Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782 Start Date of Project: 01/01/2019 Duration: 40 months

Report on **Milestone 18**

Beta version of automatic verification software available for testing

Dissemination Level	PU
Due Date of Milestone	31/12/20 (M24)
Actual Achievement Date	**31/12/20**
Lead Beneficiary/LTP	3. SHARE ERIC
Work Package	WP4 - Innovations in Data Production
Task	Task 4.3 Applying Computer Assisted Translation in Social Surveys
Version	V1.1
Number of Pages	p.1 - p.3

Abstract:

This report documents the availability for testing of the beta version of the Automatic Verification Tool (AVT). The tool enables the user to verify translations using Bilingual Word Embeddings and to report to the translators a set of translated questions to be re-checked.

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020); H2020-INFRAEOSC-04-2018, under the agreement No. 823782

Author List

Organisation	Name	Contact Information
SHARE/MPISOC	Yuri Pettinicchi	pettinicchi@mea.mpisoc.mpg.de

1. Introduction

This report documents the availability of the Automatic Verification Tool (AVT) that is used in the translation research activities of Task 4.3 of the SSHOC project. The task team describes the role of the milestone and the means of verification.

2. Description of the Milestone

The process of translation verification in a complex survey usually takes a third of the time allocated for translation procedures and it costs a person-month for each language for a 90-minutes-long questionnaire. The aim is to preserve high standards and low operational costs. SHARE/MPISOC is developing in-house solutions to make the verification process more efficient. One approach is to use machine translation systems. Unfortunately, this approach requires large parallel corpora, which are expensive to get and/or are not available for some languages and specific domains. Therefore, a word-to-word translation system in an unsupervised manner is selected for the verification task, which allows the researcher to do the task without using parallel corpora.

Based on the unsupervised model approach introduced by Artetxe et al. (2018)[1], the SHARE/MPISOC team trained a Bilingual Word Embeddings model. The team used monolingual corpora in the source language (English) and the target language (German). The team combined more general corpus, e.g., Europarl, News Commentary, which domain specific corpus, e.g. SHARE survey questions.

Role of the Milestone

The role of this milestone is to describe the functionalities of the Automatic Verification Tool (AVT) that make use of the trained model.

Once (English-German) the Bilingual Word Embeddings model is available, 1) the AVT imports the questions and makes use of the trained model and 2) it generates the 10 best foreign language translations of each English word. If one of the best translations is in the human foreign language translation, 3) the AVT marks the

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

word pair as matched. By measuring the number of matched word pairs, 4) the AVT estimates the translation quality. As a last step, 5) the AVT stores the translation quality scores and reports them to the user.

The steps 1-5 are programmed in Python. The milestone was achieved on 31st of December 2020.

2.1. Means of verification

According to the GA, the means of verification of this milestone is the availability of the software for testing. The codes are available online in a GitHub Repository. Prerequisites are Python 3.5 and a set of dependencies enlisted in the file "requirements.txt".[2] A short demo "AVT_demo.webm" shows the installation steps and the main features of AVT.[3]

Conclusions and next steps

The AVT is already being used in translation research activities of Task 4.3. During the translation phase of the SHARE Corona questionnaire (May-June 2020), SHARE/MPISOC made use of a preliminary version of the AVT. SHARE/MPISOC also plans to make use of the beta version of the AVT in the coming translation phase of a second SHARE Corona questionnaire (March-April 2021). The results from these two test runs will provide the ground to compile guidelines as promised in deliverable D*4.11 "Report on the experience with the automatic verification programme in SHARE wave 9"*. The report will include documentation about the selection of the set of items to be verified, libraries used to perform the sanity checks, interaction with MT tools and their performance.

[2] GitHub repository: https://github.com/yichennliu/bwes_translation

[3] A short demo of Automatic Verification Tool (AVT): https://zenodo.org/record/4557069#.YDT35Ooo_Lp