Tree-based and Forest-based Translation

Yang Liu
Institute of Computing Technology
Chinese Academy of Sciences
yliu@ict.ac.cn

Liang Huang
Information Sciences Institute
University of Southern California
lhuang@isi.edu

1 Introduction

The past several years have witnessed rapid advances in syntax-based machine translation, which exploits natural language syntax to guide translation. Depending on the type of input, most of these efforts can be divided into two broad categories: (a) **string-based systems** whose input is a string, which is simultaneously parsed and translated by a synchronous grammar (Wu, 1997; Chiang, 2005; Galley et al., 2006), and (b) **tree-based systems** whose input is already a parse tree to be directly converted into a target tree or string (Lin, 2004; Ding and Palmer, 2005; Quirk et al., 2005; Liu et al., 2006; Huang et al., 2006).

Compared with their string-based counterparts, tree-based systems offer many attractive features: they are much faster in decoding (linear time vs. cubic time), do not require sophisticated binarization (Zhang et al., 2006), and can use separate grammars for parsing and translation (e.g. a context-free grammar for the former and a tree substitution grammar for the latter).

However, despite these advantages, most tree-based systems suffer from a major drawback: they only use 1-best parse trees to direct translation, which potentially introduces translation mistakes due to parsing errors (Quirk and Corston-Oliver, 2006). This situation becomes worse for resource-poor source languages without enough Treebank data to train a high-accuracy parser.

This problem can be alleviated elegantly by using packed forests (Huang, 2008), which encodes exponentially many parse trees in a polynomial space. Forest-based systems (Mi et al., 2008; Mi and Huang, 2008) thus take a packed forest instead of a parse tree as an input. In addition, packed forests could also be used for translation rule extraction, which helps alleviate the propagation of parsing errors into rule set. Forest-based translation can be regarded as a compromise between the string-based and tree-based methods, while combining the advantages of both: decoding is still fast, yet does not commit to a single parse. Surprisingly, translating a forest of millions of trees is even faster than translating 30 individual trees, and offers significantly better translation quality. This approach has since become a popular topic.

2 Content Overview

This tutorial surveys tree-based and forest-based translation methods. For each approach, we will discuss the two fundamental tasks: *decoding*, which performs the actual translation, and *rule extraction*, which learns translation rules from real-world data automatically. Finally, we will introduce some more recent developments to tree-based and forest-based translation, such as tree sequence based models, tree-to-tree models, joint parsing and translation, and faster decoding algorithms. We will conclude our talk by pointing out some directions for future work.

3 Tutorial Overview

1. Tree-based Translation
   - Motivations and Overview
   - Tree-to-String Model and Decoding
   - Tree-to-String Rule Extraction
   - Language Model-Integrated Decoding: Cube Pruning
2. Forest-based Translation
   - Packed Forest
   - Forest-based Decoding
   - Forest-based Rule Extraction
3. Extensions
   - Tree-Sequence-to-String Models
   - Tree-to-Tree Models
   - Joint Parsing and Translation
   - Faster Decoding Methods
4. Conclusion and Open Problems

Tutorial Abstracts of ACL 2010, page 2,
Uppsala, Sweden, 11 July 2010.