Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Report on results of the WP4 first evaluation phase

Deliverable number	*D4.3*
Dissemination level	*Public*
Delivery date	*31 August 2012*
Status	*Final*
Author(s)	*Pavel Pecina, Jakub Bystroň, Jan Hajič, Jaroslava Hlaváčová, Zdeňka Urešová*

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 paper reports on the results of the first evaluation phase of Khresmoi WP4. We describe and analyse the first half of the project in the area of resource acquisition and analysis for training the Khresmoi Machine Translation component. The component comprises of six translation systems (English→French, English→German, English→Czech and French→English, German→English, Czech→English) tuned for translation of text from the medical domain. The systems are evaluated on in-domain test sets using standard evaluation metrics.

Table of Contents

List of abbreviations

CS	Czech
DE	German
EN	English
EMEA	European Medicines Agency
FR	French
HTTP REST	Hypertext Transfer Protocol, Representational State Transfer
JSON	JavaScript Object Notation
MERT	Minimum Error Rate Training
MeSH	Medical Subject Headings
MT	Machine Translation
SMT	Statistical Machine Translation

1 Introduction

The Machine Translation (MT) service is an essential component of Khresmoi which provides cross-lingual capability of searching in biomedical documents. 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.

In this evaluation phase, apart from English (EN), we support three other user languages: French (FR), German (DE), and Czech (CS). These three languages serve as *target* languages when translating the summaries of search results from English and as *source* languages when translating the non-English user queries to English. In total, the service provides translation in six directions: EN→FR, EN→DE, EN→CS and FR→EN, DE→EN, CS→EN. Each translation direction is realized as a separate Statistical Machine Translation (SMT) system integrated into one web service.

The report is organized as follows. After the introduction in this section we describe details of the applied SMT system in Section 2. Section 3 reviews the data resources employed for training, tuning, and testing the translation systems. Section 4 describes the technical details of the entire translation component. Section 5 presents the main evaluation results and Section 6 concludes the report.

2 System description

The Khresmoi Machine Translation system is based on the phrase-based SMT decoder Moses (Koehn et al., 2007) and other related tools[1]. In this section, we briefly describe the basic concept of phrase-based SMT and then provide details of our systems and their parameters.

2.1 Phrase-Based Statistical Machine Translation

In phrase-based SMT an input sentence is segmented into (multiple) sequences of consecutive words which are called phrases (typically not linguistic phrases). Each phrase in each input sequence is then translated into (multiple) target language phrases. The sequence of translated phrases may also be reordered to produce the final output. Formally, the phrase-based SMT model is based on the noisy channel model and the best translation of an input sentence is searched for by maximizing the translation probability formulated as a log-linear combination of a set of feature functions and their weights.

The components of the phrase-based SMT model usually include features of the following models:

- *phrase translation model* (phrase translation probabilities, lexical weighting, and phrase penalty) which ensures that the source and target phrases are good translations of each other,
- *language model* which ensures that the translations are fluent,
- *reordering model* which allows to reorder phrases in the input sentences, and
- *word penalty* which regulates length of the translation.

Two kinds of training data are needed for training a complete system. *Parallel training data* for training the phrase translation and reordering model and *monolingual training data* for training the language model.

[1] http://www.statmt.org/moses

Parallel data comprises a set of sentences in one (source) language, each translated to the other (target) language. Such sentence pairs (in the source and target language) are called parallel sentences.

During training, the parallel sentences are aligned on word level — words which are translations of each other are linked together. This information is then used to identify phrase pairs of various length which are mutual translations and form the translation and reordering models.

Monolingual training data is a set of texts in the target language (a language model ensuring the fluency of the output is built only for the target language). Often, the target side of the parallel training data is used with some additional monolingual data (which are not as scarce as parallel data) in the target language.

The weights of the log-linear combination are optimised by Minimum Error Rate Training (MERT), proposed by Och (2003), which automatically searches for the optimal values that maximize a given translation quality measure on a development set of parallel sentences.

To reduce data sparsity, the training data (monolingual and parallel) is usually lowercased and the entire translation (decoding) procedure is performed without information about true letter casing in words and phrases. Reconstruction of the true casing in the output translation can be done by recaser, a simple SMT model which "translates" from lowercased text to text with true letter casing. The model is trained on parallel training data formed by the original and lowercased versions of monolingual training data.

2.2 Training pipeline

Before training the system, the training data is tokenized (segmented into tokens — words and punctuation marks) and lowercased. The original (non-lowercased) target sides of the parallel data and monolingual data are kept for training the Moses recaser. The lowercased versions of the target sides of the parallel data are used for training an interpolated 5-gram language model with Kneser-Ney discounting using the IRSTLM toolkit (Federico et al, 2008). Translation models are trained on the parallel training data, lowercased and filtered on sentence level; we kept all sentence pairs having less than 100 words on each side. Word alignment is done by GIZA++ (Och and Ney, 2003). The phrase pairs are extracted using the alignment parameter *grow-diag-final* with the maximum length of aligned phrases set to 7. The reordering models are trained using parameters: *distance*, *orientation-bidirectional-fe*.[2] The model weights are tuned by MERT on the development sets of parallel sentences.

2.3 Translation pipeline

Each test sentence to be translated is tokenized, lowercased, and then translated by the tuned system. After translation, the letter casing of the words in the sentence is reconstructed by the Moses recaser and extra blank spaces in the tokenized text are removed in order to produce human-readable text by the detokenization tool. The translation pipeline is realized as a web service based on the HTTP REST protocol and JSON format. The details of its implementation are provided in Section 4 of this report.

3 Data acquisition and analysis

This section provides details on parallel and monolingual data used for training, tuning, and testing the Khresmoi MT systems.

[2] For explanation of these parameters see http://www.statmt.org/moses/

SMT is an example of a machine learning application. As such, it requires the test data to be drawn from the same distribution as the training data. In practice, this implies that the training, development, and test data should be from the same domain, of the same genre and style. In Khresmoi, which operates in the area of medical data, the essential requirement is to use data from this domain. In general, availability of domain-specific data is limited. For certain domains (such as news, law, parliament proceedings, etc.) there are large amounts of data (both parallel and monolingual) publicly available with low or no cost (e.g. Europarl, JRC, Hansard, etc.). For other domains, data is very expensive or not available at all. For medicine, the only widely available (in many languages including EN, FR, DE, CS) parallel corpus is EMEA made of documents from the European Medicines Agency and containing about 300 thousand sentence pairs. However, the current SMT systems can take advantage of much larger amounts of training data and can process millions of sentence pairs of parallel training data and billions of words of monolingual training data. In order to acquire enough data for training our SMT systems, we opted to use as much in-domain data as possible together with data from some other domains which are not too specific and can contain enough general language and can potentially improve translation of less technical text in Khresmoi. This strategy is commonly used and proved to improve translation quality (e.g. Pecina et al, 2011).

3.1 Parallel training data

Statistics of the parallel training data used for training the translation and reordering models is given in Tables 1–3. For all language pairs, we used Europarl, JRC-Acquis, and News Commentary as the out-of-domain data and EMEA and MeSH as in-domain data. In addition to this, for the EN–DE translation pair, we used the MuchMore corpus and for the EN–FR pair the COPPA corpus, both as in-domain data. The total amounts of the parallel training data vary for different language pairs: we used 4.5M sentence pairs for EN–FR, 3.2M sentences pairs for EN–DE, and 1.7M sentences pairs for EN–CS. All parallel training data were provided as sentence aligned texts so no further sentence alignment was not necessary to perform. Brief description of the individual parallel corpora follows.

Europarl[3]

The Europarl (Koehn, 2005) parallel corpus is extracted from the proceedings of the European Parliament and the current version (6) includes versions in 21 European languages including EN, FR, DE, and CS.

JRC-Acquis[4]

The JRC-Acquis (Ralf et al., 2006) parallel corpus is extracted from Acquis Communautaire, the total body of European Union law applicable in its Member States, by the Language Technology group of the European Commission's Joint Research Centre. The current version of the corpus (3.0) is currently available in 22 languages including EN, FR, DE, and CS.

News Commentary[5]

The News Commentary parallel corpus comprises news and commentary texts from publicly available sources provided by organizers of the series of Workshops on Machine Translation as shared task training data. The WMT 11 version is available for the following language pairs: FR–EN, DE–EN, and CS–EN.

EMEA[6]

3 http://www.statmt.org/europarl/

4 http://langtech.jrc.it/JRC-Acquis.html

5 http://statmt.org/wmt11

The EMEA (Tiedemann, 2009) is a parallel corpus made out of documents from the European Medicines Agency currently available in 22 languages including EN, FR, DE, and CS.

MeSH[7]

Medical Subject Headings (MeSH) is a controlled vocabulary thesaurus used for indexing medical articles created and updated by the United States National Library of Medicine (NLM), it is used by the MEDLINE/PubMed article database and by NLM's catalog of book holdings. MeSH, originally in English, was translated into many other languages including FR, DE, and CS. In the Khresmoi translation component, MeSH is used as a simple dictionary of medical terms (the parallel corpus is formed by pairs of MeSH descriptors in the relevant languages).

MuchMore[8]

MuchMore Springer Bilingual Corpus is a parallel corpus of English-German scientific medical abstracts obtained from 41 medical journals from the Springer Link web sites. The corpus was created withing the MuchMore project.

COPPA[9]

COPPA is the Corpus Of Parallel Patent Applications provided by WIPO (World Intellectual Property Organization) of English-French Patent Cooperation Treaty applications (title and abstract) published between 1990 and 2010.

Corpus	Sentence pairs	English tokens	French tokens
Europarl	1,967,685	55,465,406	61,607,623
JRC-Acquis	824,059	31,908,923	35,698,556
News-Commentary	136,040	3,398,528	3,997,406
EMEA	355,546	6,024,247	7,165,806
MeSH	39,954	90,460	101,167
COPPA	1,190,005	26,273,305	31,262,412
Total	4,513 289	123,160 869	139,832,970

Table 1: Statistics of the English-French parallel training data.

Corpus	Sentence pairs	English tokens	German tokens
Europarl	1,875,269	52,823,318	50,261,019
JRC-Acquis	807,103	31,099,787	28,575,317
News-Commentary	157,286	3,849,193	3,942,253
EMEA	347,447	5,878,261	5,396,072
MeSH	37,770	84,672	67,045

6 http://opus.lingfil.uu.se/EMEA.php

7 http://www.ncbi.nlm.nih.gov/mesh/

8 http://muchmore.dfki.de/resources1.htm

9 http://www.wipo.int/patentscope/en/data

MuchMore	6,373	1,006,087	914,296
Total	3,231,248	94,741,318	89,156,002

Table 2: Statistics of the English-German parallel training data.

Corpus	Sentence pairs	English tokens	Czech tokens
Europarl	628,595	17,250,292	14,831,766
JRC-Acquis	616,394	24,022,994	20,704,877
News-Commentary	134,692	3,257,301	2,968,195
EMEA	320,034	5,542,309	5,448,756
MeSH	31,182	68,506	68,688
Total	1,730,897	50,141,402	44,022,282

Table 3: Statistics of the English-Czech parallel training data.

3.2 Monolingual training data

Monolingual data is generally less scarce also for the medical domain. All language and recaser models of the Khresmoi translation component were trained on the target sides of the parallel training data and additional in-domain data from various sources. Statistics of the data from these sources is given in Table 4–7. We used the total of 632M tokens for EN, 269M tokens for FR, 57M tokens for DE, and 172M tokens for CS. Brief description of the individual corpora follows.

BMC[10]

BMC comprises of Czech texts from Bibliographia Medica Čechoslovaca, the Czech national register of biomedical and healthcare literature since 1947.

CESART[11]

CESART Evaluation Package was produced within the French national project CESART (Evaluation of terminology extraction tools). Apart from software tools, it contains three domain-specific corpora in French, one of which is this medical corpus.

Cochrane[12]

The Cochrane dataset comprises English reviews of primary research in human health care and health policy.

DrugBank[13]

The DrugBank (Knox et al, 2011) corpus comprises bioinformatics and cheminformatics descriptions of drugs in English.

EQueR[14]

10 http://www.nlk.cz/informace-o-nlk/odborne-cinnosti/bmc/bmc-uvod?set_language=en&cl=en

11 http://catalog.elra.info/product_info.php?products_id=993&language=en

12 http://www.cochrane.org/

13 http://www.drugbank.ca/

The EQueR (Ayache, 2005) cropus contains data from the French Evaluation campaign of question-answering systems.

FMA[15]

The FMA (Rosse and Mejino, 2007) corpus contains English texts from the Foundational Model of Anatomy Ontology — a knowledge source for biomedical informatics concerned with the representation of classes or types and relationships necessary for the symbolic representation of the phenotypic structure of the human body.

Genia[16]

The GENIA (Kim et al. 2003) corpus is the primary collection of biomedical literature compiled and annotated within the scope of the GENIA project. The corpus was created to support the development and evaluation of information extraction and text mining systems for the domain of molecular biology.

GREC[17]

The Gene Regulation Event Corpus (Thompson et al, 2009) is a semantically annotated corpus of biomedical abstracts in English, originally designed for information extraction experiments.

HON

HON corpora comprises texts from the medical domain crawled using the HonBot web spider, as described in Deliverable 8.3, section 3.2.3. The source web sites are all HONcode certified and come in a variety of languages. The language of each page is automatically identified using a statistical language detection library. For the machine translation component only documents in EN, FR, DE, and CS were used.

PIL[18]

The Patient Information Leaflet Corpus (v 3.0) is a collection of several hundred documents giving instructions to patients about their medication.

Radio2wiki[19]

Radio2wiki (Lechner and Breitenseher, 2003) is the text from the german textbook on radiology diagnosis: Lehrbuch der radiologisch-klinischen Diagnostik by Lechner and Breitenseher (2003).

Corpus	Sentences	Tokens
HON	1,882,030	573,432,969
Cochrane	2,128,652	56,483,384
Drugbank	23,062	769,345
FMA	149,996	867,434
Genia	18,469	50,0424

14 http://catalog.elra.info/product_info.php?products_id=996

15 http://sigpubs.biostr.washington.edu/view/projects/Foundational_Model_of_Anatomy.html

16 http://www.nactem.ac.uk/genia/

17 http://www.nactem.ac.uk/GREC/

18 http://mcs.open.ac.uk/nlg/old_projects/pills/corpus/PIL/

19 http://universitypublisher.meduniwien.ac.at/radio2wiki/index.php/Main_Page

Grec	241	58,494
Pil	19,949	577,529
Total	**4,222,399**	**632,689,579**

Table 4: Statistics of the English monolingual training data.

Corpus	Sentences	Tokens
HON	762,868	250,302,243
Cesart	163,204	9,111,470
Equer	158,052	9,642,605
Total	**1,084,124**	**269,056,318**

Table 5: Statistics of the French monolingual training data.

Corpus	Sentences	Tokens
HON	275,055	57,726,121
Radio2wiki	2,509	88,302
Total	**277,564**	**57,814,423**

Table 6: Statistics of the German monolingual training data.

Corpus	Sentences	Tokens
HON	6,040	1,255,496
BMC	618,913	18,497,814
Total	**624,953**	**172,586,012**

Table 7: Statistics of the Czech monolingual training data.

3.3 Development and test data

Translation quality of a Machine Translation system should be evaluated on texts (and its reference translation) of the same nature as the text the system will be used to translate. In Khresmoi, such (in-domain) evaluation data should include 1) sentences from summaries of search results returned to a Khresmoi user and 2) Khresmoi user queries. Using in-domain development data for tuning parameters of the systems is equally important because it has a substantial influence on their translation quality (e.g. Pecina et al., 2011). In the first evaluation phase of the Khresmoi Machine Translation component, no representative test and development data of this nature with reference translation was available (the test and development sets will be prepared for the next phase evaluation) and we had to prepare an alternative solution. We selected a random sample of sentence pairs from the EMEA corpus and manually checked them for errors to ensure the sentences are correct translations of each other. As a result, we obtained 2,000 test sentences pairs and 1,064 development sentence pairs in EN–FR, EN–DE, and EN–CS. The test and development sentences were, of course, removed from the

training data. Moreover, we removed all sentences which share more than 80%[20] of words with any test or development sentences. This step was done because the EMEA corpus contains a lot of similar sentences which differ only in one or two words (usually a name of drug or disease) which could bias the evaluation (test data too similar to the training data), see example below:

There are no data on the use of Aliskiren in pregnant women.

There are no data on the use of Avastin in pregnant women.

There are no data on the use of Duloxetine in pregnant women.

There are no data on the use of Tasigna in pregnant women.

However, even this step cannot guarantee that the evaluation is not biased for this reason. A proper alternative would be to use independent tests sets based on Khresmoi summaries of real documents and real user queries and their manual reference translations.

4 Technical specification of the translation service

This section provides specification of the first version of the Application Programming Interface (API) of the Khresmoi Machine Translation (MT) web service developed and maintained by the Charles University in Prague (CUNI). Detailed description is provided for the system architecture, client-server communication, request and response format (including required and optional parameters), error messages, special features (including n-best list and alignment information), and current limitations.

4.1 Introduction

The MT component provides complete Machine Translation services for Khresmoi. The functionality includes translation of user queries from Czech, French, and German to English, translation of document summaries from English to Czech, French, and German, and translation of full documents from English to Czech, French, and German. Optionally, the service provides multiple translation options for a given input. The translations can also be supplied with alignment information (which links parts of the input sentence with corresponding parts of the output sentence).

4.2 Architecture

The translation service is implemented using the standard HTTP protocol and the principles of REST. REST (Representational State Transfer) is a software architecture for distributed systems where clients communicate with servers – clients initiate requests to servers (e.g. a sentence to be translated); servers process requests and return appropriate responses (e.g. a translated sentence).

The translation service server accepts requests only via the POST method (the GET method is not allowed). If the service is available, the return HTTP code is always 200 OK – even if the server is not able to translate a given input (in that case there is a special error message sent in the response, see below). HTTP error codes other then 200 OK retain their usual meaning (e.g. 500 Internal Error).

[20] The threshold of 80 % was set empirically and allows e.g. one different word in a sentence of five words and two different words in a sentence of 10 words which is typical for many similar senteces in the EMEA corpus . This filter reduced the amount of training data from this corpus by approximatelly 5 % which will be used in next versions of the translation systems if independent development and test sets are employed.

The requests and responses conform to the JSON format. JSON is a language independent format used for serializing and transmitting structured data over a network connection, primarily between a client and server. It is a simpler alternative to XML.

4.3 Request format

The possible request parameters are:

- *action*: string, function name, for testing purposes the only option is translate. Stable version provides additional service functions (required)
- *sourceLang*: string, ISO 639-1 code of the source language (cs, en, de, fr) (required)
- *targetLang*: string, ISO 639-1 code of the target language (cs, en, de, fr) (required)
- *docType*: string (reserved)
- *profileType*: string (reserved)
- *nBestSize*: integer, maximum number of candidates for translation (optional, default = 1, i.e. one best translation is provided, the maximum value is set to 10).
- userId: string, globally unique user ID (optional in the dev version, required in the stable version, IDs will be issued by CUNI upon request)
- *text*: string, text to be translated in UTF-8 character encoding (required, maximum length is limited to 100 words)

A request is validated against this schema:

```
{
        "type": "object",
        "properties": {
        "action": {"type": "string"},
        "userId": {"type": "string", "required": False},
        "sourceLang": {"type": "string"},
        "targetLang": {"type": "string"},
        "text": {"type": "string"},
        "nBestSize": {"type": "integer", "required": False},
        "alignmentInfo": {"type": "bool", "required": False},
        "docType": {"type": "string", "required": False},
        "profileType": {"type": "string", "required": False},
        },
}
```

4.4 Response format

The response structure includes one or more translation structures (depending on the presence of nBestSize parameter) or an error structure.

The translation structure consists of:

text: string, translated text in UTF-8 character encoding

inputTokens: string, space separated sequence of input tokens

outputTokens: string, space separated sequence of output tokens

wordAlignment: dictionary, alignment information (see bellow)

score: number (reserved)

translationId: string, globally unique ID of the transaction

The error fields consist of:

errorCode: number, code of the error

errorMessage: detailed description of the error

Error Codes

0: OK

1: System is temporarily down

2: System busy

3: Invalid language pair

5: Parse error, missing or invalid argument …

5 Evaluation

The evaluation is carried out on the test sets of 2000 sentence pairs using the standard automatic evaluation measures: BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Denkowski and Lavie, 2011). The results for all translation directions are presented in Table 8.

Translation direction	BLEU	NIST	METEOR
EN→FR	0.4952	9.36	0.5586
EN→DE	0.3969	8.41	0.4680
EN→CS	0.4083	8.34	0.2882
CS→EN	0.5261	9.88	0.3722
DE→EN	0.4908	9.45	0.3460
FR→EN	0.5663	10.22	0.3977

Table 8: Translation evaluation results (the BLEU scores are percentages).

The automatic evaluation measures score how the produced translations match the reference translations. BLEU calculates n-gram precision (where n ranges from 1 to 4) and all n-grams are weighted equally. NIST improves BLEU by putting more weight to rarer (more informative) n-grams.

METEOR counts not only exact matches, but also matches based on stems, synonyms, and paraphrase matches between words and phrases. For various reasons (e.g. difference in comprehension of different languages, existence of other acceptable translations which may differ from the reference ones) their scores cannot be meaningfully compared across different systems and different language pairs and we cannot claim which of our six systems is better or worse (or to what extent). Manual (human) evaluation which would allow such comparison was not planned in this first phase of the project. However, the scores of the automatic evaluation measures are roughly comparable to or even higher than results reported e.g. within the WMT11 workshop (see Callison-Burch et al., 2011) and manual analysis of a sample of translated sentences from the test sets (see the examples below) indicate that the translation quality is fair – although the actual satisfaction of Khresmoi users is yet to be tested.

Example 1:

EN→FR

EN: *Therefore, the use of MicardisPlus is not recommended.*

FR: *Par conséquent, l ' utilisation de MicardisPlus n ' est pas recommandée.*

EN→DE

EN: *Therefore, the use of MicardisPlus is not recommended.*

DE: *Daher wird die Anwendung von MicardisPlus wird nicht empfohlen.*

EN→CS

EN: *Therefore, the use of MicardisPlus is not recommended.*

CS: *Proto se použití přípravku MicardisPlus se nedoporučuje.*

FR→EN

FR: *L ' utilisation de MicardisPlus n ' est pas recommandée chez ces patients.*

EN: *The use of MicardisPlus is not recommended in these patients.*

DE→EN

DE: *Daher wird die Anwendung von MicardisPlus nicht empfohlen.*

EN: *Therefore, the use of MicardisPlus is not recommended.*

CS→EN

CS: *Proto u nich není podávání přípravku MicardisPlus doporučeno.*

EN: *Therefore, in the use of MicardisPlus is not recommended.*

Example 2:

EN→FR

EN: *If your doctor's instructions are different from the amounts on the table, follow your doctor's instructions.*

FR: *Si les instructions de votre médecin sont différentes des montants sur la table, Suivez les instructions de votre médecin.*

EN→DE

EN: If your doctor's instructions are different from the amounts on the table, follow your doctor's instructions.

DE: Wenn die Anweisungen Ihres Arztes anders sind als die Beträge auf dem Tisch, befolgen Sie die Anweisungen Ihres Arztes.

EN→CS

EN: *If your doctor's instructions are different from the amounts on the table, follow your doctor's instructions.*

CS: *Pokud pokynů lékaře se liší od částky na stole, dbejte pokynů lékaře.*

FR→EN

FR: *Si la dose prescrite par votre médecin est différente de celle indiquée dans cette table, respectez les instructions de votre médecin.*

EN: *If the dose prescribed by your doctor is different from that which is written in this table, follow the instructions of your doctor.*

DE→EN

DE: *Sollten Sie Fragen bezüglich der Dosierung haben, wenden Sie sich an Ihren Arzt.*

EN: *If you have questions about the dose, ask your doctor.*

CS→EN

CS: *Pokud jsou pokyny od Vašeho lékaře odlišné od údajů v tabulce, řiďte se pokyny lékaře.*

EN: *Instructions from your doctor if they are different to the data at the table, please follow the instructions your doctor.*

6 Conclusion

In this report, we present the first evaluation phase of the Khresmoi Machine Translation component. This component provides two types of service: 1) translation of summaries of search results from English to French, German, and Czech and 2) translation of user queries from French, German, and Czech to English. The service is realized as six separate Statistical Machine Translation systems integrated into one web service based on the HTTP REST protocol and JSON format.

The translation systems were trained on mixtures of data from the medical (e.g. European Medicine Agency documents) and other domains (e.g. news and parliament proceedings), tuned and tested on domain-specific data acquired for the purposes of the evaluation within Khresmoi. The current

evaluation results are promising although they might be biased (to some extent) by a certain similarity of the training and test data. A more proper evaluation will be carried out on independent data from Khresmoi (summaries of real documents indexed in the search system) and real user queries or in the global (extrinsic) context of the entire search workflow in the later phases of the project.

7 References

[1] Ayache, C., Grau, B., Vilnat, A., (2005). Campagne d'évaluation EQueR-EVALDA : Evaluation en QuestionRéponse. Actes de l'Atelier EQueR-EASY de TALN'05, Dourdan, France.

[2] Chris Callison-Burch, Philipp Koehn, Christof Monz and Omar Zaidan (2011). Findings of the 2011 Workshop on Statistical Machine Translation. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 22-64, Edinburgh, Scotland.

[3] Michael Denkowski and Alon Lavie (2011), "Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems", Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation.

[4] Doddington, G. (2002). Automatic evaluation of Machine Translation quality using n-gram co-occurrence statistics. In Proc. of the second international conference on Human Language Technology Research, pp 138–145, San Diego, California.

[5] Marcello Federico, Nicola Bertoldi, Mauro Cettolo (2008). IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models, Proceedings of Interspeech, Brisbane, Australia.

[6] 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.

[7] Philipp Koehn (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In Conference Proc.: the tenth Machine Translation Summit, pages 79–86, Phuket, Thailand.

[8] Gerhard Lechner, Martin Breitenseher u. a. (2003): Lehrbuch der radiologischen klinischen Diagnostik. Maudrich, ISBN 3-85175-754-8, KNO-NR: 11 08 93 84.

[9] Franz Josef Och (2003). Minimum error rate training in Statistical Machine Translation. In 41st Annual Meeting on Association for Computational Linguistics, pages 160–167, Sapporo, Japan.

[10] Franz Josef Och, Hermann Ney (2003). "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, pp. 19-51.

[11] Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi and Jun'ichi Tsujii (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics. 19(suppl. 1). pp. i180-i182, Oxford University Press. ISSN 1367-4803.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu (2002). BLEU: a method for automatic evaluation of Machine Translation. In 40th Annual Meeting on Association for Computati.

[13] Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS. (2011). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res., 39 (Database issue).

[14] Pavel Pecina, Antonio Toral, Andy Way, Vassilis Papavassiliou, Prokopis Prokopidis, and Maria Giagkou (2011). Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation. In Mikel L. Forcada, Heidi Depraetere, and Vincent Vandeghinste, editors,

Proceedings of the 15th Annual Conference of the European Associtation for Machine Translation, pages 297-304, Leuven, Belgium.

[15] Cornelius Rosse and Jose L V Mejino (2007) The Foundational Model of Anatomy Ontology, in Burger, A and Davidson, D and Baldock, R, Eds. Anatomy Ontologies for Bioinformatics: Principles and Practice, pages pp. 59-117. Springer.

[16] Andreas Stolcke (2002). SRILM-an extensible language modeling toolkit. In Proceedings of International Conference on Spoken Language Processing, pages 257–286, Denver, Colorado, USA.

[17] Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, Dániel Varga (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy.

[18] Thompson, P., Iqbal, S. A., McNaught, J. and Ananiadou, S. (2009). Construction of an annotated corpus to support biomedical information extraction. BMC Bioinformatics 10:349.

[19] Jorg Tiedemann (2009). News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In Nicolas Nicolov, Galia Angelova, and Ruslan Mitkov, editors, Recent Advances in NaturalLanguage Processing V, volume 309 of Current Issues in Linguistic Theory, pages 227–248. John Benjamins, Amsterdam & Philadelphia.