mtrain: A Convenience Tool for Machine Translation

Samuel Läubli* and Mathias Müller* and Beat Horat and Martin Volk
{læubli,mmueller,horat,volk}@cl.uzh.ch
Institute of Computational Linguistics
University of Zurich

Abstract

We present mtrain, a convenience tool for machine translation. It wraps existing machine translation libraries and scripts to ease their use. mtrain is written purely in Python 3, well-documented, and freely available.1

Machine translation libraries usually focus on core model training, while data preparation and automatic evaluation are left to the user. This presents a barrier to experimental reproducibility, rapid prototyping, and entry to the field from neighbouring disciplines. In the spirit of the Experimental Management System for Moses (Koehn, 2010), our tool is meant to automate these tasks.

mtrain is designed to handle most aspects of a machine translation experiment: it manages preprocessing, model training, and automatic evaluation. Preprocessing involves automatically splitting a data set into training, validation, and test sets; tokenization; casing; byte-pair encoding; and normalization. On top of these standard preprocessing steps, mtrain can also deal with inline XML markup and intelligently transfer XML tags to translations (Müller, 2017).

Our tool provides training automation for statistical phrase-based models with Moses (Koehn et al., 2007) and neural RNN encoder-decoder models with Nematus (Sennrich et al., 2017). After training, mtrain offers automatic evaluation of translation quality. It outputs the well-known BLEU, TER, and METEOR metrics (Clark et al., 2011). Given a folder that contains trained models, the separate component mtrans can be used to translate from files or standard input.

All steps can be configured with config files or command line options, but default settings already lead to functional baseline systems, making it easier for inexperienced users to use the tool. Going forward, we consider wrapping additional machine translation libraries that are native Python 3, such as Sockeye (Hieber et al., 2017).

References


1https://github.com/ZurichNLP/mtrain
*equal contribution

© 2018 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 357
Alacant, Spain, May 2018.