SMT at the International Maritime Organization: Experiences with Combining In-house Corpora with Out-of-domain Corpora

Bruno Pouliquen
World Intellectual Property Organization
34, chemin des Colombettes
CH-1211 Geneva, Switzerland
bruno.pouliquen@wipo.int

Marcin Junczys-Dowmunt
Adam Mickiewicz University
ul. Umultowska 87
61-614 Poznań, Poland
junczys@amu.edu.pl

Blanca Pinero
International Maritime Organization
4 Albert Embankment
London SE17SR, United Kingdom
bpinero@imo.org

Michał Ziemski
United Nations
8-14, Avenue de la Paix
CH-1211 Geneva, Switzerland
mziemski@unog.ch

Abstract

This paper presents a machine translation tool – based on Moses – developed for the International Maritime Organization (IMO) for the automatic translation of documents from Spanish, French, Russian and Arabic to/from English. The main challenge lies in the insufficient size of in-house corpora (especially for Russian and Arabic). The United Nations (UN) granted IMO the right to use UN resources and we describe experiments and results we obtained with different translation model combination techniques. While BLEU results remain inconclusive for combinations, we also analyze user preferences for certain models (when choosing betweeen IMO only or combined with UN). The combined models are perceived by translators as being much better for general texts while IMO only models seem better for technical texts.

1 Introduction

This paper describes the installation and training of TAPTA, an MT tool, for the automatic translation of IMO documents. TAPTA has been previously installed at other international organizations (Pouliquen et. al 2013). IMO is a specialized agency of the United Nations system dealing with safe and secure seas and the protection of the marine environment. It has three working languages (English, French and Spanish) with parallel corpora of ca. 60 million words each. A much smaller number of documents (conventions and reports totaling ca. 6 million words per language) are translated into the other official languages of the United Nations (Arabic, Chinese[1], Russian). IMO felt that the introduction of MT would help translators in their daily work, given similar experiences in other UN agencies and repetitive nature of their documentation due to periodic reporting. While building the SMT corpora, the specific terminology used in the maritime domain and the "house style" were also important considerations. The large imbalance in the number of parallel documents between language poses a problem which we try to solve by integrating larger parallel corpora which "complete" the IMO models (especially for Russian and Arabic)[2]. The corpora provided by the United Nations Secretariat were thought to be ideal for this purpose as the language pairs are the same and working practices in both translation services are very similar. Authorization was granted to merge the corpora.

2 Data and preprocessing

The International Maritime Organization has 6 official languages (Arabic, English, Spanish, French,

[1]Work on Chinese data has been postponed and is not described in this paper.

[2]Documents were provided by the Documentation Division (New York) of the Department for General Assembly and Conference Management, the main entity of the United Nations Secretariat charged with the production of parliamentary documentation.

Russian and Chinese), which means that, if such an organization wanted a translation tool for all language pair combinations, it would require 42 translation engines. A rule-based translation system would be extremely costly to build and maintain. A data-driven approach is usually more suitable when a big parallel corpus exists, therefore we focused on SMT.

Moses (Koehn et al. 2007) has been trained with a parallel corpora extracted consisting of IMO documents translated between January 2000 and October 2014 (ca. 20,000 documents for English, French and Spanish, about 400 documents for Russian/Arabic, see Table 1). The provided corpora have been extracted from original Word or PDF documents, identical IDs between languages allow to align documents for each language pair. We use an in-house (WIPO) sentence aligner. The tool processes each parallel text document and produces a set of aligned sentences after applying the following steps:

- Sentence splitting
- Tokenization
- Sentence alignment with our sentences aligned (based on Champollion (Ma 2006)) — produces an "aligned-segment-matching-score"
- filtering out whole documents with an average-segment-matching-score below a given threshold
- filtering out sets of consecutive segments having a low scores
- filtering out sets of consecutive segments that are sorted by alphabetical order[3]
- filtering out sentences having only one word or more than 80 words, or a source/target word ratio more than 9

3 SMT system

3.1 Baseline system

The baseline SMT system consists of an extended Moses (Koehn et al. 2007) configuration. Durrani et al. (2013) report on improvements for various language pairs when an Operation Sequence Model (OSM) is added to the phrase-based decoder. Class-based language models seem to be a good compromise between increased n-gram length and total model size. We use automatically calculated word cluster ids as classes. We had good experience with word2vec (Mikolov et al. 2012) in the context of larger SMT models and use this tool to compute 200 word classes from the target language data. The target language corpora are mapped to sequences of classes and 9-gram language model are estimated. The final phrase-tables of the larger models (English-French, English-Spanish) have been significance pruned (Johnson et al. 2007) for size reduction. In our experiments significance pruning results in no quality loss while reducing translation model size by a factor of 5. The standard 5-gram language models and the 9-gram word-class models are estimated with Modified Kneser-Ney smoothing (Chen and Goodman 1996, Heafield et al. 2013). To reduce size requirements, we use heavily quantized binary models with no noticeable quality reduction. Pruning is applied to all singleton n-grams with n equal to or greater than 3.

Lang. pair	IMO corpus Docs	Words	Segments	UN Words
en-fr	17132	53.8 M	2.60 M	316 M
en-es	16213	54.0 M	2.50 M	295 M
en-ru	318	5.6 M	0.30 M	296 M
en-ar	296	4.1 M	0.23 M	304 M
en-zh	[not available yet]			280 M

Table 1: Size of the parallel corpora used for training. The fourth and fifth columns show the training size (in millions of English words) for IMO and UN corpus.

3.2 Attempts at domain adaptation

We explore two model combination methods for both, translation models and language models: linear and log-linear interpolation. Log-linear model interpolation is natively supported in Moses via its feature function framework. Translation models and language models can be log-linearly interpolated just by adding them to the Moses configuration files. Parameter tuning then chooses the appropriate interpolation weights which are actually feature weights. Linear interpolation, though a standard method for language models, is more involved. In the case of language models, we

[3] In both IMO and UN, it is very common to sort enumerations of countries, persons, organizations, etc. by alphabetical order, which will of course often be different between languages and results in very noisy sentence alignment.

compute a new static linearly interpolated language model from IMO and UN data target language data. Interpolation weights are optimized on the dev set. In the case of translation models we use a new feature function available in Moses that allows for setting up virtual phrase tables that are in fact linearly interpolated translation models (Sennrich 2012). We use the same interpolation weights as previously determined for linear language model interpolation. The two interpolated translation models are the original IMO and UN translation models as used in stand-alone translators. Results are mixed, we report the best results for our experiments (see Table 2, Section 5.1). Log-linear interpolation is downright harmful (and therefore omitted), for the larger language pairs (en-fr and en-es) any of the interpolation methods seem to be unhelpful, improvements for en-es are within the range of optimizer instability. For the smaller models (en-ar, en-ru) we observe quite significant improvements that stem mainly from linear translation model interpolation.

4 Translating

4.1 Server configuration

The server has been installed on a virtual machine running Ubuntu, the same machine is being used for training and decoding. Server specifications are: 12 cores, 64 GB RAM and 1 TB of disk space.

The server runs several Moses decoders (one decoder is a Moses single-thread executable). Each decoder is encapsulated in a Java RMI interface server which allows to operate several concurrent decoders. Each sentence submitted is queued and sent to the next free decoder. Since both, the phrase table as well as all the language models, rely on memory mapping and shared memory, having several independent workers instead of a multi-threaded architecture does not represent much of a memory problem. Common data is shared automatically between processes. Thanks to our experience in installing the tool, we were able to install and configure the server in 2 days (not including research model parameters and specific experiments with model combinations). Training one IMO model takes ca. 20 hours.

4.2 User interface

4.2.1 Web interface: gist translation

A web interface allows users to submit short texts and access the corresponding automatic translation (with highlighting of parallel segments or words).

Figure 1: Translating with the "auto hotkey"

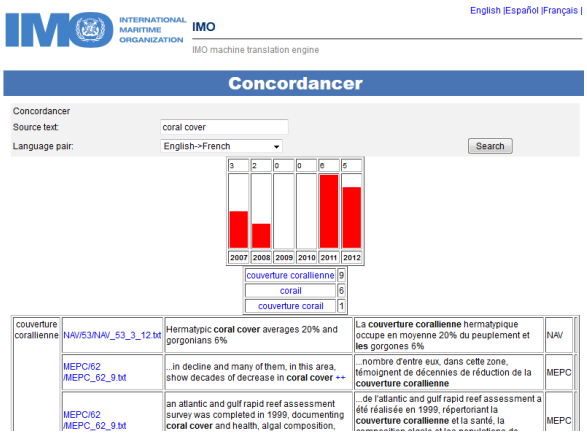

Figure 2: Concordancer for term "coral cover", the graph shows the term usage over years, next the most used translations are display, then the full parallel segments with links to original documents.

4.2.2 "Auto hotkey" access

Translators in IMO use specific software (MultiTrans Prism) and do not wish to copy-paste texts in order to use the tool. So we decided to use the "auto hotkey" open source software (`http://www.autohotkey.com`), which allows users to call the tool with a keystroke, translations are then copied to the clipboard and users can paste it into the translation in-progress (see Figure 1 for an example screen shot).

4.2.3 Concordancer

Users can access the concordancer using a Web interface or through a different "hotkey". The concordancer is based on a Lucene index containing the word aligned corpus. This concordancer displays segments containing the search term and the corresponding aligned words. A first window dis-

	IMO only	Combined	Google
en-fr	54.24	54.03	32.58
en-es	52.68	52.99	35.18
en-ru	58.77	60.20	20.56
en-ar	41.20	44.18	16.58
en-zh		[not available]	

Table 2: BLEU scores for each language pair, compared with a combined model and with Google translate.

plays the usage of the term by year, a second window displays the aligned words by order of frequency, the user can immediately see which translation is the most common (see Figure 2 for an example).

5 Results/Evaluation

5.1 Automatic evaluation

BLEU scores (Papineni et al. 2002) were used to compare human translations with automatic translations (one reference) on a set slightly more than 2000 sentences which have been set apart before model training.

5.2 Human perception

It is always difficult to measure user acceptance, especially at this early stage. However we can now observe that, on average, more than 1500 words are translated every day using our tool. Some users "jump" between various models (eg. users prefer IMO-only models for English-to-Spanish, but nevertheless use the combined model in more than 10% of the cases). Even though the automatic evaluation scores do not show significant improvement with combined models, translators judged combined models to be better for general texts while IMO-only models work better for more technical texts. Additional functionality such as the concordancer are readily embraced and found useful alongside the pure translation function.

6 Conclusion and future work

During our experiments, we had to face both, a scarcity problem (small IMO corpora for some languages) and a scalability problem (large UN corpora). However, our experience shows that open source solutions can sometimes provide better results than generic commercial products. Moreover, sharing the tool between these organizations facilitates sharing of corpora and the spread of MT in international organizations. User comments include that the Web interface is intuitive and the "auto-hotkey" is an easy and fast way of accessing translations; integration like this requires very little training and this training can be done internally. Future work includes: better integration into the users' environment and a biannual retraining of all the models. We believe the model combination technique can still be improved. An area to explore would be to "automatically" choose the best model to translate a given document/sentence.

Acknowledgements

The authors wish to thank Mrs Olga O'Neil, Director, Conference Division, IMO, for making this collaboration possible, and Ms Cecilia Elizalde, UNHQ, for her invaluable assistance in obtaining the UN corpora.

References

Pouliquen B., C. Elizalde, M. Junczys-Dowmunt, C. Mazenc, and J. Garca-Verdugo. 2013. Large-scale Multiple Language Translation Accelerator at the United Nations, *Proc. of MT Summit 2013*.

Chen S. F. and J. Goodman. 1996. An Empirical Study of Smoothing Techniques for Language Modeling, in *Proc. of ACL 1996*.

Durrani, N., A. Fraser, H. Schmid, H. Hoang, and P. Koehn. 2013. Can Markov Models over Minimal Translation Units Help Phrase-based SMT? in *ACL*.

Heafield, K, I. Pouzyrevsky, J. H. Clark, and P. Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation, in *Proc. of ACL 2013*.

Johnson, H., J. D. Martin, G. F. Foster, and R. Kuhn. 2007. Improving Translation Quality by Discarding Most of the Phrasetable, *Proc. of EMNLP 2007*.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris C. Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Proc. of ACL 2007*.

Koehn, Phillip. 2010. *Statistical Machine Translation.* Textbook, Cambridge University Press.

Ma, Xiaoyi. 2006. Champollion: A Robust Parallel Text Sentence Aligner. *Proc. of LREC-2006*.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space, *CoRR, vol. abs/1301.3781*.

Papineni, K., S. Roukos, T. Ward, and WJ Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation, *Proc. of ACL 2002*.

Sennrich R. 2012, Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. *Proc. of EACL 2012*.