Statistical Machine Translation Part I: Hands-On Introduction

Stephan VOGEL
InterACT, LTI
Carnegie Mellon University
407 South Craig Street, Pittsburgh, PA 15213
stephan.vogel@cs.cmu.edu

Abstract

Statistical machine translation (SMT) is currently one of the hot spots in natural language processing. Over the last few years dramatic improvements have been made, and a number of comparative evaluations have shown, that SMT gives competitive results to rule-based translation systems, requiring significantly less development time. This is particularly important when building translation systems for new language pairs or new domains.

This workshop is intended to give an introduction to statistical machine translation with a focus on practical considerations. Participants should be able, after attending this workshop, to set out building an SMT system themselves and achieving good baseline results in a short time.

The tutorial will cover the basics of SMT:

- architecture of an SMT system
- word alignment models, esp. IBM1 and HMM models
- phrase alignment, from Viterbi path and direct phrase alignment models
- decoder, including recombination, pruning, n-best list generation
- integrating output from other MT engines (multi engine translation)
- data processing: checking, cleaning, normalizing the data
- evaluation, especially automatic evaluation (Bleu, NIST, ...), including significance analysis

Theory will be put into practice. STTK, a statistical machine translation tool kit, will be introduced and used to build a working translation system. STTK has been developed by the presenter and co-workers over a number of years and is currently used as the basis of CMU's SMT system. It has also successfully been coupled with rule-based and example based machine translation modules to build a multi engine machine translation system. The source code of the tool kit will be made available.

Biography

Stephan Vogel is research scientist at the Language Technologies Institute, Carnegie Mellon University. He is also affiliated to InterACT, the International Center for Advanced Communication Technologies, a joint center between the University of Karlsruhe, Germany, and Carnegie Mellon University. After receiving a MSc in Physics from Philips University of Marburg, and a MPhil in History and Philosophy of Science from the University of Cambridge, England, he did his doctoral studies at the University of Aachen (RWTH), where he started to work on Statistical Machine Translation. This remained the focus of his research. Since he joined CMU in 2001 he built a SMT research team, which now consists of more then 10 PhD and Master students, working on a number of text and speech translation projects.