Morphological, Syntactical and Semantic Knowledge in Statistical Machine Translation

Marta R. Costa-jussà*, Chris Quirk§
*Institute for Infocomm Research
§Microsoft Research
martaruizcostajussa@gmail.com
chrisq@microsoft.com

1 Overview

This tutorial focuses on how morphology, syntax and semantics may be introduced into a standard phrase-based statistical machine translation system with techniques such as machine learning, parsing and word sense disambiguation, among others.

Regarding the phrase-based system, we will describe only the key theory behind it. The main challenges of this approach are that the output contains unknown words, wrong word orders and non-adequate translated words. To solve these challenges, recent research enhances the standard system using morphology, syntax and semantics.

Morphologically-rich languages have many different surface forms, even though the stem of a word may be the same. This leads to rapid vocabulary growth, as various prefixes and suffixes can combine with stems in a large number of possible combinations. Language model probability estimation is less robust because many more word forms occur rarely in the data. This morphologically-induced sparsity can be reduced by incorporating morphological information into the SMT system. We will describe the three most common solutions to face morphology: preprocessing the data so that the input language more closely resembles the output language; using additional language models that introduce morphological information; and post-processing the output to add proper inflections.

Syntax differences between the source and target language may lead to significant differences in the relative word order of translated words. Standard phrase-based SMT systems surmount reordering/syntactic challenges by learning from data. Most approaches model reordering inside translation units and using statistical methodologies, which limits the performance in language pairs with different grammatical structures. We will briefly introduce some recent advances in

Tutorials, NAACL-HLT 2013, pages 16–18,
Atlanta, Georgia, June 9 2013. ©2013 Association for Computational Linguistics

SMT that use modeling approaches based on principles more powerful flat phrases and better suited to the hierarchical structures of language: SMT decoding with stochastic synchronous context free grammars and syntax-driven translation models.

Finally, semantics are not directly included in the SMT core algorithm, which means that challenges such as polysemy or synonymy are either learned directly from data or they are incorrectly translated. We will focus on recent attempts to introduce semantics into statistical-based systems by using source context information.

The course material will be suitable both for attendees with limited knowledge of the field, and for researchers already familiar with SMT who wish to learn about modern tendencies in hybrid SMT. The mathematical content of the course include probability and simple machine learning, so reasonable knowledge of statistics and mathematics is required. There will be a small amount of linguistics and ideas from natural language processing.

2 Outline

1. Statistical Machine Translation
    - Introduction to Machine Translation approaches
    - Phrase-based systems
2. Morphology in SMT
    - Types of languages in terms of morphology
    - Enriching source language
    - Inflection generation
    - Class-based language models
3. Syntax in SMT
4. Semantics in SMT
    - Sense disambiguation
    - Context-dependent translations

3 Speaker Bios

Marta R. Costa-jussà[1], Institute for Infocomm Research (I2R), is a Telecommunication's Engineer by the Universitat Politècnica de Catalunya (UPC, Barcelona) and she received her PhD from the UPC in 2008. Her research experience is mainly in Automatic Speech Recognition, Machine Translation and Information Retrieval. She has worked at LIMSI-CNRS (Paris), Barcelona Media Innovation Center (Barcelona) and the Universidade de Sao Paulo (São Paulo). Since December 2012 she is working at Institute for Infocomm Research (Singapore) implementing the IMTraP project ("Integration of Machine Translation Paradigms") on Hybrid Machine Translation, funded by the European Marie Curie International Outgoing European Fellowship program. She is currently organizing the ACL Workshop HyTRA 2013 and she will be teaching a summer school course on hybrid machine translation at ESSLLI 2013.

Chris Quirk[2], Microsoft Research. After studying Computer Science and Mathematics at Carnegie Mellon University, Chris joined Microsoft in 2000 to work on the Intentional Programming project, an extensible compiler and development framework. He moved to the Natural Language Processing group in 2001, where his research has mostly focused on statistical machine translation powering Microsoft Translator, especially on several generations of a syntax directed translation system that powers over half of the translation systems. He is also interested in semantic parsing, paraphrase methods, and very practical problems such as spelling correction and transliteration.

[1] http://www.costa-jussa.com

[2] http://research.microsoft.com/en-us/people/chrisq/