KCONNECT

www.kconnect.eu

Adaptation to Hungarian, Swedish, and Spanish

Deliverable number	*D1.4*
Dissemination level	*Public*
Delivery date	*31 January 2016*
Status	*Final*
Author(s)	*Jindřich Libovický, Aleš Tamchyna, Pavel Pecina*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644753 (KConnect)

Executive Summary

This report presents the Khresmoi Translator adapted to new language pairs. Prior KConnect, the Khresmoi Translator supported translation for English–Czech, English–German, and English–French. While the direction into English was specifically tuned for translation of user search queries, the opposite direction was aimed at translating full sentences. The new language pairs include English–Hungarian, English–Spanish, and English–Swedish and the systems are tuned for query translation as well as sentence translation. The English–Spanish pair replaced the originally planned English–Polish because of the current consortium requirements. Support for English–Polish translation will be added later during the course of the project. This report describes the newly developed translation systems, it provides technical details of training and tuning and present results of translation quality evaluation.

Table of Contents

1 Introduction 4
2 Data 4
2.1 Parallel data 4
2.2 Monolingual data 6
2.3 Development and test data 6
3 System description 7
3.1 SMT components 7
3.2 Khresmoi Translator 9
4 Evaluation 9
5 Conclusion 10
6 References 10

List of Abbreviations

(S)MT	(Statistical) Machine Translation
TM	Translation Model
LM	Language Model

1 Introduction

Khresmoi Translator is a key component of Khresmoi tools and services. It provides capability for cross-lingual search and access to medical documents via two functionalities: It allows 1) translation of user search queries from their preferred language to the language of documents and 2) translation of document sentences into the preferred language of the users.

Within the Khresmoi project, the translation system was developed for three language pairs: English–Czech, English–German, and English–French. English, as the central language of documents in Khresmoi, was used for indexing and queries in English were processed directly. However, users were also allowed to formulate their queries in Czech, German, or French. Such queries were translated into English and further processed by the retrieval system. The retrieved documents were presented back to the users in a form of a list containing document titles and summaries, eventually translated into the users' preferred languages. The Khresmoi translator supported translation of queries from Czech, German, and French into English and translation of summary sentences (and titles) from English to Czech, German, and French.

The system was built on the current state-of-the-art technologies, especially Moses [2] – an open-source toolkit for Statistical Machine Translation (SMT) [2] and MT Monkey, an open-source framework for deploying SMT as a web service [3]. The system and its evaluation was described in Khresmoi deliverables D4.3 [6], D4.6 [7], and D4.7 [8].

One of the main goals of KConnect, is to develop procedures for easy adaptation of the Khresmoi tools and services to new languages. In D1.2 [1] we described Eman Lite, an easy-to-use tool for training new MT systems for Khresmoi Translator. In this report we present such systems for English–Hungarian, English–Spanish, and English–Swedish. The systems are tuned for both query translation and sentence translation. Originally, we planned to support English–Polish instead of English–Spanish, but the current requirements of industrial partners in the consortium changed the preferences. Support for English–Polish translation will be added later during the course of the project.

In this report, we describe the newly developed translation systems, including technical details and data used for training, tuning and testing. Finally, we also present results of empirical evaluation of translation quality of all the systems.

2 Data

The data for training the new systems were extracted from the sources collected in WP1 during M1-M6 of the project. The acquired resources are described in D1.2 [1] and include: parallel data for training the translation models and monolingual data for training the language modes. Both types of data are either in-domain (medical) or out-of-domain (general), the following sections provide a brief summary.

2.1 Parallel data

The parallel data was compiled from various public sources and sources that the consortium members have a licence for. While training an SMT system, parallel data are used for extracting a 'phrase table' which is later used for retrieval of hypotheses how can be parts of a sentence translated.

The most important source of parallel data in the medical domain is the EMEA corpus which was made out of the documents published by the European Medicines Agency and is therefore available in all official languages of the EU. Use of this data is entirely free. The second biggest source of medical-domain data is the UMLS which is a multilingual metathesaurus of medical terminology maintained by the U.S. National Medical Library. The Spanish part of the database is substantially bigger than for the other three languages. In addition to that, we were able to collect Spanish parallel data from the U.S. official medical webs and web pages of international institutions like WHO. For Swedish and

Hungarian, the consortium was able collect bilingual abstracts of medical papers. Minor sources of data were titles of Wikipedia articles and subtitles of medical TV shows collected from publically available sources on the Internet. The in-domain corpora are combined with general-domains parallel data. From these, the most important are corpora collected from the proceeding of the European parliament and European legislation which is in all official EU languages. Another big corpus we use is collected from the movie subtitles publically available on the Internet. There were also some language specific sources – the Hunglish corpus by Budapest Institute of Technology and Spanish UN corpus made out of proceedings of UN. Size of the training resources is summarized in Tables 1 and 2 (in terms of millions of tokens).

	English–Spanish	English–Hungarian	English–Swedish
EMEA	14.8	14.3	14.8
UMLS	56.6	3.8	1.4
Health Web Crawl	3.7	–	–
Subtitles	1.7	–	1.0
MeSH	0.12	1.7	0.56
SnomedCT	–	–	5.9
HU Abstracts	–	0.5	–
ECDC	0.04	0.04	0.04
WIKI titles - medical	0.01	0.002	0.005
Total	**76**	**18**	**23**

Table 1: Statistics of the parallel data in medical domain (millions of tokens on the English side)

	English–Spanish	English–Hungarian	English–Swedish
WIKI titles - general	1.7	0.2	0.8
UN - general	241	–	–
EU Bookshop	14.2	12.2	51.9
Europarl	56.4	17.0	52.5
JRC-Acquis	18.4	12.6	18.0
Opensubtitles	912	510	206
News Commentary	4.7	–	–
Hunglish	–	29	–
DGT-Acquis	173	142	148
Total	**1420**	**652**	**402**

Table 2: Statistics of the parallel data in general domain (millions of tokens on the English side)

2.2 Monolingual data

The second major component of an SMT system is a language model. The purpose of having a language model in the system is to be able to build a coherent sentence from the candidates' phrases provided by the phrase table. The language model operates in the target language only. In addition to target side of the parallel data, we use more data because the monolingual data are usually easier to obtain.

For Polish and Hungarian we collected additional monolingual by crawling websites for medical professionals. For Swedish, we downloaded a collection of medical texts from Läkartidningen. We add also texts from Wikipedia for all the languages. In case of English, we were able to collect enough in-domain data that there is no need to interpolate them with any out-of-domain data.

	Spanish	Hungarian	Swedish	English
SV Med. Journals	–	–	20	–
Precognox crawl	–	12	–	–
CLEF 2014	–	–	–	685
Trip	–	–	–	2120
BioMedCentral	–	–	–	367
Cochrane	–	–	–	67
Drugbank	–	–	–	0.89
FMA	–	–	–	0.88
Genia	–	–	–	0.56
HON	–	–	–	1158
Wikipedia med.	1.7	0.6	0.8	0.5
Total	**2**	**13**	**21**	**4280**

Table 3: Statistics of the monolingual data in general domain (millions of tokens)

source	Spanish	Hungarian	Swedish
Wikipedia	474	98	158

Table 4: Statistics of the monolingual data from Wikipedia (millions of tokens)

2.3 Development and test data

For the purposes of tuning the parameters of the MT systems and evaluating the final translation quality, we had to acquire realistic samples of data the systems will be translating during application time (user search queries, document summary sentences, both from the medical domain). Such data needs to be unseen by the systems (i.e., not involved in the training process) and provided with accurate translations (i.e., manual).

Within the Khresmoi project, such data was acquired for the three original language pairs. The query data were sampled from query logs provided by medical search engines and the summary sentences were extracted from English medical articles. All were then translated into Czech, German, and French. The entire process of acquisition and translation of the data is described in [4] and [5]. Within KConnect, we extended the data sets by adding parallel translations into the new languages (Hungarian, Spanish, Swedish, and Polish). The translation was conducted by a professional translation company following the instructions developed with the Khresmoi project and described in [4] and [5]. The translators were medical professionals, native speakers of the target languages or highly skilled in translating medical texts. The translations were proofread by native speakers (not necessarily medical professionals) to ensure fluency and absence of grammatical and or technical errors. Random checks were performed by KConnect partners' representatives before finalizing the data sets and using them for tuning and testing the systems. The data sets will be released to the public under the CC-BY-NC license. A brief statistics of the data sets are given in Tables 5 and 6.

set	queries	English	Czech	German	French	Hungarian	Spanish	Swedish
dev	5,08	1,084	1,128	1,041	1,335	1,022	1,220	967
test	1,000	2,067	2,121	1,951	2,490	1,927	2,312	1,847

Table 5: Statistics of the query dataset (number of queries and words per language).

set	sentences	English	Czech	German	French	Hungarian	Spanish	Swedish
dev	500	10,576	9,299	10,327	12,901	10,143	11,372	9,543
test	1,000	21,996	19,376	21,648	27,416	19,739	24,096	19,974

Table 6: Statistics of the summary dataset (number of sentences and words per language).

3 System description

3.1 SMT components

The systems trained for the new language pairs are similar to those we developed for the Khresmoi translator. They are based on the domain-adaptation techniques that we described in [4]. The main idea is based on splitting the training data into two subsets: in-domain and out-of-domain (also called general domain) and training two types of models: in-domain models trained on the in-domain data, and general-domain models trained on the out-of-domain data. These models are then interpolated in a linear fashion using parameters estimated on the development data (to minimize its perplexity) and used directly by the Moses decoder. This procedure is done with both the translation models and language models. This idea is not novel and has been shown several times to outperform the straightforward approach when all the data is concatenated into a single data set (no interpolation).

The information whether a data set is in-domain or out-of-domain is usually obtained as explicit (based on the source, e.g., the UMLS metathesaurus or the EMEA corpus are medical-domain) or estimated automatically, using e.g. a binary classifier (classes: medical vs non-medical).

We trained two English language models for data selection. The first, in-domain model was trained on 1 million sentence pairs, extracted randomly from HON and Cochrane corpora. The out-of-domain model was trained on a 1 million sentence random sample from Gigaword. LM vocabulary was restricted

– we excluded singletons and words containing non-latin characters. LM perplexity difference was used as the criterion for data selection.

We divided each parallel corpus into two parts based on whether the perplexity difference was negative (in-domain data) or positive (general domain). In each part, we estimated the word alignment using fast_align and extracted a standard Moses phrase table. For each translation direction, the phrase tables were linearly interpolated using TMcombine; interpolation weights were tuned to minimize the perplexity of the development set. The statistics of parallel data used for training are given in Table 7.

Systems which translate into English utilize English monolingual data in LM training. Due to the large size of the English data, we selected 50 million sentences with the lowest perplexity difference and trained a LM on this subset. For systems that translate out of English, we use all the available monolingual data. The statistics of monolingual training data are given in Table 8.

In all translation directions, the LM trained on monolingual data was linearly interpolated with LMs trained on the target side of the parallel data (positive and negative); the final LM for each direction is thus an interpolation of three LMs. We used the SRILM toolkit for the interpolation. All LMs are 5-gram models with modified Kneser-Ney smoothing, estimated using KenLM.

Finally, relative weights of the individual models in the final MT system were optimized using MERT towards BLEU.

language pair (L1-L2)	**domain**	**sentences**	**L1 tokens**	**L2 tokens**
English–Hungarian	in	3,108,230	26,903,806	23,590,650
English–Hungarian	out	1,407,617	21,545,274	17,590,953
English–Spanish	in	13,968,285	133,504,199	156,566,133
English–Spanish	out	6,347,827	183,27,2621	201,196,490
English–Swedish	in	1,640,200	16,809,804	14,726,100
English–Swedish	out	221,779	1,007,932	874,560

Table 7: Parallel training data statistics.

language	**domain**	**sentences**	**tokens**
Hungarian	mix	6,509,907	116,532,403
Spanish	mix	18,253,966	475,267,482
Swedish	mix	11,407,713	178,793,607
English	in	50,000,000	1,120,091,680

Table 8: Monolingual training data statistics.

3.2 Khresmoi Translator

The new SMT systems have been deployed into the Khresmoi translator and extend the set of supported services by twelve in total. We newly support translation for three new language pairs, both from English and to English and have separate services for query translation and sentence translation.

The web utilises the MTMonkey [3]. Figure 3 shows an overview of the entire architecture.

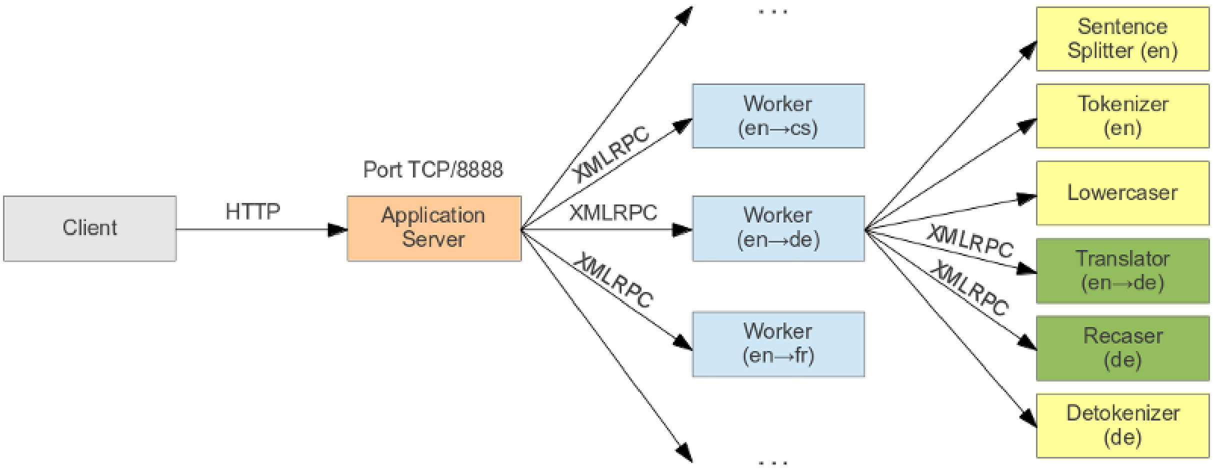

Figure 1: MTMonkey architecture overview.

The translation service is implemented using the standard HTTP protocol, JSON format, and the REST principles. A client initiates a request to the server (text to be translated); server distributes the request to a worker handling translation for the given translation pair; the worker splits the input text into sentences, performs tokenization, mapping to lowercase letters (lowercasing), translation, reconstruction of letter cases (recasing), and removal of unnecessary white spaces. Finally, the worker returns the translated text to the server which sends it back to the client.

The service is available at this URL: http://cuni1-khresmoi.ms.mff.cuni.cz:8080/khresmoi and can be tested, e.g., using the `curl` tool:

```
curl -i -H "Content-Type: application/json" -X POST -d
'{"action":"translate", "sourceLang":"en", "targetLang":"de", "text":
"This is a medical translation test." }'
http://cuni1-khresmoi.ms.mff.cuni.cz:8080/khresmoi
```

A fully working demo is available here: http://quest.ms.mff.cuni.cz/khresmoi/demo/

And an up-to-date documentation of the API here: https://github.com/ufal/mtmonkey

4 Evaluation

The evaluation has been carried out on the test sets described in Section 3. The results are presented in Table 9 in terms of BLEU scores. BLEU [9] is a standard metrics for translation quality evaluation which scores how machine the translated text matches a reference translations. The achieved scores cannot be meaningfully compared across different language pairs and different test datasets (mainly due to the difference in comprehension of the languages, existence of other translations which may be acceptable but differ from the reference translation) but with the exception of English→Hungarian translation of queries and summary sentences, the scores are comparable to those achieved by the original systems in Khresmoi Translator for English→Czech, English→German, and English→French [7,8]. The main reason why the English→Hungarian systems achieve much lower scores than the other systems are probably the morphological and syntactical properties of Hungarian when this acts as the target language).

direction	query	summary
English→Spanish	35.25±4.89	34.81±1.12
English→Hungarian	15.64±6.79	9.92±0.73
English→Swedish	33.65±5.83	37.05±1.22
Spanish→English	43.02±5.29	51.80±1.34
Hungarian→English	38.16±5.18	21.06±1.08
Swedish→English	44.79±5.50	50.87±1.32

Table 9: Evaluation results of the new systems (BLEU scores with a 5% confidence intervals).

5 Conclusion

In this report, we presented new language pairs supported by the Khresmoi Translator. The system is newly able to translate between English–Hungarian, English–Spanish, and English–Swedish in both directions, and both modes (query translation and sentence translation). Support for English–Polish will be added later. Translation quality is mostly comparable to the original language pairs in the Khresmoi Translator. Quality of English→Hungarian translation will be further analysed. Our future work will focus on training new systems for English–Polish and improving translation quality of the current ones.

6 References

[1] Aleš Tamchyna, Jindřich Libovický, Pavel Pecina: D1.2 Toolkit and Report for Translator Adaptation to New Languages (First Version). KConnect deliverable. www.kconnect.eu. 2015.

[2] 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 ACL on Interactive Poster and Demonstration Sessions, pages 177–180, Prague, Czech Republic.

[3] Aleš Tamchyna, Ondřej Dušek, Rudolf Rosa, and Pavel Pecina. MTMonkey: A Scalable Infrastructure for a Machine Translation Web Service. In The Prague Bulletin of Mathematical Linguistics, No. 100, pp. 31–40, 2013.

[4] Zdeňka Urešová, Ondřej Dušek, Jan Hajič, and 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, pp. 3244-3247, Reykjavik, Iceland, 2014.

[5] Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Johannes Leveling, Philipp Koehn, Christof Monz, Pavel Pecina, Matt Post, Herve 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, pp. 12-58, Baltimore, USA, 2014.

[6] Pavel Pecina, Jakub Bystroň, Jan Hajič, Jaroslava Hlaváčová, Zdeňka Urešová. D4.3 - Report on results of the WP4 first evaluation phase. Khresmoi Deliverable. http://www.khresmoi.eu. 2013.

[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 Deliverable. http://www.khresmoi.eu. 2014.

[8] Ondřej Dušek, Jan Hajič, Jaroslava, Hlaváčová, Michal Novák, Pavel Pecina, Rudolf Rosa, Aleš Tamchyna, Zdeňka Urešová, Daniel Zeman. D4.7 - Report on results of the WP4 second evaluation phase. Khresmoi Deliverable. http://www.khresmoi.eu. 2014.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. BLEU: a method for automatic evaluation of Machine Translation. In 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. 2002