Topics in Statistical Machine Translation

Kevin Knight
Information Sciences Institute
University of Southern California
knight@isi.edu

Philipp Koehn
School of Informatics
University of Edinburgh
pkoehn@inf.ed.ac.uk

1 Introduction

In the past, we presented tutorials called "Introduction to Statistical Machine Translation", aimed at people who know little or nothing about the field and want to get acquainted with the basic concepts. This tutorial, by contrast, goes more deeply into selected topics of intense current interest. We aim at two types of participants:

1. People who understand the basic idea of statistical machine translation and want to get a survey of hot-topic current research, in terms that they can understand.
2. People associated with statistical machine translation work, who have not had time to study the most current topics in depth.

We fill the gap between the introductory tutorials that have gone before and the detailed scientific papers presented at ACL sessions.

2 Tutorial Outline

Below is our tutorial structure. We showcase the intuitions behind the algorithms and give examples of how they work on sample data. Our selection of topics focuses on techniques that deliver proven gains in translation accuracy, and we supply empirical results from the literature.

1. QUICK REVIEW (15 minutes)
   - Phrase-based and syntax-based MT.
2. ALGORITHMS (45 minutes)
   - Efficient decoding for phrase-based and syntax-based MT (cube pruning, forward/outside costs).
   - Minimum-Bayes risk.
   - System combination.
3. SCALING TO LARGE DATA (30 minutes)
   - Phrase table pruning, storage, suffix arrays.
   - Large language models (distributed LMs, noisy LMs).
4. NEW MODELS (1 hour and 10 minutes)
   - New methods for word alignment (beyond GIZA++).
   - Factored models.
   - Maximum entropy models for rule selection and re-ordering.
   - Acquisition of syntactic translation rules.
   - Syntax-based language models and target-language dependencies.
   - Lattices for encoding source-language uncertainties.
5. LEARNING TECHNIQUES (20 minutes)
   - Discriminative training (perceptron, MIRA).

Tutorial Abstracts of ACL-IJCNLP 2009, page 2,
Suntec, Singapore, 2 August 2009. ©2009 ACL and AFNLP