Grant Agreement Number: 257528

KHRESMOI
www.khresmoi.eu

Report on results of the WP4 second evaluation phase

Deliverable number	*D4.7*
Dissemination level	*Public*
Delivery date	*June 2014*
Status	*Final*
Author(s)	*Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Michal Novák, Pavel Pecina, Rudolf Rosa, Aleš Tamchyna, Zdeňka Urešová, Daniel Zeman*

Abstract

This document reports on the results of the second evaluation phase of WP4 focused primarily on translation of medical document summaries and partially also on translation of medical queries. The evaluation was carried out within the shared task organized by the Charles University in Prague (CUNI) and Dublin City University (DCU) as a part of the Ninth Workshop on Statistical Machine Translation (WMT) collocated with the 52nd Annual Meeting of the Association for Computational Linguistics taking place in Baltimore, Maryland, USA in June 2014. The Khresmoi translation system is described in the CUNI's system description paper published in the WMT proceedings and attached to D4.6 *Machine translation techniques for presentation of summaries*. We refer to the evaluation results described in the shared task overview paper which is attached to this deliverable as an appendix. We also report on our new and improved results of the translation of medical document summaries, which were achieved after the shared task on the same data sets.

_____

Table of Contents

List of abbreviations

CS	Czech
DE	German
EN	English
FR	French
CUNI	Charles University in Prague
DCU	Dublin City University
MT	Machine Translation
SMT	Statistical Machine Translation
WMT	Workshop on Statistical Machine Translation

1 Introduction

The machine translation (MT) effort in Khresmoi is focused on two goals: 1) translation of user search queries from Czech, German, and French into English to support cross-lingual search capability and 2) translation of summaries of retrieved documents from English to users' preferred languages (Czech, German, and French). The first development phase of the Khresmoi translation component resulted in D4.3 *Report on results of the WP4 first evaluation phase* [1] which analysed the resources acquired for training of the Khresmoi translation component and evaluated the baseline systems using test sets sampled from the EMEA corpus [2], the only in-domain parallel corpus from the medical domain available in all languages of interest to Khresmoi. At that stage, no specific distinction was made between translations of search queries and document summaries, though the two tasks are quite different: Unlike document summaries, the queries do not form sentences.

In the second phase of the project, we aimed at tuning the translation systems for the two tasks separately. First, we focused on query translation and created our own task-specific data for development and testing purposes [3]. The results of this effort were described in D4.5 *Report on query expansion techniques* [4] and published as a journal paper [5].

Further, we focused on the summary translation and created another test set developed specifically for this task. As a part of Khresmoi dissemination activities, we organized a shared task on medical text translation at the Ninth Workshop of Statistical Machine Translation 2014 and provided the translation test sets (both for query translation and summary translation) for development and testing purposes of the task. Our team (CUNI) provided baseline systems for all the language pairs in both subtasks. The systems were described in paper [6] presented at the workshop and attached to the D4.6 *Machine translation techniques for presentation of summaries*. The main improvements of the systems with respect of the previous versions are twofold: First, we were able to acquire large training data and improved the method to select the pseudo-relevant subsets. Second, we used the created development data sets for tuning of the systems for this very specific domain. Details can be found in [6].

In this document, we first report in Section 2 on the results presented in the shared-task overview paper [7] attached to this deliverable as an appendix (Section 5). In Section 3, we describe our new results of the translation of medical document summaries which improved those presented at the workshop [6, 7]. Section 4 concludes the document.

2 WMT 2014 Medical Translation Task

The medical translation task was organized as a part of the Ninth Workshop of Statistical Machine Translation (WMT 2014), which organizes similar tasks (usually focused on the news domain) annually since 2005. The goal was to investigate the applicability of MT to translate domain-specific and genre-specific texts between EN↔CS, EN↔DE, and EN↔FR. Following the MT goals in Khresmoi, the shared task was split into two subtasks: 1) translation of sentences from summaries of medical articles, 2) translation of queries entered by users of medical search engines. We provided test sets for development and testing purposes as well as links to in-domain and out-of domain data for training. The participants were asked to train/tune their systems using the provided resources (constrained task) or any additional resource (unconstrained task) and submit translations of the test data within 5 days. In Khresmoi, we are interested in the translation of queries from CS, DE, and FR into EN and translating the summaries from EN to CS, DE, and FR. In the shared task, we allowed translation in all directions in both the subtasks.

The results of the WMT 2014 Medical Translation Task are described in Section 5 of the task overview paper [8] published in the proceedings of WMT 2014 and attached to this document. The

evaluation was performed using automatic evaluation measures, such as BLEU [10], TER [11], PER [12], and CDER [13]. There were the total of eight teams participating in the task and their results varied depending on the subtask and translation direction. Most of the systems were based on the teams' systems applied to the standard translation task (news domain) and trained on the data provided or its subsets.

In the query translation subtask, with one exception, our systems (CUNI) performed the best according to the automatic measures ignoring letter casing, which is natural in IR. The exception was the FR→EN translation direction, where the best result was achieved by the team of the Dublin City University (members that are not involved in the Khresmoi project). The difference, however, was not statistically significant.

In the summary translation subtask, the best overall results were achieved by the team of the University of Edinburgh (UEDIN), which won for DE→EN, EN→CS, and EN→FR, followed by the team of the University of Macao (UN-DA), which performed on par with UEDIN in all the other translation directions. All these systems outperformed the baselines.

3   Khresmoi summary translation improved

As we have already stated, CUNI was organizing the WMT 2014 Medical Translation Task and participated mainly to provide baseline results for both the subtasks and all translation directions. In the constrained task, the CUNI baseline systems were based on the standard Moses phrase-based toolkit [9] and linear interpolation of two kinds of language models and phrase tables: one trained on data selected from the in-domain resources and the other one selected from the out-of-domain resources. The selection was based on measuring cross-perplexity of each sentence (sentence pair for parallel data) given language models trained on medical domain texts and language models trained on non-medical-domain text [14]. For training the language and translation models, we only exploited sentences (sentence pairs for parallel data) which had lower perplexity given the in-domain model compared to the perplexity given the out-of-domain model. Such data can be seen as "medical-like" and thus is more suitable for training medical-domain systems. The constrained and unconstrained systems differ in the training data only. For the constrained systems, we performed the data selection from all allowed training data. The unconstrained ones take advantage of additional web-crawled monolingual data used to train the language models, and additional parallel non-medical data. Details are given in the system description paper [6].

Using this approach reduced the size of training data quite significantly and this was probably the reason why our systems did not perform very well in the summary translation subtask at WMT. We discarded a lot of training data because it was not very similar to the medical-domain texts but (as we can see from the shared task results) it still could help to improve translation. In our next experiments, we learnt from our WMT participation and decided not to ignore such data, but rather used it to train additional models (language models and phrase tables), which we added to the interpolation. The final language models and phrases tables are thus built as linear interpolation of four components: medical-like data selected from the in-domain resources (1) plus its complement, i.e. all remaining data (2), medical-like data selected from the out-of-domain resources (3) and its complement again (4). The coefficients of the linear interpolation were tuned on the official development sets for each language pair. As show in the following table, these methods improved our results and the resulting systems are now deployed in the translation component of the Khresmoi systems (Khresmoi for Everyone, Khresmoi Professional). We should also admit that in most cases the improvements are not better than the best systems participating in the WMT 2014 medical shared task.

System	Translation	BLEU	1-PER	1-TER	1-CDER

WMT	CS → EN	33.18±1.15	51.48±1.15	66.00±1.03	55.30±0.96
NEW	**CS→ EN**	**35.45±1.19**	**53.41±1.14**	**67.10±1.08**	**57.27±0.99**
WMT	DE→ EN	33.14±1.19	50.98±1.06	65.88±1.04	54.74±0.94
NEW	**DE→ EN**	**33.45±1.16**	**51.63±1.07**	**66.40±1.04**	**55.12±0.92**
WMT	FR→ EN	36.84±1.17	54.56±1.13	66.43±1.07	59.14±0.90
NEW	**FR→ EN**	**37.76±1.23**	**55.30±1.18**	**66.84±1.10**	**59.95±0.9**

Table 1: Results of the new summary translation systems (NEW-CUNI) compared with our best submissions to the WMT 2014 Medical Translation Task (WMT-CUNI). All scores are reported as percentages.

4 Conclusion

In this report, we provided an overview of the results of the WP4 second evaluation phase. The evaluation was conducted within a shared task organized as a part of the Ninth Workshop on Statistical Machine Translation collocated with the 52nd Annual Meeting of the Association for Computational Linguistics in Baltimore, Maryland, USA in June 2014. The shared task focused on the translation of medical texts and was organized by the Khresmoi team from the Charles University in Prague and Dublin City University. The Charles University team participated in the shared task and provided (strong) baseline systems for all the language pairs in both subtasks (summary translation and query translation). In this report, we reviewed the results of the shared task and reported on improvements of the Khresmoi translator achieved after the shared task.

5 Appendix

Paper [8] is attached to this deliverable as an appendix. Its Section 5 presents the results of the medical translation shared task of the WMT 2014.

6 References

[1] Pavel Pecina, Jakub Bystroň, Jan Hajič, Jaroslava Hlaváčová, Zdeňka Urešová. D4.3 Report on results of the WP4 first evaluation phase. *Khresmoi project deliverable*, 2012.

[2] Jörg Tiedemann (2009). News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Recent Advances in Natural Language.* Processing V, volume 309 of Current Issues in Linguistic Theory, pages 227–248. John Benjamins, Amsterdam & Philadelphia.

[3] 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. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, 2014.

[4] Lorraine Goeuriot, Liadh Kelly, Johannes Leveling, Hugo Mougard, Gareth Jones, Pavel Pecina, Jan Hajič. D4.5 Report on query expansion techniques. *Khresmoi project deliverable*, 2013.

[5] 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.

[6] Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Michal Novák, Pavel Pecina, Rudolf Rosa, Aleš Tamchyna, Zdeňka Urešová, and Daniel Zeman. Machine Translation of Medical Texts in the Khresmoi Project. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 221-228, Baltimore, USA.

[7] Ondřej Dušek, Jan Hajič, Jaroslava, Hlaváčová, Michal Novák, Pavel Pecina, Rudolf Rosa, Aleš Tamchyna, Zdeňka Urešová, Daniel Zeman. D4.6 Machine translation techniques for presentation of summaries. *Khresmoi project deliverable*. 2014.

[8] Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Johannes Leveling, Philipp Koehn, Christof Monz, Pavel Pecina, Matt Post, Hervé Saint-Amand, Radu Soricut, Lucia Specia and Aleš Tamchyna. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12-58, Baltimore, USA.

[9] 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)*, pages 177-180, Association for Computational Linguistics, Stroudsburg, PA, USA.

[10] 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, Pennsylvania, USA.

[11] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA, USA.

[12] Christoph Tillmann, Stephan Vogel, Hermann Ney, A. 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.

[13] Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. Cder: Efficient mt evaluation using block movements. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 241–248, Trento, Italy.

[14] Robert C. Moore, William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220– 224, Uppsala, Sweden.