Multilingual Test Sets for Machine Translation of Search Queries for Cross-Lingual Information Retrieval in the Medical Domain

Zdeňka Urešová, Ondřej Dušek, Jan Hajič, Pavel Pecina

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 11800 Prague 1, Czech Republic
{uresova,odusek,hajic,pecina}@ufal.mff.cuni.cz

Abstract

This paper presents development and test sets for machine translation of search queries in cross-lingual information retrieval in the medical domain. The data consists of the total of 1,508 real user queries in English translated to Czech, German, and French. We describe the translation and review process involving medical professionals and present a baseline experiment where our data sets are used for tuning and evaluation of a machine translation system.

Keywords: multilinguality; information retrieval; machine translation

1. Introduction

A number of problems being tackled within the Khresmoi project[1] required previously unavailable data resources. This was also the case of cross-lingual information retrieval (CLIR), which involves machine translation of short, Google-style search queries. Translating search queries is an inherently difficult task for current statistical machine translation (SMT) systems due to various reasons, the main being that search queries are typically very short and lack context, which helps substantially in current state-of-the-art phrase-based SMT systems. The medical domain adds another difficulty due to scarcity of large in-domain language resources, especially parallel texts. On the other hand, the existence of medical ontologies, such as MeSH (Rogers, 1963) and UMLS (U.S. National Library of Medicine, 2009), represents an opportunity for interesting experiments using these resources.

However, in order to evaluate such experiments, a representative evaluation set is needed that reflects both the medical domain and the genre of the texts to be translated, i.e., short search queries. In this paper, we describe the creation of such a resource – a multilingual evaluation data set of 1,508 search queries, manually translated from English to German, French, and Czech and thoroughly reviewed. The data set is now available through the LINDAT/Clarin repository and has already been used in a large-scale CLIR experiment described in Pecina et al. (2014) as well as for the WMT 2014 Medical Translation Task[2].

We first give a brief overview of related work in domain and genre adaptation for SMT (Section 2.) and then describe the process of creating our multilingual data sets (Section 3.). We also include a report of a baseline SMT experiment using our data sets in Section 4.

[1] http://www.khresmoi.eu
[2] http://www.statmt.org/wmt14/medical-task

2. Related work

2.1. Statistical machine translation

In phrase-based SMT, e.g., the Moses system (Koehn et al., 2007), an input sentence is split into phrases 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 ways of translation and reordering, the system searches for the best translation variant by maximizing the probability of the target sentence given the source sentence in a log-linear combination of feature functions, each of them being associated with one weight parameter.

The phrase translation model and reordering model are trained using probabilistic word alignment on bilingual pairs of sentences. The target language model is trained on (typically) larger amounts of monolingual data. Feature weights influence translation quality and are usually optimized using Minimum Error Rate Training, which minimizes a given error measure (e.g., BLEU or PER, position-independent word error rate, see Section 4.).

2.2. Domain and genre adaptation

In the case of Khresmoi, it is obvious that some domain and genre adaptation is necessary; while the domain part is obvious (medicine), by genre we mean the type of texts to be translated – in our case short, Google-like queries posted by both medical professionals and general public alike. We will first introduce some work related to these two adaptation needs.

For domain adaptation, quantity, and to a certain extent also the quality of the in-domain data is essential. However, experiments have to be performed in order to find out which combination or incorporation technique works best: sometimes, using just the in-domain data is the best option, but in most cases, a combination of in-domain and (large) general domain data gives the best results. One of the most successful techniques, the benefits of which we can confirm, is *pseudo in-domain data* acquisition and selection, described, e.g., by Mansour et al. (2011).

For genre, similar techniques can be used as in the case of domain adaptation, but typically much less data is available, especially within the domain of interest. Moreover, while domain adaptation concentrates on lexical coverage, genre differences are mostly in the structure of the texts in question, such as grammar, length, and use of punctuation. As the most relevant related work, we can mention Nikoulina et al. (2012), who worked on a similar problem of SMT adaptation to short queries in the CLEF 2009 task.

3. The data sets

3.1. Data source

The source side of the presented data set consists of randomly selected 749 real English queries asked by the general public through the Health on the Net Foundation website[3], and another 759 queries asked in English by healthcare professionals in the Trip database[4] (Meats et al., 2007). All the queries have been manually cleaned of random artifacts (e.g., texts which could not be interpreted as search queries or contained nonsensical strings such as *asdfghj* etc.). Spelling errors were preserved, but a correction has been added if the true meaning could be identified unambiguously. All the queries have been translated into German, French, and Czech. It was decided at the beginning that the resulting data set should be of a maximum quality; an appropriate funding was reserved for this purpose. It was clear that the process will need several rounds of translations and revisions by native speakers as well as medical experts.

3.2. Manual translation

The manual translation of the queries was performed in several rounds: the initial translation and up to three rounds of reviews and corrections.

The initial translations were performed by native speakers of Czech, French, and German, which were not medical experts. They were asked to provide spelling corrections for the source terms (if required), translation into their language, and possibly additional comments. In this first round, no specific instructions have been given to the translators but one: to provide a translation, not an interpretation, e.g., avoid adding explanatory comments in parentheses. The translators were also asked to perform two additional tasks:

- filter out queries in other languages than English,
- correct spelling (but not grammar) for queries which appeared to be in English but had spelling errors.

In the second round, reviewers (medical experts) were asked to mark and correct questionable translation and provide further comments if needed. They were given more specific instructions, based on our own internal review of the first translations:

- preserve the original (non-)syntax: translate as a phrase if the query appears to have syntax, otherwise translate the words one by one, not introducing any grammatical structure; e.g., *colon cancer* should be translated as *rakovina tlustého střeva/Dickdarmkrebs/cancer du côlon* (noun phrase), but *pain cancer* should result in *bolest rakovina/Schmerz Krebs/douleur cancer* (no syntax),
- translate abbreviations "naturally": keep the English original if it is used in the target language as well (e.g., *EEG, CRP*), use target language conventions for the meaning of the source abbreviation, i.e., use abbreviation in the target language if the abbreviation is commonly used, such as *JIP/ITS/USI* for *ICU* (meaning *Intensive Care Unit*), but use full text if that is the norm, e.g., *ultrazvukové vyšetření v reálném čase/Echtzeit-Ultraschall/ultasons en temps réel* for *RTU* (meaning *Real-Time Ultrasonography*),
- review the correction of the English original, too: the spelling corrections were sometimes unnecessary or wrong, thus the reviewers were asked to post-correct them, marking such cases clearly and possibly retranslating the query if the correction had changed its meaning (e.g., *bleeding diathesis* corrected to *bleeding diasthesis* and post-corrected back to *bleeding diathesis*),
- operator treatment: some queries contained logical query operators (*AND*, *OR*, possibly lowercased); these should have been identified (distinguished from conjunctions in their usual meaning) and left intact, as in *žíravý AND stent/caustique AND stent/kaustisch AND Stent* for *caustic AND stent*.

After the review round, an adjudication process followed: disagreement between the translator and the reviewer was identified and outright translation errors found by the reviewers were corrected directly. Disputed entries have been distributed to a different set of medical professionals for an additional review.

Finally, the remaining discrepancies were resolved by a final round of reviews performed by an independent person, taking all of the translation and revision information collected so far into consideration.

3.3. Statistics

We have split the translated data sets into two sections for development and testing purposes, respectively. The queries entered by general public and healthcare professionals (cf. Section 3.1.) are distributed almost evenly in both sections. The overall statistics of the data sets, including word counts in the individual languages, are given in Table 1.

4. Baseline experiments

We performed a basic SMT experiment to evaluate the effect of our data sets on in-domain translation, comparing the performance of an SMT system tuned on general-domain data to a system tuned using the development section of our query data sets.

4.1. Machine translation system

We used the Moses SMT toolkit (Koehn et al., 2007) in our setup, trained on plain tokenized texts with no additional

[3] http://www.hon.ch

[4] http://www.tripdatabase.com

section	queries	public / professionals	Czech	German	French	English	len_{EN}
development	508	249 / 259	1,128	1,041	1,335	1,084	2.13
test	1,000	500 / 500	2,121	1,951	2,490	2,067	2.07

Table 1: Statistics of our development and test data sets – number of queries included and their sources, total number of tokens per language, and average number of tokens in the English originals (len_{EN}).

development set	Czech–English		German–English		French–English	
	BLEU	1-PER	BLEU	1-PER	BLEU	1-PER
general	26.59±4.42	55.25±3.38	23.03±3.87	54.76±3.52	32.67±5.17	65.73±3.23
query	35.73±5.60	66.21±2.18	29.50±4.92	60.40±2.51	37.84±5.32	71.78±2.33

Table 2: BLEU and 1-PER scores of the baseline systems tested on medical queries and tuned on development sets of different domains, including 95% confidence intervals estimated using (plain) bootstrap resampling (Koehn, 2004).

factors. Our experiments involve three language pairs, with Czech, German, and French as the source and English as the target language. The parallel and monolingual training data came from a general domain (news texts, legislation, web crawl etc.) in all our setups, consisting of 10 million parallel and 30 million monolingual sentences for each language pair.

The parameters of the SMT systems have been tuned using Minimum Error Rate Training (Och, 2003) towards BLEU score (Papineni et al., 2002) on two different development data sets. We used the WMT 2012 translation task data set (Callison-Burch et al., 2012), consisting of 3,003 news sentences, as a general-domain development set and the development section of our query corpus as an in-domain development set.

4.2. Experiments and results

We compared the performance of the SMT systems tuned on both development sets (see Section 4.1.) on the test section of our query corpus. BLEU score and position-independent word error rate (PER) (Tillmann et al., 1997) were used for evaluation. PER is similar to word error rate, i.e., the Levenshtein distance (Levenshtein, 1966) computed on words (not characters), but it does not penalize word reordering. We include PER in the evaluation as it should better fit IR systems which typically ignore query word order. The scores of both system variants for all three language pairs are given in Table 2. PER is reported as 1-PER so that higher scores indicate better translations. Both metrics are reported as percentages.

Despite the higher variance caused by the small test set size, the differences between the two setups for all language pairs were confirmed for both metrics to be statistically significant at 95% level using paired bootstrap resampling (Koehn, 2004). The SMT system tuned on the query development set shows a remarkable improvement, considering that different weights of feature functions resulting from the tuning are the only difference between the two setups.

5. Conclusions

We have presented a new multilingual dataset (English, Czech, German and French) for the purpose of evaluation of machine translation systems within a cross-lingual information retrieval task in the medical domain, where short Google-like queries are expected. The translations from English to the other three languages have been very carefully checked by medical professionals and only confirmed after consensus has been reached.

We have also described a set of experiments that show the usability and usefulness of this dataset, which we believe is the first for this domain, genre, and the two user bases (medical professionals and general public). More in-depth experiments, including settings which gave better MT results than the public translation engines of Google Translate and Microsoft Bing, are presented in Pecina et al. (2014).

The data is now in the public domain[5] under the CC-BY-NC 3.0 license, thanks to an agreement with the original query log providers.

Acknowledgments

This work was supported by the EU FP7 project Khresmoi (contract no. 257528). The language resources presented in this work are distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth, and Sports of the Czech Republic (project LM2010013). This research was partially supported by SVV project number 260 104.

6. References

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004*

[5] http://hdl.handle.net/11858/00-097C-0000-0022-D9BF-5

Conference on Empirical Methods in Natural Language Processing, pages 388–395, Barcelona, Spain.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.

Saab Mansour, Joern Wuebker, and Hermann Ney. 2011. Combining translation and language model scoring for domain-specific data filtering. In *International Workshop on Spoken Language Translation*, pages 222–229, San Francisco, CA, USA.

Emma Meats, Jon Brassey, Carl Heneghan, and Paul Glasziou. 2007. Using the Turning Research Into Practice (TRIP) database: how do clinicians really search? *Journal of the Medical Library Association*, 95(2):156–163.

Vassilina Nikoulina, Bogomil Kovachev, Nikolaos Lagos, and Christof Monz. 2012. Adaptation of statistical machine translation model for cross-lingual information retrieval in a service context. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 109–119, Avignon, France.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA.

Pavel Pecina, Ondřej Dušek, Lorraine Goeuriot, Jan Hajič, Jaroslava Hlaváčová, Gareth Jones, Liadh Kelly, Johannes Leveling, David Mareček, Michal Novák, Martin Popel, Rudolf Rosa, Aleš Tamchyna, and Zdeňka Urešová. 2014. Machine translation for multilingual information retrieval in medical domain. *Artificial Intelligence in Medicine*, 1(Accepted for publication).

Frank Rogers. 1963. Medical subject headings. *Bulletin of the Medical Library Association*, 51:114–116.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Alex Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In *Proceedings of the Fifth European Conference on Speech Communication and Technology*, pages 2667–2670, Rhodes, Greece.

U.S. National Library of Medicine. 2009. UMLS reference manual. Metathesaurus. Bethesda, MD, USA.