Models of Synchronous Grammar Induction for SMT

Phil Blunsom
Phil.Blunsom@comlab.ox.ac.uk
University of Oxford

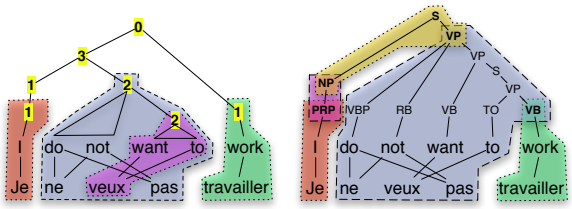
The last decade of research in Statistical Machine Translation (SMT) has seen rapid progress. Unfortunately these successes have not been uniform; closely related language pairs can be translated with a high degree of precision, while for distant pairs the result is far from acceptable.

Models which have been most successful for translating between structurally divergent language pairs have been based on synchronous grammars. A critical component of these translation models is their *grammar* which encodes translational equivalence and licenses reordering between tokens in the source and target languages. There is considerable scope for improving beyond current techniques for automatically acquiring synchronous grammars from bilingual corpora, which seek to find either extremely simple grammars with only one non-terminal or else rely on treebank-trained parsers. The simple grammars are incapable of representing the substitutability of a constituent, while the richer grammars limit the systems' portability to new target languages (effectively limiting us to translating into/out of English) while enforcing a restrictive notion of linguistic constituency (Figure 1).

Clearly there is a need for research into the unsupervised induction of synchronous grammar based translation models. We propose the pragmatic approach of embracing existing algorithms for inducing unlabelled SCFGs (e.g. the popular Hiero model), and then using state-of-the-art hierarchical non-parametric Bayesian models to independently learn syntactic classes for translation rules in the grammar.

The agenda of the workshop will address two goals: (1) to implement and systematically investigate the performance of current synchronous grammar based SMT systems, focusing on the role of constituency and syntactic classes in informing translation structure; (2) to develop scalable unsupervised algorithms for assigning labels to translation rules in synchronous grammars. These algorithms will be implemented within the *Joshua* decoder with the intention of providing an open source implementation of the existing and proposed synchronous grammar SMT systems.

1) Systematic comparison of synchronous grammar based SMT Firstly we will extend the Joshua decoder to handle a range of the currently proposed synchronous grammar based SMT models. Through experimentation on a range of parallel corpora (small and large, hi and low density languages), we will systematically explore the question of what the most effective models of synchronous syntax for SMT are.



(a) Unsupervised syntax (b) Treebank syntax

Figure 1. What is the best syntactic structure for SMT?

2) Unsupervised learning of labelled synchronous grammars The second goal will be to apply unsupervised Bayesian techniques, such as Latent Dirichlet Allocation (LDA) and hierarchical non-parametric models (HDPs, HPYPs), to the task of assigning equivalence classes to phrase translations. Inspired by work in monolingual PCFG learning, we will investigate generative models which describe the production of phrase translations in terms of sequences of tokens (or word classes) and their observed contexts.

We have put together an enviable superset of faculty and graduate students from which the senior members of the workshop will be drawn: Joy Ying Zhang (CMU), Trevor Cohn (Sheffield), Alex Clark (RHUL), Chris Dyer (UMD), Adam Lopez (Edinburgh), Yang Liu (CAS), Zhifei Li (JHU) and Andreas Zollmann (CMU).

The success of the workshop will impact widely on both the machine translation community, and the field of grammar induction. By investigating our stated goals in the context of a CLSP workshop we will provide a deep understanding of synchronous grammar based SMT; both generating interest and furthering research into widely applicable unsupervised techniques for synchronous grammar induction. Such techniques promise to bring high performance SMT models, only currently applicable to working with English, to the full range of languages of the world. In addition we will provide a benchmark implementation of synchronous grammar based SMT models ready for wide adoption within the SMT research community.