JAIST Repository
https://dspace.jaist.ac.jp/

Title	Rule selection for syntax-based Vietnamese-English statistical machine translation
Author(s)	Bui, Thanh Hung
Citation	
Issue Date	2010-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/9143
Rights	
Description	Supervisor:Professor Akira Shimazu, 情報科学研究科, 修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Rule selection for syntax-based Vietnamese-English statistical machine translation

Bui Hung Thanh (0810207)
School of Information Science
Japan Advanced Institute of Science and Technology
August 10, 2010

Keywords: Syntax-based SMT, Hierarchical phrase-based model, Rule selection, Maximum entropy-based rule selection

Abstract

The syntax-based statistical machine translation model uses rules with hierarchical structures as translation knowledge, which can capture long-distance reorderings. Typically, a translation rule consists of a source side and a target side. However, the source side of a rule usually corresponds to multiple target-sides in multiple rules. Therefore, during decoding, the decoder should select the correct target-side for a given source side. This is rule selection.

Rule selection is of great importance to syntax-based statistical machine translation systems. This is because a rule contains not only terminals (words or phrases), but also nonterminals and structural information. During decoding, when a rule is selected and applied to a source text, both lexical translations (for terminals) and reorderings (for nonterminals) are determined. Therefore, rule

Copyright © 2010 by Bui Hung Thanh

selection affects both lexical translation and phrase reorderings. However, most of the current syntax-based systems ignore contextual information when they select rules during decoding, especially the information covered by nonterminals. This makes it difficult for the decoder to distinguish rules. Intuitively, information covered by nonterminals as well as contextual information of rules is believed to be helpful for rule selection.

In this work, rule selection for syntax-based Vietnamese-English statistical machine translation, we propose a maximum entropy-based rule selection model for syntax-based statistical machine translation. The maximum entropy-based rule selection model combines local contextual information around rules and information of sub-trees covered by variables in rules. Therefore, our model allows the decoder to perform context-dependent rule selection during decoding. We incorporate the maximum entropy-based rule selection model into a state-of-the-art syntax-based Vietnamese-English statistical machine translation model. Experiments show that our approach achieves significant improvements over the baseline system.

This thesis is organized into three main parts. The first chapter presents the introduction and overview of the thesis. The second and the third chapters summarize the related theories by a literature review, giving a detailed exposition of the theory of statistical machine translation and rule selection for syntax-based statistical machine translation. By discussing the experimental output, the last chapter summarizes this thesis and proposes further work.