MTTK: An Alignment Toolkit for Statistical Machine Translation

Yonggang Deng [1]
Center for Language and Speech Processing[1]
Johns Hopkins University
Baltimore, MD 21218
dengyg@jhu.edu

William Byrne[1,2]
Machine Intelligence Lab [2]
Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, UK
wjb31@cam.ac.uk

Abstract

The MTTK alignment toolkit for statistical machine translation can be used for word, phrase, and sentence alignment of parallel documents. It is designed mainly for building statistical machine translation systems, but can be exploited in other multi-lingual applications. It provides computationally efficient alignment and estimation procedures that can be used for the unsupervised alignment of parallel text collections in a language independent fashion. MTTK Version 1.0 is available under the Open Source Educational Community License.

1 Introduction

Parallel text alignment procedures attempt to identify translation equivalences within collections of translated documents. This can be be done at various levels. At the finest level, this involves the alignment of words and phrases within two sentences that are known to be translations (Brown et al., 1993; Och and Ney, 2003; Vogel et al., 1996; Deng and Byrne, 2005). Another task is the identification and alignment of sentence-level segments within document pairs that are known to be translations (Gale and Church, 1991); this is referred to as sentence-level alignment, although it may also involve the alignment of sub-sentential segments (Deng et al.,) as well as the identification of long segments in either document which are not translations. There is also document level translation which involves the identification of translated document pairs in a collection of documents in multiple languages. As an example, Figure 1 shows parallel Chinese/English text that is aligned at the sentence, word, and phrase levels.

Parallel text plays a crucial role in multi-lingual natural language processing research. In particular, statistical machine translation systems require collections of sentence pairs (or sentence fragment pairs) as the basic ingredients for building statistical word and phrase alignment models. However, with the increasing availability of parallel text, human-created alignments are expensive and often unaffordable for practical systems, even at a small scale. High quality automatic alignment of parallel text has therefore become indispensable. In addition to good alignment quality, several other properties are also desirable in automatic alignment systems. Ideally, these should be general-purpose and language independent, capable of aligning very different languages, such as English, French, Chinese, German and Arabic, to give a few examples of current interest. If the alignment system is based on statistical models, the model parameters should be estimated from scratch, in an unsupervised manner from whatever parallel text is available. To process millions of sentence pairs, these models need to be capable of generalization and the alignment and estimation algorithms should be computationally efficient. Finally, since noisy mismatched text is often found in real data, such as parallel text mined from web pages, automatic alignment needs to be robust. There are systems available for these purposes, notably the GIZA++ (Och and Ney, 2003) toolkit and

Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 265–268,
New York City, June 2006.

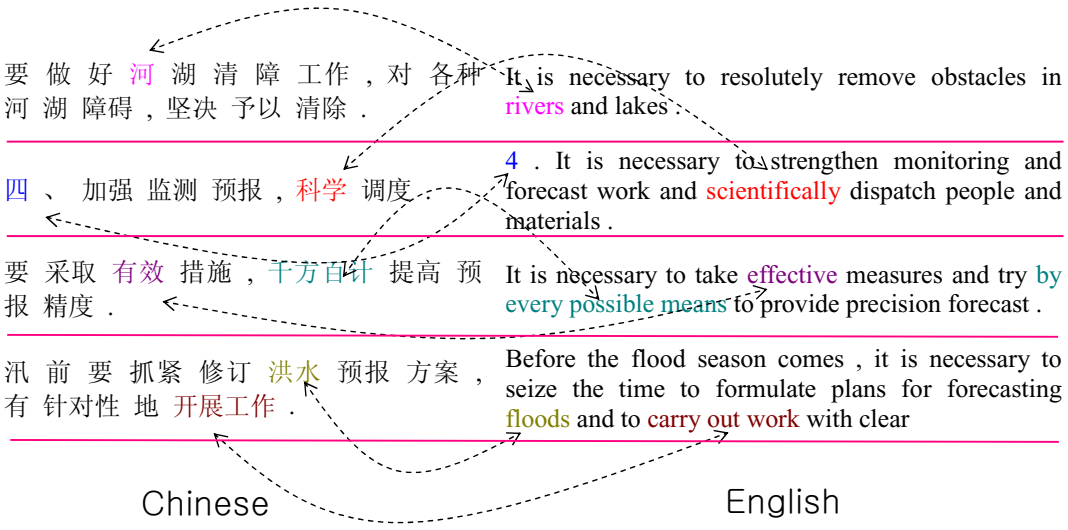

Figure 1: Chinese/English Parallel Corpus Aligned at the Sentence, Word, and Phrase Levels: horizontal lines denote the segmentations of a sentence alignment and arrows denote a word-level mapping.

the Champollion Toolkit (Ma et al., 2004).

This demo introduces MTTK, the Machine Translation Toolkit. The toolkit can be used to train statistical models and perform parallel text alignment at different levels. Target applications include not only machine translation, but also bilingual lexicon induction, cross lingual information retrieval and other multi-lingual applications.

2 MTTK Components

MTTK is a collection of C++ programs and Perl and shell scripts that can be used to build statistical alignment models from parallel text. Respective of the text to be aligned, MTTK's functions are categorized into the following two main parts.

2.1 Chunk Alignment

Chunk alignment aims to extract sentence or sub-sentence pairs from parallel corpora. A chunk can be multiple sentences, a sentence or a sub-sentence, as required by the application. Two alignment procedures are implemented: one is the widely used dynamic programming procedure that derives monotone alignment of sentence segments (Gale and Church, 1991); the other is divisive clustering procedure that begins by finding coarse alignments that are then iteratively refined by successive binary splitting (Deng et al.,). These two types of alignment procedures complement each other. They can be used together to improve the overall sentence alignment quality.

When translation lexicons are not available, chunk alignment can be performed using length-based statistics. This usually can serve as a starting point of sentence alignment. Alignment quality can be further improved when the chunking procedure is based on translation lexicons from IBM Model-1 alignment model (Brown et al., 1993). The MTTK toolkit also generates alignment score for each chunk pair, that can be utilized in post processing, for example in filtering out aligned segments of dubious quality.

2.2 Word and Phrase Alignment

After a collection of sentence or sub-sentence pairs are extracted via chunk alignment procedures, statistical word and phrase alignment models can be estimated with EM algorithms. MTTK provides implementations of various alignment, models including IBM Model-1, Model-2 (Brown et al., 1993), HMM-based word-to-word alignment model (Vogel et al., 1996; Och and Ney, 2003) and HMM-based word-to-phrase alignment model (Deng and Byrne, 2005). After model parameters are estimated, the Viterbi word alignments can be derived. A novel computation performed by MTTK is the genera-

tion of model-based phrase pair posterior distributions (Deng and Byrne, 2005), which plays an important role in extracting a phrase-to-phrase translation probabilities.

3 MTTK Features

MTTK is designed to process huge amounts of parallel text. Model parameter estimation can be carried out parallel during EM training using multiple CPUs. The entire parallel text is split into parts. During each E-step , statistics are collected parallel over each part, while in the M-steps, these statistics are merged together to update model parameters for next iteration. This parallel implementation not only reduces model training time significantly, it also avoids memory usage issues that arise in processing millions of sentence pairs, since each E-Step need only save and process co-occurrence that appears in its part of the parallel text. This enables building a single model from many millions of sentence pairs.

Another feature of MTTK is language independence. Linguistic knowledge is not required during model training, although when it is available, performance can be improved. Statistical parameters are estimated and learned automatically from data in an unsupervised way. To accommodate language diversity, there are several parameters in MTTK that can be tuned for individual applications to optimize performance.

4 A Typical Application of MTTK in Parallel Text Alignment

A typical example of using MTTK is give in Figure 2. It starts with a collection of document pairs. During pre-processing, documents are normalized and tokenized into token sequences. This preprocessing is carried out before using the MTTK, and is usually language dependent, requiring, for example, segmenting Chinese characters into words or applying morphological analyzing to Arabic word sequences.

Statistical models are then built from scratch. Chunk alignment begins with length statistics that can be simply obtained by counting the number of tokens on in each language. The chunk aligning procedure then applies dynamic programming to derive a sentence alignment. After sorting the generated sentence pairs by their probabilities, high quality sentence pairs are then selected and used to train a translation lexicon. As an input for next round chunk alignment, more and better sentence pairs can be extracted and serve as training material for a better translation lexicon. This bootstrapping procedure identifies high quality sentence pairs in an iterative fashion.

To maximize the number of training words for building word and phrase alignment models, long sentence pairs are then processed further using a divisive clustering chunk procedure that derives chunk pairs at the sub-sentence level. This provides additional translation training pairs that would otherwise be discarded as being overly long.

Once all usable chunk pairs are identified in the chunk alignment procedure, word alignment model training starts with IBM Model-1. Model complexity increases gradually to Model-2, and then HMM-based word-to-word alignment model, and finally to HMM-based word-to-phrase alignment model (Deng and Byrne, 2005). With these models, word alignments can be obtained using the Viterbi algorithm, and phrase pair posterior distributions can be computed in building a phrase translation table.

In published experiments we have found that MTTK generates alignments of quality comparable to those generated by GIZA++, where alignment quality is measured both directly in terms of Alignment Error Rate relative to human word alignments and indirectly through the translation performance of systems constructed from the alignments (Deng and Byrne, 2005). We have used MTTK as the basis of translation systems entered into the recent NIST Arabic-English and Chinese-English MT Evaluations as well as the TC-STAR Chinese-English MT evaluation (NIST, 2005; TC-STAR, 2005).

5 Availability

MTTK Version 1.0 is released under the Open Source Educational Community License[1]. The tools and documentation are available at http://mi.eng.cam.ac.uk/∼wjb31/distrib/mttkv1/ .

[1] http://www.opensource.org/licenses/ecl1.php

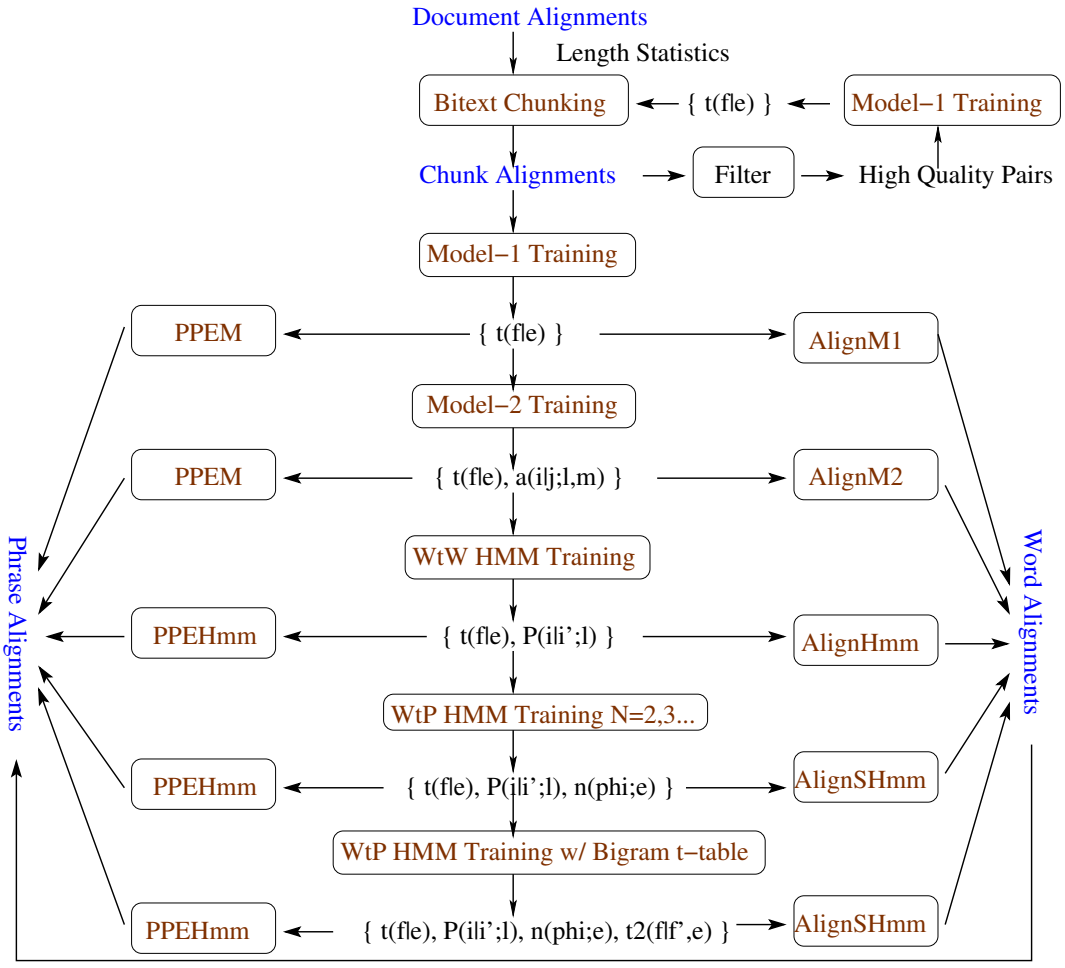

Figure 2: A Typical Unsupervised Translation Alignment Procedure with MTTK.

6 Acknowledgements

Funded by ONR MURI Grant N00014-01-1-0685.

References

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19:263–312.

Y. Deng and W. Byrne. 2005. Hmm word and phrase alignment for statistical machine translation. In *Proc. of HLT-EMNLP*.

Y. Deng, S. Kumar, and W. Byrne. Segmentation and alignment of parallel text for statistical machine translation. *Journal of Natural Language Engineering*. to appear.

W. A. Gale and K. W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184.

X. Ma, C. Cieri, and D. Miller. 2004. Corpora & tools for machine translation. In *Machine Translation Evaluation Workshop*, Alexandria, VA. NIST.

NIST, 2005. *The NIST Machine Translation Evaluations Workshop*. North Bethesda, MD, June. http://www.nist.gov/speech/tests/summaries/2005/mt05.htm.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

TC-STAR, 2005. *TC-STAR Speech-to-Speech Translation Evaluation Meeting*. Trento, Italy, April. http://www.tc-star.org/.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM based word alignment in statistical translation. In *Proc. of the COLING*.