Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation

John Tinsley, Mary Hearne, and Andy Way

National Centre for Language Technology
Dublin City University, Ireland
{*jtinsley, mhearne, away*}*@computing.dcu.ie*

Abstract. Given much recent discussion and the shift in focus of the field, it is becoming apparent that the incorporation of syntax is the way forward for the current state-of-the-art in machine translation (MT). Parallel treebanks are a relatively recent innovation and appear to be ideal candidates for MT training material. However, until recently there has been no other means to build them than by hand. In this paper, we describe how we make use of new tools to automatically build a large parallel treebank and extract a set of linguistically motivated phrase pairs from it. We show that adding these phrase pairs to the translation model of a baseline phrase-based statistical MT (PBSMT) system leads to significant improvements in translation quality. We describe further experiments on incorporating parallel treebank information into PBSMT, such as word alignments. We investigate the conditions under which the incorporation of parallel treebank data performs optimally. Finally, we discuss the potential of parallel treebanks in other paradigms of MT.

1 Introduction

The majority of research in recent years in machine translation (MT) has centred around the phrase-based statistical approach. This paradigm involves translating by training models which make use of sequences of words, so-called phrase pairs, as the core translation model of the system [1]. These phrase pairs are extracted from aligned sentence pairs using heuristics over a statistical word alignment. While phrase-based models have achieved state-of-the-art translation quality, evidence suggests there is a limit as to what can be accomplished using only simple phrases, for example, satisfactory capturing of context-sensitive reordering phenomena between language pairs [2]. This assertion has been acknowledged within the field as illustrated by the recent shift in focus towards more linguistically motivated models.

Aside from the development of fully syntax-based models of MT, [3–6] to list a few, there have been many extensions and improvements to the phrase-based model which have endeavoured to incorporate linguistic information into the translation process. Examples of these can be seen in the work of [7] and [8] who make use of syntactic supertags and morphological information respectively. [9, 10] describes a phrase-based model which makes use of generalised templates while [11] exploit semantic information in the form of phrase-sense disambiguation. All of these approaches have a