Artificial Intelligence in Medicine xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Adaptation of machine translation for multilingual information retrieval in the medical domain

Pavel Pecina[a,*], Ondřej Dušek[a], Lorraine Goeuriot[b], Jan Hajič[a], Jaroslava Hlaváčová[a], Gareth J.F. Jones[b], Liadh Kelly[b], Johannes Leveling[b], David Mareček[a], Michal Novák[a], Martin Popel[a], Rudolf Rosa[a], Aleš Tamchyna[a], Zdeňka Urešová[a]

[a] Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, Malostranské nám. 25, 118 00 Prague 1, Czech Republic
[b] CNGL Centre for Global Intelligent Content, School of Computing, Dublin City University, Glasnevin, Dublin 9, Ireland

ARTICLE INFO

Article history:
Received 3 July 2013
Received in revised form 15 January 2014
Accepted 28 January 2014

Keywords:
Statistical machine translation
Domain adaptation of statistical machine translation
Intelligent training data selection for machine translation
Compound splitting
Cross-language information retrieval
Medical query translation

ABSTRACT

Objective: We investigate machine translation (MT) of user search queries in the context of cross-lingual information retrieval (IR) in the medical domain. The main focus is on techniques to adapt MT to increase translation quality; however, we also explore MT adaptation to improve effectiveness of cross-lingual IR.
Methods and data: Our MT system is Moses, a state-of-the-art phrase-based statistical machine translation system. The IR system is based on the BM25 retrieval model implemented in the Lucene search engine. The MT techniques employed in this work include in-domain training and tuning, intelligent training data selection, optimization of phrase table configuration, compound splitting, and exploiting synonyms as translation variants. The IR methods include morphological normalization and using multiple translation variants for query expansion. The experiments are performed and thoroughly evaluated on three language pairs: Czech–English, German–English, and French–English. MT quality is evaluated on data sets created within the Khresmoi project and IR effectiveness is tested on the CLEF eHealth 2013 data sets.
Results: The search query translation results achieved in our experiments are outstanding – our systems outperform not only our strong baselines, but also Google Translate and Microsoft Bing Translator in direct comparison carried out on all the language pairs. The baseline BLEU scores increased from 26.59 to 41.45 for Czech–English, from 23.03 to 40.82 for German–English, and from 32.67 to 40.82 for French–English. This is a 55% improvement on average. In terms of the IR performance on this particular test collection, a significant improvement over the baseline is achieved only for French–English. For Czech–English and German–English, the increased MT quality does not lead to better IR results.
Conclusions: Most of the MT techniques employed in our experiments improve MT of medical search queries. Especially the intelligent training data selection proves to be very successful for domain adaptation of MT. Certain improvements are also obtained from German compound splitting on the source language side. Translation quality, however, does not appear to correlate with the IR performance – better translation does not necessarily yield better retrieval. We discuss in detail the contribution of the individual techniques and state-of-the-art features and provide future research directions.

1. Introduction

The development of health information search and retrieval techniques is an important research topic. Indeed, it has been found that almost 70% of search engine users in the US have conducted a web search for information about a specific disease or health problem [1]. Given that much medical content is written in the English language, research to date in the medical space has predominantly focused on monolingual English retrieval. However, given the large number of non-English speaking users of the Internet and the lack of content in their native language, support for them to search and utilize these English sources is required if the value of the information available on the Internet is to be fully realized [2]. In a recent study, Lopes and Ribeiro [3] assessed the effect of translating health queries for users with different levels of English language proficiency. Their results confirmed that users with even

* Corresponding author. Tel.: +420 221914332.
E-mail address: pecina@ufal.mff.cuni.cz (P. Pecina).

0933-3657/$ – see front matter © 2014 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.artmed.2014.01.004

basic competence of English can benefit from a system which automatically retrieves English content based on a non-English query, or at least suggests English translations of the non-English queries.

Support for search of English language content by non-native English speakers is one of the major goals of the large integrated EU-funded Khresmoi project.[1] Among other goals, including joint text and image retrieval of radiodiagnostic records, the Khresmoi project aims to develop technology for transparent cross-lingual search of medical sources, for both professionals and laypeople, with the emphasis primarily on publicly available web sources. While a sophisticated search interface is being developed for the needs of medical professionals, the final application for the general public should be as simple as possible to operate and similar to the well-known interfaces of web search engines in use today with the addition of cross-lingual functionality.

The languages supported by the Khresmoi project are English (EN), Czech (CS), German (DE), and French (FR). Queries come from Czech, German, and French and are machine-translated to English. This reflects the real availability of data, which is predominantly available in English, and query translation needs of non-native speakers of English. Our focus in this paper is on the machine translation (MT) part of the cross-lingual search and retrieval task, while using a standard information retrieval (IR) technique for the search and retrieval part, in order to pinpoint contributions and problems with using MT for query translation from the three languages selected (Czech, German, and French) into English and its influence on the resulting quality of retrieved sets of documents.

Our MT system is based on Moses [4], a state-of-the-art statistical MT system. The IR experiments are performed using the Lucene search engine[2] on the CLEF eHealth 2013 dataset for the languages specified above, directed towards retrieving English documents only. Since MT is only an intermediate component of the whole system pipeline, we proceed in two steps. We first independently tune MT to produce the best possible translations of queries (Section 2) and then use various techniques to modify and expand the translated queries for improved IR performance (Section 3). The methods applied in Section 2 include: in-domain training and tuning, intelligent training data selection, optimization of phrase table configuration, exploiting synonyms to construct translation variants, and decompounding (splitting) of complex German words on the source language side, which normally appear as unknown words. For evaluation of translation quality itself, we use BLEU – the de facto standard automatic evaluation metric [5], which compares MT output against manual reference translation and accounts both for adequacy and fluency (word order) of the machine translation. We also report inverse position-independent word error rate [6], called PER, another automatic evaluation metric which compares words in the MT output and the reference translation but without taking the word order into account and thus might be better suited to application of MT in IR, where word order is often ignored. In selected experiments, the automatic evaluation is supplemented by manual assessment of the results performed by medical professionals.

The results of our MT for experiments for queries show that we are able to outperform results of Google Translate, the best freely available MT service on the web. We also find that using synonyms to enrich training data with translation variants does not improve the MT performance; however, decompounding of complex German words slightly improves the translation, at least according to BLEU. In Section 3, we evaluate query translation in a cross-lingual IR setting using standard methods on the CLEF eHealth 2013 Task 3 test collection. Here, despite achieving superior performance on the query MT task, as described in Section 2, we do not outperform the retrieval results obtained by using queries translated by Google Translate. In the last section, we perform a summary analysis of the overall results, the results of the individual techniques for improving MT performance and their integration into an IR system, and give suggestions for further work.

2. Machine translation for medical queries

In this section, we describe the application of phrase-based statistical machine translation (SMT) to the translation of medical queries with the goal of producing accurate and fluent translations. This task differs from typical MT applications in two aspects: the *domain* and the *genre* of the input text. The domain, which reflects what the text is about, is very specific, characterized by a large and specialized vocabulary which does not occur in general texts. The genre, which indicates the general style, is also very distinctive. The input text is generally not in the traditional form of complete and coherent sentences, but rather in a form of short sequences of more or less independent terms. Such a situation requires application of special techniques to adapt the SMT system, including training data selection, model configuration, and parameter optimization. We also apply some standard additional methods to improve SMT quality in this task, including morphological normalization of the input text, splitting of complex compounds in German input, and exploitation of synonyms obtained from in-domain lexicons and dictionaries.

This section continues with a brief introduction to SMT and an overview of related research, followed by a detailed description of the data and the translation system used in this work. We then present details of the MT experiments carried out with details of results and a detailed analysis of our findings.

2.1. State-of-the-art and related work

In this section, we describe basic principles of phrase-based SMT (the most widely used paradigm in SMT) and review other related works to provide a complete background for our experiments.

2.1.1. Phrase-based statistical machine translation

In phrase-based SMT (e.g., the Moses system [4]), an input sentence is split into phrases (sequences of consecutive words) that are translated one-by-one and eventually reordered to produce the output translation. As there are typically many ways to split a sentence into phrases and many possibilities for translation and reordering, the system searches for the best translation variant $\hat{\mathbf{e}}$ by maximizing the probability of the target sentence $\mathbf{e}$ given source sentence $\mathbf{f}$ in a log-linear combination of feature functions h_i with associated weights λ_i:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \sum_{i=1}^{n} \lambda_i \log h_i(\mathbf{e}, \mathbf{f})$$

The computational complexity of this decoding approach is reduced by pruning the space of translation hypotheses using a heuristic beam-search algorithm [7] that explores the space represented as a graph by expanding the most promising nodes only. The feature functions include predictions of the *phrase translation model*, which captures probabilistic relations of source phrases to target phrases, thus ensuring that the individual phrases correspond to each other, the *target language model*, which estimates the fluency of the output sentence, the *reordering model* to capture different phrase order in the two languages, and *word penalty* to penalize translations that are too long or too short.

The phrase translation model and reordering model are trained using probabilistic word alignment [8] in parallel (i.e., bilingual

[1] http://www.khresmoi.eu/.
[2] http://lucene.apache.org/.