Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Machine translation techniques for presentation of summaries

Deliverable number	*D4.6*
Dissemination level	*Public*
Delivery date	*April 2014*
Status	*Final*
Author(s)	*Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Michal Novák, Pavel Pecina, Rudolf Rosa, Aleš Tamchyna, Zdeňka Urešová, Daniel Zeman*

This project is supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development.

Abstract

This document reports on the Khresmoi activity to provide machine translation of document summaries. We develop systems for translation of English summaries into Czech, French, and German. The systems are trained on combination of medical-domain and general-domain data, using state-of-the-art domain-adaptation techniques (e.g., linear interpolation of medical-domain and general-domain models), and tuned and tested on a set of manually translated sentences sampled from summaries of real Khresmoi-indexed documents. The development and test data created for this purpose was made available for a shared-task on machine translation of medical texts, organized by the Charles University in Prague (CUNI) within the Ninth Workshop of Statistical Machine Translation (WMT) 2014 as a part of the Khresmoi dissemination effort. We attach a paper describing our system participating in this task accepted for publication in the proceedings of WMT 2014.

Table of Contents

List of abbreviations

CS	Czech
DE	German
EN	English
FR	French
CUNI	Charles University in Prague
MT	Machine Translation
SMT	Statistical Machine Translation

1. Introduction

The Machine Translation (MT) service is the component providing cross-lingual capability to search in biomedical documents in Khresmoi. The service allows 1) to present summaries of search results returned to the user in a chosen language and 2) to translate non-English user queries to English which is the central language used for indexing and searching in Khresmoi. While the query translation service was described in D4.5 Report on Query Expansion Techniques and research papers [1,2], the summary translation service is discussed in this document. In Section 2, we describe the data set created for the purposed of tuning and testing the MT systems. This resource was made available as development and test data for the Medical Translation Task at the Ninth Workshop of Statistical Machine Translation (WMT) 2014, described in Section 3. Our MT systems are sketched in Section 4 and in more detail described in the attached paper [3] which was accepted for publication in the proceedings of WMT 2014. Section 5 concludes this deliverable.

2. Khresmoi Summary Translation Test Data

For the purposes of development and testing our systems for summary translation, we developed the Khresmoi summary translation test data for the medical domain. This data set consists of 1500 English sentences randomly sampled from automatically generated summaries of (English) documents from the CLEF 2013 eHealth collection and found to be relevant to 50 CLEF 2013 eHealth Task 3 topics. The collection comprises of about one million HON-certified pages webcrawled by Khresmoi in 2012. The sampled sentences were manually checked and any out-of-domain and/or ungrammatical sentences were removed. The sentences were translated by medical experts into CS, FR, and DE. The translations were further reviewed.

The resulting data set is available under the terms of the Creative Commons Attribution-Noncommercial (CC-BY-NC) license, ver. 3.0 unported from the Lindat/Clarin repository: http://hdl.handle.net/11858/00-097C-0000-0023-866E-1 (version 1.1).

3. WMT 2014 Medical Translation Task

The Medical Translation Task (http://www.statmt.org/wmt14/medical-task/) has been organized as a featured shared task of the Ninth Workshop of Statistical Machine Translation (WMT) which will take place in Baltimore, MD, USA in June 2014 as an event collocated with the 52nd Annual Meeting of the Association for Computational Linguistics. WMT is a prestigious annual event organizing highly competitive shared tasks in machine translations for almost a decade.

For its internal development and testing purposes within Khresmoi, CUNI developed two test sets (Khresmoi Query Translation Test Data and Khresmoi Summary Translation Test Data) and offered them to be used and development and test data for a featured shared task within WMT 2014.

3.1 Shared task description

The goal of the tasks was to investigate the applicability of current MT techniques to the translation of domain-specific and genre-specific texts between EN↔CS, EN↔DE, and EN↔FR. The task was split into two subtasks:

1. translation of sentences from summaries of medical articles,
2. translation of queries entered by users of medical information search engines.

We invited both beginners and established research groups to participate in this task. On top of the development/test resources, we provided links to additional in-domain data and out-of domain data for training. The participants were asked to train/tune their system using the provided resources (constrained task) or any additional resource (unconstrained task).

3.2 Shared task schedule

Task announcement: December 12, 2013

Release of development test sets: January 6, 2014

Release of test sets: March 10, 2014

Submission of translations: March 14, 2014

Submission of papers: April 1, 2014

Notification of paper acceptance: April 21, 2014

Submission of camera-ready versions: April 28, 2014

Workshop: June 26–27, 2014

3.3 Shared task results

Detailed results of the shared task will be described in the workshop overview paper "Findings of the 2014 Workshop on Statistical Machine Translation" [4] published in the workshop proceedings.

4. Khresmoi summary translation system

Our MT systems for translation from EN to CS, DE, and FR are thoroughly described in the attached CUNI WMT 2014 system description paper [3] together with other systems for summary translation (from Czech, German, and French to English) and systems for query translation (both directions, from English and to English). Since we were involved in the organization of the Medical Translation Task, our primary goal was to set up strong baselines for both its subtasks (summary translation and query translation) and for all translation directions. Our systems were based on the phrase-based Moses toolkit [5] and standard methods for domain adaptation including selection of pseudo-in-domain training data and interpolation of in-domain and general-domain models optimized to minimize cross-perplexity on the development data, see the attached paper [3] for more details. Within Khresmoi, we are still actively working on the summary translation systems and will report our results within the scope of the project.

5. Conclusions

This deliverable provides an overview of the Khresmoi work on machine translation of medical document summaries. We developed a unique resource for development and testing of MT systems for the medical domain. It contains 1500 real English sentences translated into Czech, German, and French by medical professionals. This data was made available for the purposes of the shared task on medical translation organized by CUNI as a part of the Ninth Workshop of Statistical Machine Translation. We referred to the CUNI's system description paper providing details of the systems developed for summary (and query) translation within Khresmoi. The systems are still being developed and improved.

6. Appendix

Paper [3] is attached this deliverable as an appedix.

7. References

[1] Pavel Pecina, Ondřej Dušek, Lorraine Goeuriot, Jan Hajič, Jaroslava Hlaváčová, Gareth J. F. Jones, Liadh Kelly, Johannes Leveling, David Mareček, Michal Novák, Martin Popel, Rudolf Rosa, Aleš Tamchyna, Zdeňka Urešová. Adaptation of machine translation for multilingual information retrieval in the medical domain. To appear in Artificial Intelligence in Medicine, Elsevier, 2014.

[2] Zdeňka Urešová, Ondřej Dušek, Jan Hajič, Pavel Pecina. Multilingual Test Sets for Machine Translation of Search Queries for Cross-Lingual Information Retrieval in the Medical Domain. To appear in Proceedings of the Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 2014.

[3] Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Michal Novák, Pavel Pecina, Rudolf Rosa, Aleš Tamchyna, Zdeňka Urešová, Daniel Zeman. Machine Translation of Medical Texts in the Khresmoi Project. To appear in Proceedings of the ACL 2014 Ninth Workshop of Statistical Machine Translation, Baltimore, USA, 2014.

[4] Findings of the 2014 Workshop on Statistical Machine Translation. To appear in Proceedings of the ACL 2014 Ninth Workshop of Statistical Machine Translation, Baltimore, USA, 2014.

[5] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL '07). Association for Computational Linguistics, Stroudsburg, PA, USA, 177-180.