Moses: Open Source Toolkit for Statistical Machine Translation

Philipp Koehn
Hieu Hoang
Alexandra Birch
Chris Callison-Burch
University of Edinburgh[1]

Marcello Federico
Nicola Bertoldi
ITC-irst[2]

Brooke Cowan
Wade Shen
Christine Moran
MIT[3]

Richard Zens
RWTH Aachen[4]

Chris Dyer
University of Maryland[5]

Ondřej Bojar
Charles University[6]

Alexandra Constantin
Williams College[7]

Evan Herbst
Cornell[8]

[1] pkoehn@inf.ed.ac.uk, {h.hoang, A.C.Birch-Mayne}@sms.ed.ac.uk, callison-burch@ed.ac.uk. [2]{federico, bertoldi}@itc.it. [3] brooke@csail.mit.edu, swade@ll.mit.edu, weezer@mit.edu. [4] zens@i6.informatik.rwth-aachen.de. [5] redpony@umd.edu. [6] bojar@ufal.ms.mff.cuni.cz. [7] 07aec_2@williams.edu. [8] evh4@cornell.edu

Abstract

We describe an open-source toolkit for statistical machine translation whose novel contributions are (a) support for linguistically motivated factors, (b) confusion network decoding, and (c) efficient data formats for translation models and language models. In addition to the SMT decoder, the toolkit also includes a wide variety of tools for training, tuning and applying the system to many translation tasks.

1 Motivation

Phrase-based statistical machine translation (Koehn et al. 2003) has emerged as the dominant paradigm in machine translation research. However, until now, most work in this field has been carried out on proprietary and in-house research systems. This lack of openness has created a high barrier to entry for researchers as many of the components required have had to be duplicated. This has also hindered effective comparisons of the different elements of the systems.

By providing a free and complete toolkit, we hope that this will stimulate the development of the field. For this system to be adopted by the community, it must demonstrate performance that is comparable to the best available systems. Moses has shown that it achieves results comparable to the most competitive and widely used statistical machine translation systems in translation quality and run-time (Shen et al. 2006). It features all the capabilities of the closed sourced Pharaoh decoder (Koehn 2004).

Apart from providing an open-source toolkit for SMT, a further motivation for Moses is to extend phrase-based translation with factors and confusion network decoding.

The current phrase-based approach to statistical machine translation is limited to the mapping of small text chunks without any explicit use of linguistic information, be it morphological, syntactic, or semantic. These additional sources of information have been shown to be valuable when integrated into pre-processing or post-processing steps.

Moses also integrates confusion network decoding, which allows the translation of ambiguous input. This enables, for instance, the tighter integration of speech recognition and machine translation. Instead of passing along the one-best output of the recognizer, a network of different word choices may be examined by the machine translation system.

Efficient data structures in Moses for the memory-intensive translation model and language model allow the exploitation of much larger data resources with limited hardware.

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 177–180,
Prague, June 2007.

2 Toolkit

The toolkit is a complete out-of-the-box translation system for academic research. It consists of all the components needed to preprocess data, train the language models and the translation models. It also contains tools for tuning these models using minimum error rate training (Och 2003) and evaluating the resulting translations using the BLEU score (Papineni et al. 2002).

Moses uses standard external tools for some of the tasks to avoid duplication, such as GIZA++ (Och and Ney 2003) for word alignments and SRILM for language modeling. Also, since these tasks are often CPU intensive, the toolkit has been designed to work with Sun Grid Engine parallel environment to increase throughput.

In order to unify the experimental stages, a utility has been developed to run repeatable experiments. This uses the tools contained in Moses and requires minimal changes to set up and customize.

The toolkit has been hosted and developed under sourceforge.net since inception. Moses has an active research community and has reached over 1000 downloads as of 1st March 2007.

The main online presence is at

http://www.statmt.org/moses/

where many sources of information about the project can be found. Moses was the subject of this year's Johns Hopkins University Workshop on Machine Translation (Koehn et al. 2006).

The decoder is the core component of Moses. To minimize the learning curve for many researchers, the decoder was developed as a drop-in replacement for Pharaoh, the popular phrase-based decoder.

In order for the toolkit to be adopted by the community, and to make it easy for others to contribute to the project, we kept to the following principles when developing the decoder:

- Accessibility
- Easy to Maintain
- Flexibility
- Easy for distributed team development
- Portability

It was developed in C++ for efficiency and followed modular, object-oriented design.

3 Factored Translation Model

Non-factored SMT typically deals only with the surface form of words and has one phrase table, as shown in Figure 1.

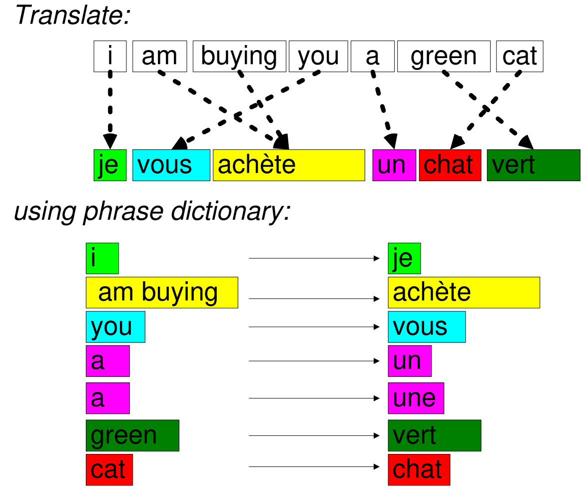

Figure 1. Non-factored translation

In factored translation models, the surface forms may be augmented with different factors, such as POS tags or lemma. This creates a factored representation of each word, Figure 2.

$$\begin{pmatrix} je \\ PRO \\ je \\ 1st \end{pmatrix}\begin{pmatrix} vous \\ PRO \\ vous \\ 1st \end{pmatrix}\begin{pmatrix} achet \\ VB \\ acheter \\ 1st/present \end{pmatrix}\begin{pmatrix} un \\ ART \\ un \\ masc \end{pmatrix}\begin{pmatrix} chat \\ NN \\ chat \\ sing/masc \end{pmatrix}$$

$$\Downarrow$$

$$\begin{pmatrix} i \\ PRO \\ i \\ 1st \end{pmatrix}\begin{pmatrix} buy \\ VB \\ tobuy \\ 1st/present \end{pmatrix}\begin{pmatrix} you \\ PRO \\ you \\ 1st \end{pmatrix}\begin{pmatrix} a \\ ART \\ a \\ sing \end{pmatrix}\begin{pmatrix} cat \\ NN \\ cat \\ sing \end{pmatrix}$$

Figure 2. Factored translation

Mapping of source phrases to target phrases may be decomposed into several steps. Decomposition of the decoding process into various steps means that different factors can be modeled separately. Modeling factors in isolation allows for flexibility in their application. It can also increase accuracy and reduce sparsity by minimizing the number dependencies for each step.

For example, we can decompose translating from surface forms to surface forms and lemma, as shown in Figure 3.

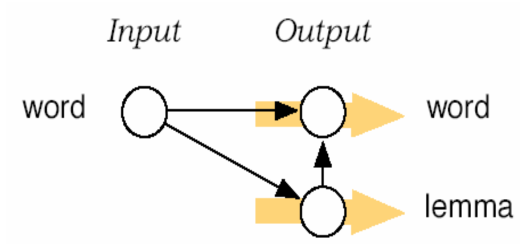

Figure 3. Example of graph of decoding steps

By allowing the graph to be user definable, we can experiment to find the optimum configuration for a given language pair and available data.

The factors on the source sentence are considered fixed, therefore, there is no decoding step which create source factors from other source factors. However, Moses can have ambiguous input in the form of confusion networks. This input type has been used successfully for speech to text translation (Shen et al. 2006).

Every factor on the target language can have its own language model. Since many factors, like lemmas and POS tags, are less sparse than surface forms, it is possible to create a higher order language models for these factors. This may encourage more syntactically correct output. In Figure 3 we apply two language models, indicated by the shaded arrows, one over the words and another over the lemmas. Moses is also able to integrate factored language models, such as those described in (Bilmes and Kirchhoff 2003) and (Axelrod 2006).

4 Confusion Network Decoding

Machine translation input currently takes the form of simple sequences of words. However, there are increasing demands to integrate machine translation technology into larger information processing systems with upstream NLP/speech processing tools (such as named entity recognizers, speech recognizers, morphological analyzers, etc.). These upstream processes tend to generate multiple, erroneous hypotheses with varying confidence. Current MT systems are designed to process only one input hypothesis, making them vulnerable to errors in the input.

In experiments with confusion networks, we have focused so far on the speech translation case, where the input is generated by a speech recognizer. Namely, our goal is to improve performance of spoken language translation by better integrating speech recognition and machine translation models. Translation from speech input is considered more difficult than translation from text for several reasons. Spoken language has many styles and genres, such as, formal read speech, unplanned speeches, interviews, spontaneous conversations; it produces less controlled language, presenting more relaxed syntax and spontaneous speech phenomena. Finally, translation of spoken language is prone to speech recognition errors, which can possibly corrupt the syntax and the meaning of the input.

There is also empirical evidence that better translations can be obtained from transcriptions of the speech recognizer which resulted in lower scores. This suggests that improvements can be achieved by applying machine translation on a large set of transcription hypotheses generated by the speech recognizers and by combining scores of acoustic models, language models, and translation models.

Recently, approaches have been proposed for improving translation quality through the processing of multiple input hypotheses. We have implemented in Moses confusion network decoding as discussed in (Bertoldi and Federico 2005), and developed a simpler translation model and a more efficient implementation of the search algorithm. Remarkably, the confusion network decoder resulted in an extension of the standard text decoder.

5 Efficient Data Structures for Translation Model and Language Models

With the availability of ever-increasing amounts of training data, it has become a challenge for machine translation systems to cope with the resulting strain on computational resources. Instead of simply buying larger machines with, say, 12 GB of main memory, the implementation of more efficient data structures in Moses makes it possible to exploit larger data resources with limited hardware infrastructure.

A phrase translation table easily takes up gigabytes of disk space, but for the translation of a single sentence only a tiny fraction of this table is needed. Moses implements an efficient representation of the phrase translation table. Its key properties are a *prefix tree* structure for source words and *on demand loading*, i.e. only the fraction of the phrase table that is needed to translate a sentence is loaded into the working memory of the decoder.

For the Chinese-English NIST task, the memory requirement of the phrase table is reduced from 1.7 gigabytes to less than 20 mega bytes, with no loss in translation quality and speed (Zens and Ney 2007).

The other large data resource for statistical machine translation is the language model. Almost unlimited text resources can be collected from the Internet and used as training data for language modeling. This results in language models that are too large to easily fit into memory.

The Moses system implements a data structure for language models that is more efficient than the canonical SRILM (Stolcke 2002) implementation used in most systems. The language model on disk is also converted into this binary format, resulting in a minimal loading time during start-up of the decoder.

An even more compact representation of the language model is the result of the *quantization* of the word prediction and back-off probabilities of the language model. Instead of representing these probabilities with 4 byte or 8 byte floats, they are sorted into bins, resulting in (typically) 256 bins which can be referenced with a single 1 byte index. This quantized language model, albeit being less accurate, has only minimal impact on translation performance (Federico and Bertoldi 2006).

6 Conclusion and Future Work

This paper has presented a suite of open-source tools which we believe will be of value to the MT research community.

We have also described a new SMT decoder which can incorporate some linguistic features in a consistent and flexible framework. This new direction in research opens up many possibilities and issues that require further research and experimentation. Initial results show the potential benefit of factors for statistical machine translation, (Koehn et al. 2006) and (Koehn and Hoang 2007).

References

Axelrod, Amittai. *"Factored Language Model for Statistical Machine Translation."* MRes Thesis. Edinburgh University, 2006.

Bertoldi, Nicola, and Marcello Federico. *"A New Decoder for Spoken Language Translation Based on Confusion Networks."* Automatic Speech Recognition and Understanding Workshop (ASRU), 2005.

Bilmes, Jeff A, and Katrin Kirchhoff. *"Factored Language Models and Generalized Parallel Backoff."* HLT/NACCL, 2003.

Koehn, Philipp. *"Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models."* AMTA, 2004.

Koehn, Philipp, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, Alexandra Constantin, Christine Corbett Moran, and Evan Herbst. "*Open Source Toolkit for Statistical Machine Translation"*. Report of the 2006 Summer Workshop at Johns Hopkins University, 2006.

Koehn, Philipp, and Hieu Hoang. *"Factored Translation Models."* EMNLP, 2007.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. *"Statistical Phrase-Based Translation."* HLT/NAACL, 2003.

Och, Franz Josef. *"Minimum Error Rate Training for Statistical Machine Translation."* ACL, 2003.

Och, Franz Josef, and Hermann Ney. *"A Systematic Comparison of Various Statistical Alignment Models."* Computational Linguistics 29.1 (2003): 19-51.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. *"BLEU: A Method for Automatic Evaluation of Machine Translation."* ACL, 2002.

Shen, Wade, Richard Zens, Nicola Bertoldi, and Marcello Federico. *"The JHU Workshop 2006 Iwslt System."* International Workshop on Spoken Language Translation, 2006.

Stolcke, Andreas. *"SRILM an Extensible Language Modeling Toolkit."* Intl. Conf. on Spoken Language Processing, 2002.

Zens, Richard, and Hermann Ney. *"Efficient Phrase-Table Representation for Machine Translation with Applications to Online MT and Speech Recognition."* HLT/NAACL, 2007.