Data Selection for Machine Translation Domain Adaptation

陈博兴
Research Officer
National Research Council Canada

摘要：

In this talk, we will first give an overview of research on data selection for machine translation domain adaptation. Then, we will introduce a recently proposed method which uses semi-supervised convolutional neural networks (CNNs) to select in-domain training data. This approach is particularly effective when only tiny amounts of in-domain data are available, which makes fine-grained topic-dependent translation adaptation possible. This method performs significantly better than several state-of-the-art data selection methods on several public domain test sets. Finally, we will talk about the ongoing work which extends the CNN-based method to select in-domain data with good translation quality.

简历：

Boxing Chen is a Research Officer at the National Research Council Canada (NRC). He works on natural language processing, mainly focus on machine translation. Prior to NRC, he was a Senior Research Fellow at the Institute for Infocomm Research in Singapore, a Postdoc at FBK-IRST in Italy and a Postdoc at the University of Grenoble in France. He received his PhD degree from Chinese Academy of Science in 2003. He has co-authered more than 40 papers in NLP conferences and journals. His teams ranked the first place in the NIST 2012 OpenMT Chinese-to-English translation, the first place in the IWSLT 2007 and 2005 Chinese-to-English spoken language translation evaluation.

homepage: https://sites.google.com/site/chenboxing/