

Lattice Rescoring Methods for Statistical Machine Translation

Graeme Blackwood

July 24, 2010

Summary

Modern statistical machine translation (SMT) systems include multiple interrelated components, statistical models, and processes. Translation is often factored as a cascaded series of modules such that the output of one module serves as the input to the next; this is the SMT pipeline. Simplifying assumptions, limited training data, and pruning during search mean that the hypothesis produced by a typical SMT decoder may not represent the best translation. Since any errors will be propagated through the SMT pipeline, it is better to avoid hard decisions by passing on as much information as possible to subsequent modules. The focus, then, is less on finding the single-best translation and more on being able to generate a rich space of likely translations that can be exploited through subsequent rescoring and combination techniques. The large size of the search space in SMT means that it is not always possible to apply more complex models in translation decoding; such models are normally applied to a translation lattice, a space efficient representation of many translation alternatives with scores.

This thesis develops a robust inventory of large-scale lattice rescoring methods that improve the quality of statistical machine translation. These rescoring methods include (i) sentence-specific, high-order language models estimated over multi-billion word corpora, (ii) stochastic segmentation transducers that model the phrasal segmentation process in phrase-based SMT, (iii) efficient large-scale lattice minimum Bayes-risk decoding procedures based on weighted path counting transducers, (iv) multi-input and multi-source lattice combination techniques that synthesise multiple sources of translation knowledge, and (v) a novel decoding framework based on segmentation of a word lattice into regions of high and low confidence that supports targeted application of modelling techniques intended to address particular deficiencies in translation. Efficient realisations of these lattice rescoring methods are described in terms of general purpose weighted finite state transducer operations.

A second theme of this thesis concerns the exploitation of monolingual corpora. Although monolingual data is much more widely available than parallel data, in SMT it is typically only used for building word-based language models. However, there are other complementary ways in which this data can be used to improve translation quality. Two novel lattice rescoring methods for exploiting monolingual corpora - phrasal segmentation models that learn the segmentation of sequences of words into sequences of translatable phrases, and monolingual coverage constraints that address the often overlooked issue of machine translation fluency - are proposed in this thesis.