PBML

logo

The Prague Bulletin of Mathematical Linguistics

NUMBER ??? MARCH 2015 1–16

Domain Adaptation for Machine Translation with Instance Selection

Ergun Biçici

ADAPT CNGL Centre for Global Intelligent Content
School of Computing
Dublin City University

Abstract

Domain adaptation for machine translation (MT) can be achieved by selecting training instances close to the test set from a larger set of instances. We consider 7 different domain adaptation strategies and answer 7 research questions, which give us a recipe for domain adaptation in MT. We perform English to German statistical MT (SMT) experiments in a setting where test and training sentences can come from different corpora and one of our goals is to learn the parameters of the sampling process. Domain adaptation with training instance selection can obtain 22% increase in target 2-gram recall and can gain up to 3.55 BLEU points compared with random selection. Domain adaptation with feature decay algorithm (FDA) not only achieves the highest target 2-gram recall and BLEU performance but also perfectly learns the test sample distribution parameter with correlation 0.99. Moses SMT systems built with FDA selected 10K training sentences is able to obtain F_1 results as good as the baselines that use up to 2M sentences. Moses SMT systems built with FDA selected 50K training sentences is able to obtain 1 F_1 point better results than the baselines.

1. Introduction

Machine translation (MT) performance is affected by tokens unseen in the training set, which may be due to specific use of vocabulary or grammatical structures observed in the test domain of interest. In this paper, we develop a recipe for domain adaptation for MT by comparing different strategies for the selection of training instances close to the test set from larger sets of in-domain (ID) and out-of-domain (OOD) training data. Each corpus has some characteristic distribution of vocabulary

© 2015 PBML. All rights reserved. Corresponding author: ergun.bicici@computing.dcu.ie
Cite as: Ergun Biçici. Domain Adaptation for Machine Translation with Instance Selection. The Prague Bulletin of Mathematical Linguistics No. ???, 2015, pp. 1–16. doi: HERE-WILL-BE-A-DOI-NUMBER