

# Large-data Statistical Machine Translation with Hadoop

Chris Dyer (redpony@umd.edu)  
Dept. of Linguistics, University of Maryland  
October 25, 2007

## 1. Problem

Modern statistical machine translation (SMT) is driven by large quantities of aligned bilingual sentence pairs (so-called bitexts), from which translation models are automatically learned. I propose to develop a framework to reduce the effort involved in using extremely large quantities of training data to develop SMT systems. This task decomposes into several sub-problems, which can be addressed independently: generation of word alignments, estimation of a translation model, estimation of a language model, decoding a tuning set with the estimated models (this requires efficient access to models of potentially very large size), optimization of model parameters according to some loss function, and decoding of an evaluation set with the estimated model.

Currently, the research community deals with large data in three ways. First, some solutions for efficiently handling large amounts of training data have been developed, for example, in the domain of language model estimation and representation [1,2]. However, since cluster architectures are quite diverse, these solutions, if publicly available at all, tend to be ad-hoc and environment-dependent. Second, and more commonly, non-parallel implementations of SMT model estimators (such as GIZA++ and the Moses training suite) are applied to large data sets resulting in extremely long experiment run-times, which limits the kinds of experiments that can be run. Finally, many researchers circumvent these problems entirely by using small corpora. The use of small corpora for research is so widespread that many papers draw conclusions from systems trained on orders of magnitude less training data than is actually available (e.g., [3,4]).

The first phase of this project will focus on more efficient translation model estimation since this task is particularly well suited for Hadoop and because there currently is no available distributed solution to this problem. Improvements to word alignment will also be investigated.

## 2. Resources

The basis of this project will be the Moses decoder tool suite (<http://www.statmt.org/moses>), an open-source toolkit that is widely used in the MT community. However, it was not designed for clusters, and hence suffers from the problems discussed above. The LDC provides bitexts in various languages (e.g., Chinese-English), which will provide training data for our experiments.

## 3. The MapReduce Perspective

Translation models specify the probability of translating a phrase in the source language ( $f$ ) into a phrase in the target language ( $e$ ) and vice-versa. The probabilities represented generally include  $P(e|f)$ ,  $P(f|e)$ ,  $P_{lex}(f|e)$ , and  $P_{lex}(e|f)$ , where  $P_{lex}$  is the lexical translation problem (given the maximal word alignment, what is the word-by-word translation probability). Hadoop is well suited to

estimating these probabilities since it can easily compute the phrase counts and marginal counts necessary to compute the maximum likelihood estimate (MLE).

Word alignment is another resource-intensive problem for which no non-trivially parallel solution is available. Although models of considerable complexity have been proposed, an HMM alignment model bootstrapped with IBM Model 1 probabilities performs exceptionally well in an MT task [5,6]. Each iteration of the EM algorithm that is used to train the HMM model can be conceived of as a MapReduce task. The map step applies the training data to the existing model (E step, part 1), and the reduce step computes aggregate counts for each state and emission (E step, part 2) and normalizes (M step). The map process can be partitioned by sentence, and the reduce process can be partitioned on the states in the HMM, suggesting that this mapping can take full advantage of a large cluster.

## 4. Possible Extensions

The MLE estimate for translation probabilities is simple to compute, but it has been demonstrated that smoothing techniques can result in better performance [7]. The task of estimating smoothed probability models is further complicated by the restrictions that MapReduce imposes; however, a systematic exploration of this space would be revealing and useful for domains beyond MT, such as language modeling.

Another possible extension would be to reproduce experimental results from the literature that were carried out on small data sets (e.g., [3]) and see how well they generalize to much larger amounts of training data.

## 5. References

- [1] Federico, M., Cettolo, R. (2007). Efficient handling of n-gram language models for SMT. In *WMT 2007*.
- [2] Talbot, D., Osborne, M. (2007). Smoothed Bloom filter language models: Tera-scale LMs on the Cheap. In *EMNLP/CoNLL 2007*.
- [3] Ma, Y., Way, A. (2007). Bootstrapping word alignment via word packing. In *ACL, 2007*.
- [4] Venugopal, A., Zollman, A., Vogel, S. (2007). An efficient two-pass approach to synchronous-CFG driven statistical MT. In *NAACL/HLT 2007*.
- [5] Och, F., Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*.
- [6] Lopez, A., Resnik, P. (2006). Word-based alignment, phrase-based translation: What's the link? In *AMTA 2006*.
- [7] Foster, G., Kuhn, R., Johnson, J.H. (2006) Phrasetable smoothing for SMT. In *EMNLP 2006*

