dictionaries or machine translation systems. This raises the familiar problems in machine translation such as lack of lexical coverage and lexical translation ambiguity. We propose an alternative method that avoids direct translation of queries by implicit encoding of cross-lingual semantic relations in a vector space model. We modify the *Random Indexing* model in such way that distributional similarity between terms or documents can be measured across different languages. The key idea is to assign identical index vectors to source language terms and their target language translations during construction of the models. This requires only a translation dictionary and large source/target language text corpora, which do not need to be aligned. The dictionary does not need to have full coverage as long as it includes the most frequent words. Additional target languages can be added without having to re-train the existing models. So far we have implemented this C*ross-Lingual Random Indexing* model and conducted some pilot experiments on retrieval of semantically related terms in Norwegian and English, with encouraging results. Our contribution furthermore discusses the data required for further training and evaluating such models, as well as the need for expert knowledge in data labeling.

Keywords: Cross-Lingual; Health Data; Information Retrieval

Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis - CLEFeHealth 2012 Workshop, edited by Hanna Suominen

1 Introduction

As the health sector is growing, clinicians are confronted with an ever-growing amount of information stored in electronic format. In many countries, most of this information is free text rather than structured data, and a large part of it is stored in electronic health record (EHR) systems. The total amount of data stored in such systems throughout hospitals is large and continuously growing. As a result, clinicians have to deal with large amounts of textual data, searching for information that satisfies their informational needs. As language use in the medical domain is rather specialized, generic information retrieval tools are suboptimal for the task. This calls for specialized information retrieval systems that incorporate domain-specific knowledge. In addition, resources for relatively small languages are limited so that physicians speaking and writing in those languages often have to resort to information sources in English or one of the other major languages spoken. This means that the query language and the resource language often differ, requiring some form of cross-lingual information retrieval (CLIR).[1]

Within this general context, our goal is to aid physicians in their daily work of searching for and navigating relevant clinical practice guidelines (CPGs) and underlying recommendations. The point of departure is a patient record written in Norwegian, possibly combined with a free text query. A physician is looking for recommendations in CPGs written in English. The recommendations are to be retrieved and ranked according to their relevance to (specified parts of) the patient's medical history. This abstract describes a possible approach to solving this problem using semantic language models based on Random Indexing for statistical analysis of large amounts of text. As this is work in progress, we focus on a description and discussion of the approach and required data, leaving experimental results to future work.

2 Approach

CLIR aims at identifying relevant documents in a language other than that of the query.[1] Most approaches rely on translating the query to the target language using bilingual dictionaries or machine translation systems. This raises the familiar problems in machine translation such as lack of lexical coverage and lexical translation ambiguity. We propose an alternative method that avoids direct translation of queries by implicit encoding of cross-lingual semantic relations in a vector space model. This is motivated by the cross-lingual distributional similarity hypothesis:[2] If in a text of one language two words A and B co-occur more often than expected by chance, then in a text of another language those words that are translations of A and B should also co-occur more frequently than expected. For example, suppose a query-term in Norwegian frequently co-occurs with certain neighboring terms as observed in a Norwegian text corpus, then English terms that frequently co-occur with *English translations of the same Norwegian neighboring terms* (according to a bilingual dictionary) in an English text corpus are likely to have a high contextual similarity to the Norwegian query term. Notice that we do not make the (stronger) assumption that the Norwegian query term and the cross-lingually co-occurring English terms are proper translations of each other, but merely that the two are semantically related, and that this semantic relation can be exploited in CLIR.

In order to exploit cross-lingual distributional similarity for CLIR, we use the Random Indexing (RI) model for distributional similarity.[3] RI incrementally builds a vector space model of information units (terms in this case) in the following way. First, a unique *index vector* (a sparse high-dimensional vector consisting of mainly zeros with a couple of randomly chosen +1/-1 components) is assigned to each entry in the translation dictionary; the associated translations receive the same index vector. A term in the source language will thus have the same index vector as its translations in the target language. Next, a *context vector* is created for all terms in either language. This is accomplished by sliding a context window over the text corpus. For each term in the center of the window, the index vectors of the neighboring words within the window are added to its context vector. A term's context vector therefore encodes the terms it co-occurs with. Context vectors for larger text spans such as sentences, sections or documents are subsequently created by summing, and possible weighting (cf. TF*IDF), the context vectors of all terms contained. Structured domain knowledge can here be applied to further adjust the weightings of these vectors. This procedure assumes only a translation dictionary and large source/target language text corpora that does not need to be aligned in terms of content. The dictionary does not need to have full coverage as long as it includes a substantial part of the most frequent words, and words that are not in the dictionary will still be assigned context vectors. Additional target languages can be added in a similar way without having to re-train the existing models. For retrieval, a query is converted to a query vector by summing the context vector of all query terms in the source language. The query vector is then directly matched against pre-computed document vectors in the target language(s) using cosine similarity, resulting in a ranking of the most similar documents.

3 Discussion

So far we have implemented the outlined approach and conducted some pilot experiments on retrieval of semantically related terms in Norwegian and English, with encouraging results. Our short-term goals are to evaluate on the CLEF data for CLIR, followed by tests on real health data and for the specific task of searching for and ranking recommendations in existing CPGs automatically. The later requires three types of data for training:

1. a set of health records in the source language;
2. a set of CPGs written in the target language, containing recommendations relevant w.r.t. the health records (preferably structured so that individual recommendations are clearly separated, e.g. similar to the structure of the AT9 guidelines);
3. a bilingual dictionary covering the most frequent terms, preferably including domain-specific terminology.

Evaluation requires a gold standard consisting of pairs of health records, or a subset of clinical notes from these, and an associated set of ranked recommendations originating from the CPGs. Since a health record contains notes with different timestamps and written by physicians with different roles, it might be necessary to be selective when it comes to which clinical notes to use as basis for the query. Due to the sensitive content in health records, such data is difficult to acquire. Nor is it an easy task to get access to the necessary expert knowledge required to develop such a gold standard. Development of these evaluation data for Norwegian-English is currently under way.

4 Conclusion

There is a potential in aiding clinical work through retrieving and ranking CPGs using CLIR. The method described here is a possible way of accomplishing such a task. However, there is a need for establishing a suitable dataset for training and validating this and similar approaches.

Acknowledgements

This work was partly funded by the EviCare project (`http://www.evicare.no`) and by the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement nr 248307 (PRESEMT).

References

1. Kishida K. Technical issues of cross-language information retrieval: a review. *Information Processing & Management*. 2005;41(3):433-455.
2. Rapp R. Identifying word translations in non-parallel texts. Association for Computational Linguistics: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. 1995:320-322.
3. Sahlgren M. An introduction to random indexing. *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*. 2005:1-9.