Research and Innovation Action

Social Sciences & Humanities Open Cloud

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

Deliverable 4.6

Guidelines for further use of MT systems in social surveys

Dissemination Level	PU
Due Date of Deliverable	30/04/2022 (M40)
Actual Submission Date	26/04/2022
Work Package	WP4 - Innovations and Data Production
Task	Task 4.2 Preparing tools for the use of Computer Assisted Translation
Type	Report
Approval Status	Approved by EC - 27 April 2022
Version	V1.0
Number of Pages	p.1 – p.14

Abstract:

This report describes a set of guidelines to follow when developing an MT system for the domain of social surveys. Partners describe the whole pipeline, including in-domain data gathering, preparation and augmentation and its impact on the resulting MT system performance. Additionally, partners present a use-case, where the previous baseline (non-adapted) MT models were applied.

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

History

Version	Date	Reason	Revised by
0.1	14/03/2022	Initial Submission	
0.2	22/03/2022	WP4 Review	Diana Zavala-Rojas
1.0	22/03/2022	Addressing Review	Dušan Variš

Author List

Organisation	Name	Contact Information
CLARIN/CUNI	Dušan Variš	varis@ufal.mff.cuni.cz
CLARIN/CUNI	Jan Hajič	hajic@ufal.mff.cuni.cz
WageIndicator	Dani Ceccon	dceccon@gmail.com
UPF	Danielly Sorato	danielly.sorato@upf.edu

Executive Summary

This report describes guidelines that can be applied for training specialized neural machine translation (NMT) systems aimed at translation in a narrow textual domain, namely the domain of social surveys, requiring a specialized MT model that is able to handle domain-specific terminology. The work presented in this report demonstrates how relatively low-resource in-domain corpora can be used to prepare these specialized models. All described models are compatible with the packaged MT framework described in Deliverable D4.5 and the best performing models are available at the Lindat repository. The code used in the training pipeline (for experiment reproduction) is available on GitHub distributed under Mozilla Public License 2.0.[1]

Partners also describe the full translation pipeline including file sharing and preprocessing that was used to help with automatic translation of the Covid-19 surveys into English. While the description of the pipeline is general enough to be used in other future projects, the code published by partners on GitHub serves only as an example of a task-specific solution.

[1] https://www.mozilla.org/en-US/MPL/2.0/ (accessed March 2022)

Abbreviations and Acronyms

BLEU	Bilingual Evaluation Understudy
CLI	Command-line Interface
CUNI	Charles University
EOSC	European Open Science Cloud
ESS	European Social Survey
EVS	European Values Study
MCSQ	Multilingual Corpus of Survey Questionnaires
MT	Machine Translation
NMT	Neural Machine Translation
SHARE	Survey of Health Ageing and Retirement in Europe
SSH	Social Sciences and Humanities
T2T	Tensor2tensor
TRAPD	Translation, Review, Adjudication, Pretesting and, Documentation
WIS	WageIndicator Survey
WMT	Workshop on Machine Translation

Table of Contents

1.Introduction

Automatic translation of social surveys presents a challenge even when faced with the modern MT systems. The main reason behind the unsatisfactory performance of these systems stems mainly from the requirement for a specialized MT that can capture the nuances of the domain of social surveys - most of the contemporary MT are however often trained on a mixture of various domains (e.g., news, movie subtitles, etc.) and often do not even encounter social survey parallel training data or have access only to a limited volume of such data. This lack of data is even more impactful when preparing MT systems based on the state-of-the-art neural machine translation (NMT) approaches which are documented to be highly "data-hungry".

This report provides a description of guidelines that can help when preparing NMT systems for a domain-specific translation, namely automatic translation of surveys. Partners describe the full training pipeline, including in-domain data collection and preparation, model training and the subsequent evaluation of the resulting NMT systems. The code for reproducing the experiments is available online at Github.[2]

This report expands on the work presented in Deliverable D4.5 which presented a neural machine translation (NMT) package that can be easily deployed and supports user-made NMT models based on the popular Transformer architecture (Vaswani et al., 2017). The best performing in-domain models are described in Section 3.2.Lastly, partners also present a real-life use case of the presented MT models translating the English coronavirus pandemic surveys for the WageIndicator Foundation.

2. The MCSQ Dataset

The Multilingual Corpus of Survey Questionnaires (MCSQ) is the first publicly available corpus of survey questionnaires. In its third version (entitled Rosalind Franklin), the MCSQ contains approximately 766,000 sentences and more than 4 million tokens, comprising 306 distinct questionnaires designed in the source (British) English language and their translations into Catalan, Czech, French, German, Norwegian, Portuguese, Spanish, and Russian, adding to 29 country-language combinations (e.g., Switzerland-French).

The MCSQ encompasses more than 40 years of survey research from large-scale comparative survey projects that provide cross-national and cross-cultural data to the Social Sciences and Humanities (SSH). Namely, the European Social Survey (ESS), the European Values Study (EVS), the Survey of Health Ageing and Retirement in Europe (SHARE), and the Wage Indicator Survey (WIS).

All questionnaires in the MCSQ are composed of survey items. A survey item is a request for an answer with a set of answer options, and may include additional textual elements guiding interviewers and

[2] https://github.com/ufal/SSHOC/tree/main/Task-4.2 (Accessed March 2022)

clarifying the information that should be understood and provided by respondents. Except in the case of the WIS, the translation process was implemented according to the Translation, Review, Adjudication, Pretesting and, Documentation (TRAPD) method, a team approach for the translation of survey questionnaires. Questionnaires included in the MCSQ were obtained from the survey projects' archives in distinct formats such as spreadsheets, XML, PDF files. The PDF files had to undergo an additional step of conversion to plain texts before going through the preprocessing pipeline. Finally, the texts were extracted from the input files and preprocessed, sentences aligned with respect to the English source and annotated with Part-of-speech (POS) and Named Entity Recognition (NER) tags. The corpus and all its metadata are freely available for visualization and download through an especially tailored user interface[3] and the CLARINO repository[4].

3. NMT Training Pipeline

The focus of the experiments was on exploring various uses of the available data from the domain of social surveys and their effect on the performance of the MT systems trained for the purpose of translating social surveys. Experiments were aimed mainly at the English-German and English-Russian language pairs.

3.1 Data Preprocessing

Although the task of survey translation is an instance of domain-specific MT, partners tried all the available bilingual data including both in-domain (MCSQ) and general-domain corpora. The general-domain corpora were provided by the Workshop on Machine Translation (WMT) conference and include various sources such as legal, news, Wikipedia entries, etc.[5] The general-domain parallel data is used only for model training.

Partners use the in-domain MCSQ survey data for both training and evaluation. After shuffling the respective in-domain De-En and Ru-En parallel corpora, 2,000 sentence pairs were removed from the in-domain training data and use 1,000 sentence pairs for validation and test set that partners use during development and final evaluation respectively.

[3] http://easy.mcsq.upf.edu/ (Accessed March 2022)

[4] https://repo.clarino.uib.no/xmlui/handle/11509/142 (Accessed March 2022)

[5] https://statmt.org/wmt21/translation-task.html (Accessed March 2022)

Table 1: Sizes of the available training corpora (number of sentence pairs). For the synthetic in-domain data, partners use all the available English-side sentences from the MCSQ corpus, resulting in the same size for both English-German and English-Russian.

	general-domain	in-domain (MCSQ)	synthetic in-domain (MCSQ)
English-German	11,590,399	36,078	648,737
English-Russian	12,890,374	17,793	648,737

Table 1 shows the comparison between the sizes of the available training bilingual corpora. Since the in-domain corpora are much smaller in volume compared to the general-domain corpora (tens of thousands of in-domain sentence pairs vs millions of general-domain sentence pairs) partners also create a synthetic in-domain parallel corpora by gathering all English sentences available in the MCSQ corpus and translating them automatically using an MT system. The details of the MT systems used for translating the English sentences are described in the following section.

Each of the training corpora is also shuffled and filtered: partners remove all sentence pairs where either the source-side or target-side sentence is an empty string and partners also remove sentence pairs with identical source and target sentence. Partners do not apply any special preprocessing, such as lowercasing or punctuation normalization - partners only perform tokenization to subword tokens using wordpiece (Sennrich et al., 2016b). Partners clip all training sentences to 150 subword tokens each. The code used for preparing the training data is available on GitHub.[6]

3.2 Experiment

The experiments were performed using the Tensor2tensor[7] (T2T, Vaswani et al. 2018) sequence learning framework implemented using TensorFlow[8] (Abadi et al. 2016) machine learning library for Python. Partners use the Transformer architecture with the "transformer_big_single_gpu" parameter settings. Each model is trained for 500,000 training iterations. Batch size is set to 2,900 tokens per batch. AdaFactor (Shazeer and Stern, 2018) is used for model optimization. A linear learning rate warmup is applied for 8,000 training steps followed by reciprocal square root learning rate decay. Dropout (Srivastava et al., 2014) is not applied during training.

During the inference, the last single checkpoint is used and beam search with beam size 4 and length penalty 1.0 is applied during decoding. The system performance is compared using an automatic

[6] https://github.com/ufal/SSHOC/tree/main/Task-4.2 (Accessed March 2022)
[7] https://github.com/tensorflow/tensor2tensor (Accessed March 2022)
[8] https://www.tensorflow.org/ (Accessed March 2022)

evaluation on the test set extracted from the MCSQ dataset. BLEU (Papineni et al., 2002) automatic MT evaluation metric is used - it is implemented in the MultEval tool (Clark et al. 2011) that supports bootstrap resampling.[9]

Several NMT systems are compared, each trained using a different part of the available training data. *WMT* model (baseline) uses only the general-domain parallel data. *MCSQ* model was trained only using a small amount of in-domain parallel data. *WMT+MCSQ* is a model trained on the combination of the in-domain and general-domain corpora. Table 2 shows the comparison between the MT models trained on the authentic parallel data.

Table 2: Comparison of the baseline systems using either in-domain data, general-domain data or a combination of both. The best result for each language pair is highlighted in ***bold.***

	En->De	De->En	En->Ru	Ru->En
WMT	16.3	18.1	13.7	13.8
MCSQ	**67.5**	**63.8**	**64.3**	**45.7**
WMT+MCSQ	52.3	57.1	46.6	39.4

As mentioned in the previous section, partners also tried to enhance in-domain data by preparing additional synthetic in-domain dataset. Partners gathered all English sentences within the MCSQ dataset and translated them using the models *WMT* and *MCSQ* models, in a similar fashion as suggested by (Sennrich et al., 2016a), labeling the resulting datasets WMT-BT and MCSQ-BT respectively. While the former should be better performing due to the amount of general-domain training data the latter can benefit from the matching domains. Partners then retrained the previous *WMT*, *MCSQ* and *WMT+MCSQ* setups using the additional synthetic training data.

Table 3: Comparison of the systems trained using the authentic MCSQ data in combination with the artificially translated English-side of the MCSQ corpora. MCSQ base (MCSQ from Table 2) is included as a baseline. The best result for each language pair is highlighted in bold.

	En->De	De->En	En->Ru	Ru->En
MCSQ base	**67.5**	63.8	**64.3**	45.7
MCSQ-with-WMT-BT	14.1	**75.0**	11.2	**74.7**
MCSQ-with-MCSQ-BT	66.4	74.7	45.9	45.5

[9] https://github.com/jhclark/multeval (Accessed March 2022)

Table 2 shows comparison between the baseline systems. As expected, the general-domain models perform poorly when applied to the domain of social surveys. On the other hand, the systems trained on the in-domain data showed a significant improvement despite being trained on a very limited amount of data. In comparison, when the models are trained on the combination of the in-domain and general-domain data, the overall performance drops (WMT+MCSQ) despite the models being trained on larger dataset implying the importance of the strict domain adaptation (MCSQ dataset only).

Table 3 shows the results of further enhancement of the MCSQ data by the synthetic in-domain parallel data. The back-translation of the "target" English MCSQ sentences into the source German or Russian language (used for improving De->En, Ru->En models) consistently helps to improve the system performance regardless of the original data used for the BT translation model training. On the other hand, the forward-translation (translation of the same English MCSQ sentences into German or Russian but treating the English side as the source-side) approach to synthetic data creation is much more sensitive to the domain of the MT model used to produce such synthetic parallel data. Partners hypothesize that this is mainly due to the back-translated data helping to improve the MT decoder by providing proper in-domain target-side training sentences while not hurting the encoder performance with the synthetic source-side sentences. This is not the case for the forward-translated data; in this case the target-side errors introduced by the MT model likely affect the training process, fitting the decoder to a target-side sentence distribution that differs from the true sentence distribution in the MCSQ dataset. The use of an in-domain MT model for forward-translation helps to tackle this problem, however, the models trained using the additional synthetic forward-translated data still do not outperform the in-domain MCSQ base model.

The best performing models are available at Lindat repository under the CC BY-NC-SA 4.0 Creative Commons license.[10] These models are compatible with the dockerized framework described in the Deliverable D4.5.

4. Use Case: Covid-19 Survey Translation

The WageIndicator Foundation is running a survey worldwide to collect information about how the pandemic has affected people's lives and work: the Living and Working in Coronavirus Times Survey[11]. This survey includes a textbox where respondents can type in their feelings/opinions/comments on the survey topic. By the end of October 2021, around 79000 people had responded to the questionnaire. Of these entries, around 13000 provide a quote, which can be in English or in other languages. In 2020, CUNI

[10] En-De: http://hdl.handle.net/11234/1-4680; En-Ru: http://hdl.handle.net/11234/1-4681 (Accessed March 2022)

[11] The Living and Working in Coronavirus Times Survey: https://wageindicator.org/salary/living-and-working-in-times-of-the-coronavirus/frequently-asked-question-corona-work-life-project-wageindicator (Accessed March 2022)

worked on a script which automatically translated those texts and provided WageIndicator with a new variable in the same dataset which had all texts in English. The translated text was used only for presentation purposes (it was never analyzed). One example of how it was used is presented in the image below:

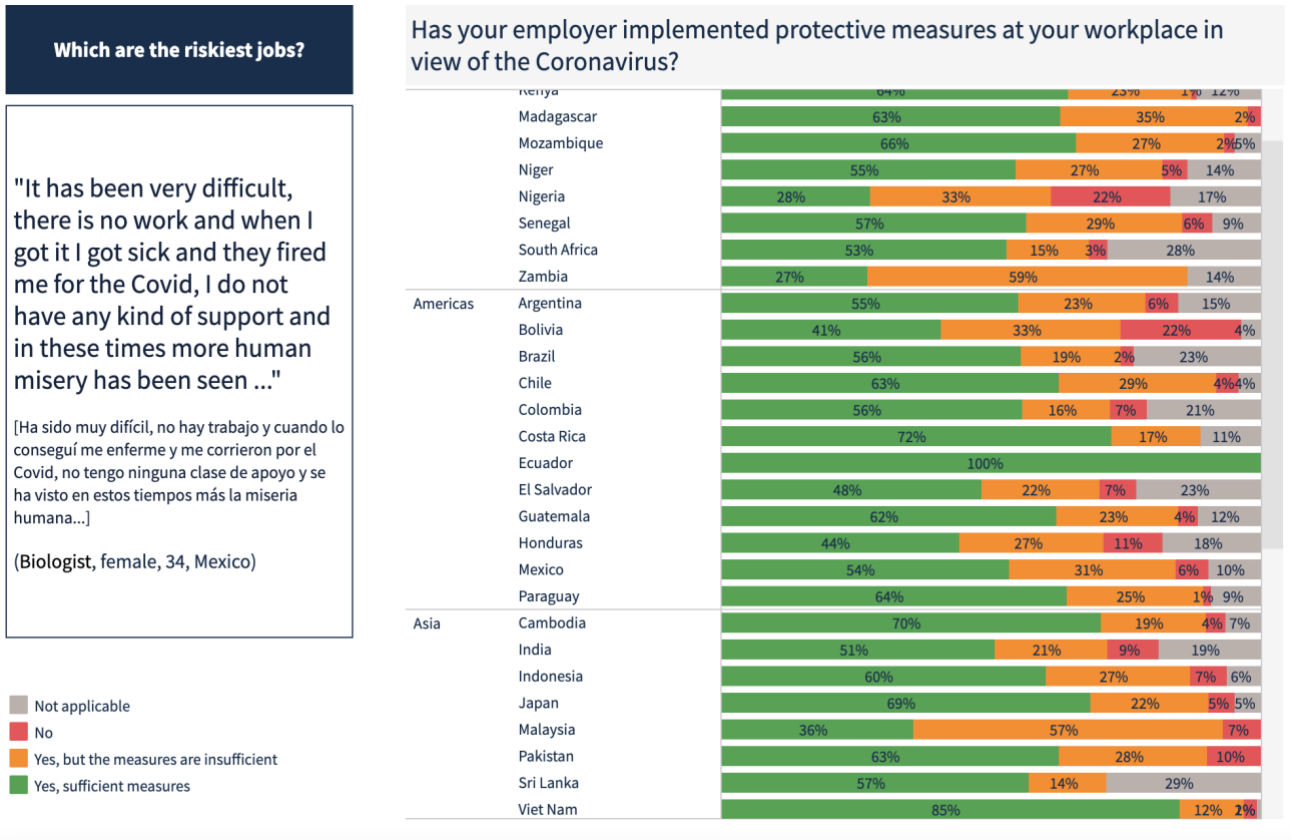

Figure 1: Presentation of the results of the Living and Working in Coronavirus Times Survey in WageIndicator.org.

4.1 Translation pipeline

Based on the requirements of the project partners, CUNI have also provided a translation pipeline to automate the translation process of specific survey files used within the project. The code for the automation is available as an example at the GitHub repository.[12] The example code is not universal since there are project-specific constants (filepaths, etc.). It should thus only serve as a guideline for preparing future solutions to similar translation problems.

[12] https://github.com/ufal/SSHOC/blob/main/Task-4.2/examples/translation_pipeline.sh (Accessed March 2022)

As a first step of the pipeline, a shared DropBox directory is set up. The shared directory has two purposes: one, partners need to monitor whether the user has uploaded any new files that need to be translated by the MT services and two, it serves as an intermediate storage of the translation outputs. Besides these two reasons, partners also chose DropBox because it also provides the developers an easy-to-use command-line interface (CLI) which simplifies the automation process.

After setting up the shared directory, the contents of the directory are monitored for any new input files. The pipeline used during the project expects a MS spreadsheet (a .xlsx file) because it is the most used file format by the project partners. For the purposes of text processing, the spreadsheet file is converted into a plaintext .csv file using the xlsx2csv utility.[13] After this conversion, a column containing the text for translation is identified via a user-defined label and extracted. The information about the source language is also extracted from each entry if present. Special characters and newlines are allowed inside the .csv entries - these have to be specified when converting the original file to be correctly escaped.

In the next step, the extracted text is sent to the Lindat translation service using the service's API.[14] The translation service is currently limited only to a few European languages - if the language is not supported by the service, a warning message is printed. Another option is sending the input text to the other translation service that supports the language - for comparison purposes, the partners experimented with Google Translate service, however, the service has only a limited use free-of-charge and after the expiration of the trial period requires a paid subscription.

Finally, the translated text is inserted into the input .csv file as a new column. Next, the file is converted back to MS spreadsheet format and uploaded back to the DropBox shared directory. The script is designed to be executed periodically (e.g., by using the cron utility) so it can check for new input files and translate them or to halt when no new files are present.

[13] https://github.com/dilshod/xlsx2csv (Accessed March 2022)

[14] https://lindat.mff.cuni.cz/services/transformer/docs (Accessed March 2022)

5. Conclusion

This report describes the training pipeline used to prepare domain specific NMT systems for automatic translation of social surveys. Multiple approaches towards exploiting available in-domain and out-of-domain training corpora were compared and methods for utilising additional in-domain monolingual data were successfully applied.

The results show that the domain-specific MT systems are much better suited for the translation of the social surveys, however, they can underperform when the available training data is limited. A significant improvement can be gained by creating synthetic parallel data using back-translation and forward-translation. Although the benefits of the additional synthetic parallel data created by back-translation lead to significant improvement regardless of the domain of the MT model used for back-translation, this effect is not so clear when a forward-translation synthetic data is added to the training dataset. Additionally, a domain mismatch has a clear negative effect on the quality of the forward-translated synthetic training data.

6. References

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." *ArXiv:1603.04467 [Cs]*.

Clark, Jonathan H., Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. "Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability." Pp. 176–81 in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "Bleu: A Method for Automatic Evaluation of Machine Translation." Pp. 311–18 in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016a. "Improving Neural Machine Translation Models with Monolingual Data." Pp. 86–96 in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016b. "Neural Machine Translation of Rare Words with Subword Units." Pp. 1715–25 in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics.

Shazeer, Noam, and Mitchell Stern. 2018. "Adafactor: Adaptive Learning Rates with Sublinear Memory Cost." Pp. 4596–4604 in *Proceedings of the 35th International Conference on Machine Learning*. PMLR.Srivastava, Nitish, Geo\uc0\u64256{}rey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov., 2014 \uc0\u8220{}Dropout: A Simple Way to Prevent Neural Networks from Over\uc0\u64257{}tting.\uc0\u8221{} 30.

Vaswani, Ashish, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. "Tensor2Tensor for Neural Machine Translation." *ArXiv:1803.07416 [Cs, Stat]*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *ArXiv:1706.03762 [Cs]*.

List of Figures

List of Tables