Stat-XFER: A General Search-based Syntax-driven Framework for Machine Translation

Alon Lavie

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
alavie@cs.cmu.edu.edu

Abstract. The CMU Statistical Transfer Framework (Stat-XFER) is a general framework for developing search-based syntax-driven machine translation (MT) systems. The framework consists of an underlying syntax-based transfer formalism along with a collection of software components designed to facilitate the development of a broad range of MT research systems. The main components are a general language-independent run-time transfer engine and decoder, along with several different tools for creating the various underlying language-pair-specific resources that are required for building a specific MT system for any given language pair. We describe the general framework, its unique properties and features, and its application to the construction of MT research prototype systems for a diverse collection of language pairs.

1 Introduction

The field of Machine Translation (MT) has dramatically shifted in the course of the past decade. Modern state-of-the-art approaches to MT rely on machine learning methods of increasing complexity and sophistication in order to automatically acquire their underlying translation models from available data resources. Phrase-based Statistical MT (PB-SMT) [1–3] has become the predominant approach in recent years. In PB-SMT, simple statistical modeling methods are used to acquire likely phrase-to-phrase translation equivalents from large volumes of sentence-parallel text corpora. In the absence of large sentence-parallel data, the statistical estimation methods break down, and the approach becomes ineffective. Vast sentence-parallel corpora exist only for a limited number of language pairs (primarily pairs of European languages, Chinese, Japanese and Arabic), severely limiting the applicability of this approach. While the amount of online resources for many languages will undoubtedly grow over time, many of the languages spoken by smaller ethnic groups and populations in the world will not have such resources within the foreseeable future. Corpus-based MT approaches will therefore not be effective for such languages for some time to come.