自动化学报 2009 35 (10): 1317-1326 ISSN: 0254-4156 CN: 11-2109/TP

论文与报告

基于同步树序列替换文法的统计机器翻译模型

蒋宏飞, 李生, 张民, 赵铁军, 杨沐昀

1. 哈尔滨工业大学计算机科学与技术学院机器智能与翻译研究室 哈尔滨 150001

2. 新加坡信息通讯研究所 新加坡 119613

收稿日期 2008-5-13 修回日期 2009-1-21 网络版发布日期 接受日期

摘要

基于短语的模型是目前发展相对成熟的一种统计机器翻译(Statistical machine translation, SMT)模型. 但基于短语的模型不包含任何结构信息, 因而缺乏有效的全局调序能力, 同时不能对非连续短语进行建模. 基于句法的模型因具有结构信息而具有解决以上问题的潜力, 因而越来越受到研究者们的重视. 然而现有的大多数基于句法的模型都因严格的句法限制而制约了模型的描述能力. 为突破这种限制并将 基于短语的模型的优点融入到句法模型中, 本文提出一种基于同步树序列替换文法(Synchronous tree sequence substitution grammar, STSSG)的统计机器翻译模型. 在此模型中, 树序列被用作为基本的翻译单元. 在这种框架下, 不满足句法限制的翻译等价对和满足句法限制的翻译等价对都可以融入句法信息并被翻译模型所使用. 从而, 两种模型的优点均得到充分利用. 在2005年度美国国家标准与技术研究所(NIST)举办的机器翻译评比的中文翻译任务语料上的实验表明, 本文提出的模型显著地超过了两个基准系统: 基于短语的翻译系统Moses和一个基于严格树结构的句法翻译模型.

关键词 统计机器翻译 句法限制 同步文法 同步树替换文法 同步树序列替换文法

分类号 TP391

Synchronous Tree Sequence Substitution Grammar for Statistical Machine Translation

JIANG Hong-Fei, LI Sheng, ZHANG Min, ZHAO Tie-Jun, YANG Mu-Yun

1. Machine Intelligence and Translation Laboratory, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, P.R. China

2. Institute for Infocomm Research, Singapore 119613, Singapore

Abstract

Phrase-based models are the state-of-the-art statistical machine translation models. However, they can not effectively handle global reordering and discontiguous phrases due to the lack of structural information. While syntax-based models have the potential to attack these problems, they suffer from the strictly syntactic constraints. To address these constraints and integrate the advantages of phrase-based models into syntax-based models, a synchronous tree sequence substitution grammar (STSSG) based statistical machine translation (SMT) model is presented in this paper. This novel model uses the tree sequence as the basic translation unit. Therefore, both the syntactic translation equivalences and the non-syntactic translation equivalences equipped with syntactic information can be utilized in the translation. Experimental results on the NIST 2005 Chinese-English machine translation data-set show that the proposed method achieves significant improvements over baseline methods including a phrasal model, Moses, and a tree-based syntax model.

Key words Statistical machine translation (SMT) syntactic constraint synchronous grammar synchronous tree substitution grammar synchronous tree sequence substitution grammar (STSSG)

DOI: 10.3724/SP.J.1004.2009.01317

通讯作者 蒋宏飞 hfjiang@mtlab.hit.edu.cn

作者个人主页 蒋宏飞; 李生; 张民; 赵铁军; 杨沐昀