Demo: Graphic-based Statistical Machine Translator

Atul Kr. Ojha, Girish Nath Jha
Jawaharlal Nehru University
New Delhi, India
{shashwatup9k, girishjha}@gmail.com

Abstract

In this demo proposal, we present Graphic-based Statistical Machine Translator (GBSMT). This tool has been developed on Moses with the purpose of visualizing GBSMT. Currently, it provides facility to train, test and evaluate statistical machine translation on phrase-based and factor-based approach for any language pair.

Keywords: Statistical Machine translation, Moses, corpus, automatic evaluation

1. Introduction

In the last two decades the field of MT has witnessed a rapid growth. Presently, researchers, developers, users and commercial organization are following Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) approaches to build their MT systems. Among these two, SMT is most popular because due to its ability to produce better results even on a small corpus as compared to NMT. The latter requires longer training time which further depends upon system configuration. For instance, if the system is trained on GPU-based machine then time taken is less than CPU which can take up to three weeks. People also use SMT because of MOSES – an open source SMT toolkit which gives permission to automatically train translation model for any language pair i.e., English and Hindi with different language model tools (Koehn, 2007).

However, a disadvantage it carries is that one needs to memorize several commands and processes to build any SMT system like: tokenization, filter to long sentences, language model, translation model, tuning and decoding etc. Missing out any of the above mentioned process or typing a wrong command, gets us an error or a bad SMT system. Such problems occur because SMT works only by command line.

Through this work, we attempt to reduce these problems. In this system, there is no need to remember commands because the same toolkit is used internally for the process of visualization which is presented briefly in the next section.

As per our knowledge, Tilde MT[1] is the only other graphical and cloud-based SMT training platform which is based on Moses toolkit. It was developed under the Let's MT Project (Vasiljevs et al., 2012). But it is available at a cost (free 30-day trial) and is mainly focused on European languages. Our training platform is primarily for Indian languages and is available for researchers at no cost. We hope to provide an impetus to researchers working on Indian Languages MT systems.

[1] https://www.tilde.com/products-and-services/machine-translation/free-trial

2. Architecture of GBSMT

Figure1 demonstrates the GBSMT structure, a web server-based application.[2] After logging in users should upload parallel (including source and target language) and monolingual (target language) files in '.txt' or '.xls/.xlsx/.ods' format. The remaining processes, thereafter, are automated described below:

(a) Pre-processing: Here, the system identifies source and target language scripts, match sentences of source and target files. Further, tokenization, extraction of tuning and test file from the parallel corpus, true-casing etc take place.

(b) Creation of Language and Translation model: Here the system creates language model and translation model from monolingual and parallel files respectively.

(c) Tuning: The system prepares tuning model through decoder method.

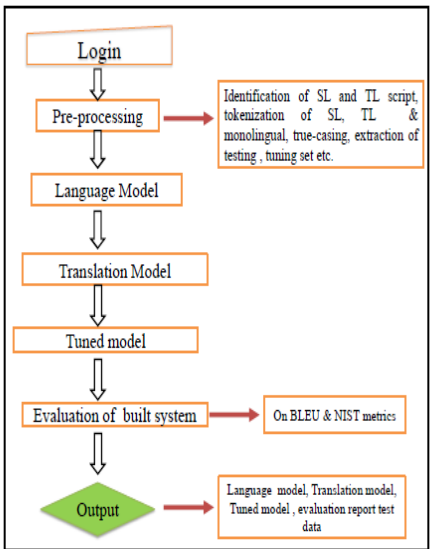

Figure1: Architecture of GBSMT

Once the above processes are over, the system is ready for testing and evaluation. It generates evaluation report from the testing set on BLEU and NIST metrics. Users

[2] http://sanskrit.jnu.ac.in/gbsmt/index.jsp

are then able to download files like Language model, Translation model, Mert file/model, evaluation report etc.

3. Summing up

GBSMT is a tool where users, developers, researchers easily train and build SMT system on window platform. At present people can use this system to build phrase-based statistical machine translation system. But in future, we will include factor-based, automatic evaluate different MT system and other methods to train SMT systems.

4. Reference

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., & Dyer, C. (2007, June). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions (pp. 177-180). Association for Computational Linguistics.

Vasiļjevs, A., Skadiņš, R., & Tiedemann, J. (2012, July). LetsMT!: a cloud-based platform for do-it-yourself machine translation. In Proceedings of the ACL 2012 System Demonstrations (pp. 43-48). Association for Computational Linguistics.