

Statistical Machine Translation Part II: Tree-Based SMT

Dekai WU

Human Language Technology Center
Hong Kong University of Science and Technology (HKUST)
Clear Water Bay, Hong Kong
dekai@cs.ust.hk

Abstract

One of the most active and promising areas of statistical machine translation (SMT) research are tree-based SMT approaches. Tree-based SMT has the potential to overcome the weaknesses of early SMT architectures which (a) do not handle long-distance dependencies well, and (b) are underconstrained in that they allow too much flexibility in word reordering.

In this tutorial, we will review the various possible approaches to tree-based SMT, ranging from the original Inversion Transduction Grammar (ITG) models to later models such as alignment templates, dependency models, tree-to-string models, tree-to-tree models, and also probabilistic EBMT models. We will discuss the theoretical relationships between approaches, with critical analysis of their strengths and weaknesses. Within this framework we will survey the emerging comparative results from intriguing new large-scale empirical studies across various language pairs. We will consider what kind of constraints and biases can or should be imposed by models on the variation between unrelated human languages, and how this can facilitate efficient algorithms for a wide range of tasks in machine learning and processing of language. We will consider both scientific and engineering implications, and investigate the potential relationships to cross-language universals.

Biography

Prof. Wu received his PhD in Computer Science from the University of California at Berkeley,

and was a postdoctoral fellow at the University of Toronto (Ontario, Canada) prior to joining HKUST in 1992. He received a BS in Computer Engineering from the University of California at San Diego (Revelle College departmental award, cum laude, Phi Beta Kappa) in 1984 and an Executive MBA from Kellogg and HKUST in 2002. He has been a visiting researcher at Columbia University in 1995-96, Bell Laboratories in 1995, and the Technische Universität München (Munich, Germany) during 1986-87. Prof. Wu serves as Associate Editor of ACM Transactions on Speech and Language Processing, Machine Translation, Journal of Natural Language Engineering, and Communications of COLIPS. He has also served as Co-Chair for EMNLP-2004, on the Editorial Board of Computational Linguistics, the Organizing Committee of ACL-2000 and WVLC-5 (SIGDAT 1997), and the Executive Committee of the Association for Computational Linguistics (ACL).