Exodus – Exploring SMT for EU Institutions

Michael Jellinghaus[1,2], **Alexandros Poulis**[1], **David Kolovratník**[3]
1: European Parliament, Luxembourg
2: Saarland University, Saarbrücken, Germany
3: Charles University in Prague, Czech Republic
micha@coli.uni-sb.de, apoulis@europarl.europa.eu, david@kolovratnik.net

Abstract

In this paper, we describe Exodus, a joint pilot project of the European Commission's Directorate-General for Translation (DGT) and the European Parliament's Directorate-General for Translation (DG TRAD) which explores the potential of deploying new approaches to machine translation in European institutions. We have participated in the English-to-French track of this year's WMT10 shared translation task using a system trained on data previously extracted from large in-house translation memories.

1 Project Background

1.1 Translation at EU Institutions

The European Union's policy on multilingualism[1] requires enormous amounts of documents to be translated into the 23 official languages (which yield 506 translation directions). To cope with this task, the EU has the biggest translation service in the world, employing almost 5000 internal staff as translators (out of which 1750 at the European Commission (EC) and 760 at the European Parliament (EP) alone), backed up by more than 2000 support staff. In 2009, the total output of the Commission's Directorate-General for Translation (DGT) and the Parliament's Directorate-General for Translation (DG TRAD) together was more than 3 million translated pages. Thus, it is not surprising that the cost of all translation and interpreting services of all the EU institutions amounts to 1% of the annual EU budget (2008 figures). According to our estimations, this is more than € 1 billion per year.

1.2 Machine Translation and Other Translation Technologies at EU Institutions

In order to make the translators' work more efficient so that they can translate more pages in the same time, a number of tools like terminology databases, bilingual concordancers, and, most importantly, translation memories are at their disposition, most of which are heavily used.

[1] http://ec.europa.eu/education/languages/eu-language-policy/index_en.htm

In real translation production scenarios, Machine Translation is usually used to complement translation memory tools (TM tool). Translation memories are databases that contain text segments (usually sentences) that are stored together with their translations. Each such pair of source and target language segments is called a translation unit. Translation units also contain useful meta-data (creation date, document type, client, etc.) that allow us to filter the data both for translation and machine translation purposes.

A TM tool tries to match the segments within a document that needs to be translated with segments in the translation memory and propose translations. If the memory contains an identical string then we have a so-called exact or 100% match which yields a very reliable translation. Approximate or partial matches are called fuzzy matches and usually, the minimum value of a fuzzy match is set to 65%–70%. Lower matches are not considered as usable since they demand more editing time than typing a translation from scratch. First experiments have shown that the quality of SMT output for certain language pairs is equal or similar to 70% fuzzy matches.

Consequently, the cases where machine translation can play a helpful role in this context is when, for a segment to be translated, there is no exact match and the available fuzzy matches do not exceed a certain threshold. This threshold in our case is expected to be 85% or lower. To this end, there exists a system called ECMT (European Commission Machine Translation; also accessible to other European institutions) which is a rule-based system.

However, only certain translation directions are covered by ECMT, and its maintenance is quite complicated and requires quite a lot of dedicated and specialized human resources. In the light of these facts and with the addition of the languages of (prospective) new member states, statistical approaches to machine translation seem to offer a viable alternative.

First of all, SMT is data-driven, i.e. it exploits parallel corpora of which there are plenty at the EU institutions in the form of translation memories. Translation memories have two main advantages over other parallel corpora. First of all, they contain almost exclusively perfectly aligned segments, as each segment is stored together with its translation, and secondly,

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 116–120,
Uppsala, Sweden, 11-16 July 2010.

they contain cleaner data since their content is regularly maintained by linguists and database administrators. SMT systems are quicker to develop and easier to maintain than rule-based systems. The availability of free, open-source software like Moses[2] (Koehn et al., 2007), GIZA++[3] (Och and Ney, 2003) and the like constitutes a further argument in their favor.

Early experiments with Moses were started by members of DGT's Portuguese Language Department as early as summer 2008 (Leal Fontes and Machado, 2009), then turned into a wider interinstitutional project with the codename Exodus, currently combining resources from European Commission's DGT and European Parliament's DGTRAD. Exodus is the first joint project of the interinstitutional Language Technology Watch group where a number of EU institutions join forces in the field of language technology.

2 Participation in WMT 2010 Shared Task

After the English-Portuguese experiments, the first language pair for which we developed a system with a sizeable amount of training data was English-to-French. This system has been developed for testing at the European Parliament. As English-to-French is also one of the eight translation directions evaluated in this year's shared translation task, we decided to participate. The reasons behind this decision are manifold: We would like to

- know where we stand in comparison to other systems,
- learn about what system adaptations are the most beneficial,
- make our project known to potential collaborators,
- compare the WMT10 evaluation results to the outcome of our in-house evaluation.

There is, however, one major difference between the evaluation as carried out in WMT10 and our in-house evaluation: The test data of WMT10 consists exclusively of news articles and is thus out-of-domain for our system intended for use within the European Parliament. This means that the impact of training our system on the in-domain data we obtain from our translation memories cannot be assessed properly, i.e. taking into consideration our specific translation production needs.

Therefore, we would like to invite other interested groups to also translate our in-domain test data with the goal of seeing how our translation scenario could benefit from their setups. Due to legal issues, however, we unfortunately cannot provide our internal training data at this moment.

[2] `http://www.statmt.org/moses/`

[3] `http://www.fjoch.com/GIZA++.html`

3 Data Used

To build our English-to-French MT system, we did not use any of the data provided by the organizers of the WMT10 shared translation task. Instead, we used data that was extracted from the translation memories at the core of EURAMIS (European Advanced Multilingual Information System; (Theologitis, 1997; Blatt, 1998)) which are the fruit of thousands of man-years contributed by translators at EU institutions who, each day, upload the majority of the segments they translate.

Initially (before pre-processing), our EN-FR corpus contained 10,446,450 segments and included documents both from the Commission and the EP from common legislative procedures. These segments were extracted in November 2009 from 7 translation memories hosted in Euramis. Currently, we do not have information about the exact document types coming from the Commission's databases. The Parliament's document types used include, among others:

- legislative documents such as draft reports, final reports, amendments, opinions, etc.,
- documents for the plenary such as questions, resolutions or session amendments,
- committee and delegation documents,
- documents concerning the ACP[4] and the EMPA[5],
- internal documents such as budget estimates, staff regulations, rules of procedure, etc.,
- calls for tender.

Any sensitive or classified documents or Commission-internal documents that do not belong to common legislative procedures have been excluded from the data.

In terms of preprocessing, we performed several steps. First, we obtained translation memory exchange (TMX) files from EURAMIS and converted them to UTF-8 text as the Euramis native character encoding is UCS-2. Then we removed certain control characters which otherwise would have halted processing, we extracted the two single-language corpora into a plain-text file, tokenized and lowercased the data. Finally, we separated the corpus into training data (9,300,682 segments), and data for tuning and testing – 500 segments each. These segments did not exceed a maximum length of 60 tokens and were excluded from the preparation of the translation and language models. The models were then trained on the remaining segments. The maximum length of 60 tokens was applied here as well.

[4] African, Caribbean and Pacific Group of States

[5] Euro-Mediterranean Parliamentary Assembly

Metric	Score
BLEU	18.8
BLEU-cased	16.9
TER	0.747

Table 1: Automatic scores calculated for Exodus in WMT10

4 Building the Models and Decoding

The parallel data described above was used to train an English-to-French translation model and a French target language model. This was done on a server running Sun Solaris with 64 GB of RAM and 8 double core CPU's @1800 Mhz (albeit shared with other processes running simultaneously).

In general, we simply used a vanilla Moses installation at this point, leaving the integration of more sophisticated features to a later moment, i.e. after a thorough analysis of the results of the present evaluation campaign when we will know which adaptations yield the most significant improvements.

For the word alignments, we chose MGIZA (Gao and Vogel, 2008), using seven threads per MGIZA instance, with the parallel option, i.e. one MGIZA instance per pair direction running in parallel. The target language model is a 7-gram, binarized IRSTLM (Federico et al., 2008). The weights of the distortion, translation and language models were optimized with respect to BLEU scores (Papineni et al., 2002) on a given held-out set of sentences with Minimum Error Rate Training (MERT; (Och, 2003)) in 15 iterations.

After the actual translation with Moses, an additional recasing "translation" model was applied in the same manner. Finally, the translation output underwent minimal automatic postprocessing based on regular expression replacements. This was mainly undertaken in order to fix the distribution of whitespace and some remaining capitalization issues.

5 Results

5.1 WMT10 Evaluation

In one of the tasks of the WMT10 human evaluation campaign, people were asked to rank competing translations. From each 1-through-5 ranking of a set of 5 system outputs, 10 pairwise comparisons are extracted. Then, for each system, a score is computed that tells how often it was ranked equally or better than the other system. For our system, this score is 32.35%, meaning it ranked 17th out of 19 systems for English-to-French. A number of automatic scores were also calculated and appear in Table 1.

5.2 Evaluation at the European Parliament

As the goal behind building our system has been to provide a tool to translators at EU institutions, we have also had it evaluated by two of our colleagues, both native speakers of French and working as professional translators of the French Language Unit at the Parliament's DG TRAD.

	Evaluator A	Evaluator B	**Overall**
Reference	1.75	2.06	**1.97**
ECMT	3.34	3.31	**3.32**
Google	3.59	3.28	**3.37**
Exodus	3.52	3.45	**3.47**

Table 2: Average relative rank (on a scale from 1 to 5)

	OK	Edited	Bad
Reference	29	30	2
ECMT	8	57	2
Google	7	33	5
Exodus	13	62	12

Table 3: Results of Editing Task ("OK" means "No corrections needed"; "Bad" means "Unable to correct")

For this purpose, we had 1742 sentences of in-house documents translated by our system as well as by the rule-based ECMT and the statistics-based Google Translate.[6,7] We developed an online evaluation tool based on the one used by the WMT evaluation campaign in the last years (Callison-Burch et al., 2009) where we asked the evaluators to perform three different tasks.

In the first one, they were shown the three automatic translations plus a human reference in random order and asked to rank the four versions relative to each other on a scale from 1 to 5. The average relative ranks can be seen in Table 2.

The second task consisted of post-editing a given translation. Again, the sentence might come from one of three MT systems, or be a human translation. The absolute number of items that did not need any corrections, had to be edited, or were impossible to correct are shown in Table 3.

For the third and last task, only translations of our own system were displayed. Here, the evaluators should simply assign them to one of four quality categories as proposed by (Roturier, 2009), and additionally tick boxes standing for the presence of 13 different types of errors in the sentence concerning word order, punctuation, or different types of syntactic/semantic problems. A total of 150 segments were judged. For the categorization results, see Tables 4 and 5.

5.3 Evaluation at the European Commission

On a side note, the Portuguese Language Department also performed a manual evaluation (Leal Fontes and Machado, 2009) which involved 14 of their managers and translators, comparing their Moses-based system to

[6] http://translate.google.com

[7] As about a third of the source documents are not public, we could not send those to Google Translate.

	Items	Proportion
Excellent	28	18.6%
Good	42	28%
Medium	45	30%
Poor	35	23.3%

Table 4: Results of Categorization Task: Quality Categories

Error type	Occurrences
Word order	
Single word	11
Sequence of words	42
Incorrect word(s)	
Wrong lexical choice	51
Wrong terminology choice	6
Incorrect form	77
Extra word(s)	21
Missing word(s)	14
Style	44
Idioms	1
Untranslated word(s)	5
Punctuation	24
Letter case	7
Other	5

Table 5: Results of Categorization Task: Error Types

ECMT and Google. Table 6 shows how many people considered Moses better, similar, or worse compared to ECMT and Google, respectively.

Moses-based SMT did well in fields where ECMT is systematically used (e.g. Justice and Home Affairs and Trade) and showed a big improvement over ECMT in terminology-intensive domains (e.g. Fisheries). As of early 2009, more than half of their translators (58%) now already use ECMT systematically in production, i.e. for all English and French originals. 85% use it for specific language combinations or for certain domains only. On a voluntary basis, they have been replacing ECMT with Moses-based SMT for the translation of day-to-day incoming documents. Over a three-month period, more than 2500 pages have been translated in this manner, and the translators of the Portuguese department declared themselves ready to switch over to an SMT system as soon as it should become available.

Compared to	Better	Similar	Worse
ECMT	7	5	2
Google	5	5	3

Table 6: Portuguese Language Department evaluation results of Moses-based MT system

6 Discussion of Results

As expected, our system did not rank among the top competitors in the WMT10 shared task. This is mainly due to the data we trained on, which is of a very specific domain (common legislative procedures of European Institutions) and relatively small in size when compared to what others used for this language combination. In addition, we more or less used Moses out-of-the-box with no sophisticated add-ons or optimization.

In the internal evaluation, our system beat neither Google Translate nor ECMT overall but it did show a similar performance. This is all the more encouraging as Exodus has been built within less than a month, while ECMT has been developed and maintained in excess of 30 years, and while Google Translate is based on manpower and computing resources that a public administration body usually cannot provide.

Finally, the successful trials of SMT software at the EC's Portuguese department seem to indicate that such a system holds enormous potential, especially when a serious adaptation to specific language combinations and domains is taken into consideration.

7 Outlook

Further use and development of SMT at EU institutions depends on the outcome of internal evaluations, among other factors. We plan to extend our activities to other language pairs, an English-to-Greek machine translation project already having started. Given a continuation of the currently promising results, Exodus will eventually be integrated into the CAT (computer-aided translation) tools used by EU translators.[8] Furthermore, we would like to release an extended EuroParl corpus not only containing parliamentary proceedings but also other types of public documents. We estimate that such a step should foster research to the benefit of both EU institutions and machine translation in general.

8 Conclusions

We have presented Exodus, a joint pilot project of the European Commission's Directorate-General for Translation (DGT) and the European Parliament's Directorate-General for Translation (DG TRAD) with the aim of exploring the potential of deploying new approaches to machine translation in European institutions.

Our system is based on a fairly vanilla Moses installation and trained on data extracted from large in-house translation memories covering a range of EU documents. The obtained models use 7-grams.

We applied the Exodus system to this year's WMT10 shared English-to-French translation task. As the test

[8] However, speed issues will have to be addressed before as the current system is not able to provide translations in real time.

data stems from a different domain than the one targeted by our system, we did not outperform the competitors. However, results from in-house evaluation are promising and indicate the big potential of SMT for European Institutions.

Acknowledgments

We would very much like to thank (in alphabetical order) Manuel Tomás Carrasco Benítez, Dirk De Paepe, Alfons De Vuyst, Peter Hjortsø, Herman Jenné, Hilário Leal Fontes, Maria José Machado, Spyridon Pilos, João Rosas, Helmut Spindler, Filiep Spyckerelle, and Angelika Vaasa for their invaluable help and support.

David Kolovratník was supported by the Czech Science Foundation under contract no. 201/09/H057 and by the Grant Agency of Charles University under contract no. 100008/2008.

References

A. Blatt. 1998. EURAMIS : Added value by integration. In *T&T Terminologie et Traduction, 1.1998*, pages 59–73.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Proceedings of Interspeech*, Brisbane, Australia.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June. Association for Computational Linguistics.

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 *Proc. of ACL Demo and Poster Sessions*, pages 177–180, Jun.

Hilário Leal Fontes and Maria José Machado. 2009. Contribution of the Portuguese Langauge Department to the Evaluation of Moses Machine Translation System. Technical report, Portuguese Language Department, DGT, European Commission, December.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, March.

Franz Josef Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proceedings of ACL*, Sapporo, Japan, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Johann Roturier. 2009. Deploying novel MT technology to raise the bar for quality: A review of key advantages and challenges. In *The twelfth Machine Translation Summit*, Ottawa, Canada, August. International Association for Machine Translation.

D. Theologitis. 1997. EURAMIS, the platform of the EC translator. In *EAMT Workshop*, pages 17–32.