MultiTraiNMT: Training Materials to Approach Neural Machine Translation from Scratch*

Gema Ramírez-Sánchez[1], Juan Antonio Pérez-Ortiz[2], Felipe Sánchez-Martínez[2], Caroline Rossi[3], Dorothy Kenny[4], Riccardo Superbo[5], Pilar Sánchez-Gijón[6], and Olga Torres-Hostench[6]

[1] Prompsit Language Engineering, gema@prompsit.com
[2] Universitat d'Alacant, {japerez,fsanchez}@ua.es
[3] Université Grenoble-Alpes, caroline.rossi@univ-grenoble-alpes.fr
[4] Dublin City University, dorothy.kenny@dcu.ie
[5] KantanMT.com, riccardos@kantanmt.com
[6] Universitat Autònoma de Barcelona, {pilar.sanchez.gijon,olga.torres.hostench}@uab.cat

Abstract. The aim of the MultiTraiNMT Erasmus+ project is to develop an open innovative syllabus in neural machine translation (NMT) for language learners and translators as multilingual citizens. Machine translation is seen as a resource to provide support to citizens when trying to acquire and develop language skills, provided they are given informed and critical training. Machine translation would thus help tackle the mismatch between the EU aim of having multilingual citizens who speak at least two foreign languages and the current situation in which they generally fall far short of this objective. The training materials consist of an open-access coursebook, an open-source NMT web application (MutNMT) for training purposes and corresponding activities.

Keywords: machine translation · neural machine translation · training · multilingual citizens · project description.

1 Introduction

The aim of the Erasmus+ strategic partnership "MultiTraiNMT – Machine Translation Training for Multilingual Citizens"[7] (2019–2022) is to develop, evaluate and disseminate open-access materials and open-source applications that will lead to the enhancement of teaching and learning about machine translation [3] among language learners, language teachers, trainee translators, translation teachers and professional translators across Europe.

MultiTraiNMT brings together experts at four European universities — Universitat Autònoma de Barcelona, Université Grenoble–Alpes, Dublin City University and Universitat d'Alacant, and two enterprises — Prompsit Language

* With the support of the Erasmus+ programme of the European Union.
[7] Project website: https://multitrainmt.eu

Translation and Interpreting Technology Online, pages 80–85,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_009

Engineering and Xcelerator Machine Translations, and is supported by numerous associate partners in education and the translation industry, all of whom are interested in teaching and learning about the use of machine translation. The partnership aims specifically at developing an innovative syllabus in machine translation (MT) in general, and neural machine translation (NMT) in particular [3]. On completion we will provide the following components which will be described in the following sections:

- An **open-access coursebook** that addresses both the technical foundations of machine learning—and especially deep learning—as used in MT, and the ethical, societal and professional implications of this technology.
- MutNMT, a **pedagogical NMT web application** that allows users to learn how NMT works, and gain insight into the internal workings of NMT systems.
- **Learning activities** related to the coursebook and MutNMT that allow language learners and translators to co-construct knowledge on NMT.

The training is designed to be followed in both asynchronous and synchronous forms. On the one hand, self-learners will be able to follow the coursebook and perform the corresponding activities. On the other hand, any interested teacher will be able to use the course and the activities in synchronous form with students. The coursebook and learning activities are designed taking into account different progress levels to approach different student profiles. Measurability, quality and progress of the project can be followed in the project website.

In short, we are developing an up-to-date syllabus on MT for use in European Higher Education and elsewhere; one that will allow students to acquire the technical and ethical skills and competences required to become informed, critical users of contemporary MT in their own language learning and translation practice. In so doing, we open up the world of machine learning to language and translation students, their teachers and others, enhancing their ability to function as technologically competent, informed citizens in a multilingual Europe.

2 The coursebook

The Creative Commons-licensed[8] open-access coursebook is organised in eight chapters. Instructors may conveniently arrange them in a different order for their courses. The modules are:

1. Multilingualism
2. Introduction to machine translation

[8] https://creativecommons.org

3. How to choose a suitable MT system and evaluation of machine translation quality
4. How to prepare and select texts for machine translation
5. How to deal with machine translation mistakes, post-editing and error fixing
6. Ethical aspects of machine translation
7. How neural machine translation works
8. Custom neural machine translation

3 The web application

MutNMT is an open-source web application[9] to train NMT for didactic purposes. It lets users train, inspect, evaluate and translate using neural engines. Contributions to other open-source projects have also been made, namely to JoeyNMT [4], a command-line tool to train NMT engines. Technical documentation is provided along with the code. Manuals for both users and instructors are available. A production installation of MutNMT is currently under evaluation.[10]

There are three profiles of users: *beginners* with default access to all basic features; *experts* with access to basic and advanced features without administration rights; and *admins* with full access to all features. Admins are able to upgrade beginners (default profile) to experts or admins by request at any time. A brief description of the main features and windows of MutNMT follows:

Data. MutNMT, as every other NMT system, needs corpora in the form of parallel data to learn from. Previewing, downloading and grabbing corpora is possible. Corpora uploaded to MutNMT retain their original licences. While free/open source corpora are recommended in MutNMT, users are allowed to upload proprietary corpora and keep them private. These corpora are shown in the application as a collection of resources as shown in figure 1.

Engines. There is also a library of engines in MutNMT, that is, already available MT systems that have been trained and shared. Of special interest are also the actions allowed: seeing the full training log of an engine, downloading the model or downloading the corpora it was trained with. Models created with MutNMT come wiht a GPL-v3 free/open-source license. While beginners can only see training reports, experts and admins are able to resume the training of an engine.

Training. This is an advanced feature for experts and admins that allows them to train NMT engines using MutNMT. Users will need to set up engine details, configuration parameters and select corpora for training a particular system.

[9] Code available on `https://github.com/Prompsit/mutnmt`
[10] `https://ntradumatica.uab.cat`

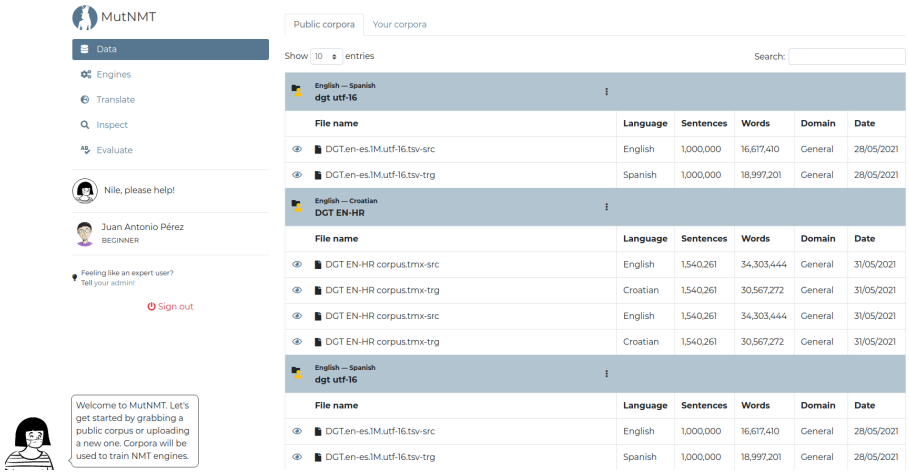

Fig. 1. A screen capture of the MutNMT's window that allows users to preview, add, and download parallel corpora.

Translate. By using the available engines, all users will be able to copy and paste a series of sentences and translate them. They will get the resulting translation in a text box and be able to export a TMX [1] out of the whole translation, thus saving pairs of source and target sentences in a standard format. Document translation is also supported and will give as a result the translation in either the original document format or in a TMX file.

Inspect. There are several options in this window, all aimed at seeing the internals of the translation engines at work such as allowing users to input a sentence and see it at different steps of processing by a particular engine: pre-processed input, hypothesis generation (n-best), pre-final output (still to be post-processed) and final output.

Evaluate. As a final step, users will be able to evaluate the output of MT comparing it to other machine translated texts or to professional human translations. MutNMT provides several automatic evaluation metrics, such as BLEU [5] and ChrF3 [6], at both document and sentence levels. All these results can also be downloaded in spreadsheet format.

4 The learning activities

Two types of learning activities are being created. On the one hand, self-learning questions are aimed at students working at their own pace; these are short-answer questions with immediate automatic feedback. On the other hand, open-answer teacher-guided activities can be customised and adapted to different contexts. After exploring different formats and repositories of learning objects, we have

opted for the open-source H5P platform,[11] as it allows each of our activities to be self-contained and easily embeddable by instructors in learning management systems such as Moodle[12] or more general environments such as Wordpress[13]. Each exercise includes metadata such as difficulty, estimated answering time, comments for instructors and considerations when adapting the text of the question to other language combinations.

5 The MultitraiNMT associate partner network

MultiTraiNMT invites higher education institutions and teachers of translation and foreign languages to join the project as associate partners/members. In order to become an associate partner, interested parties may visit the website section *Join us / Become a member* in order to download the *Associate Partners Agreement* and adapt it to their needs and interests. The aim of the network is not only to share the aforementioned coursebook, MutNMT and activities but also to create a working group to share activities, experience and best practices so that the project becomes collaborative. The partners may:

- Evaluate the use of the project coursebook in their classes.
- Test the MutNMT educational system for managing NMT engines for didactic purposes.
- Participate with project partners in the piloting of project activities on MT training or voluntary sharing of MT training activities in the project.
- Arrange with the MultitraiNMT project the certification of participants.
- Participate actively in the project multiplier events or other dissemination events.
- Participate in any other training or research activity which fosters the development of MT skills in general among multilingual citizens.

6 Conclusions

Despite recent advances in freely available NMT engines, machine translation is still often considered too complex to be understood by a non-specialist audience. The materials developed within the MultiTraiNMT project are intended to show that MT literacy can be developed across various target audiences, in line with recent proposals [2]. We have designed a course that includes activities and a platform (MutNMT) for learning NMT by doing, and we invite collaborations within our network of associate partners.

[11] `https://h5p.org/`
[12] `https://moodle.org/`
[13] `https://wordpress.org/`

References

1. Localisation industry standards: Translation Memory eXchange (TMX). https://www.etsi.org/deliver/etsi_gs/LIS/001_099/002/01.04.02_60/gs_LIS002v010402p.pdf (2013), accessed: 2021-06-24.
2. Bowker, L., Buitrago, J.: Machine translation and global research: Towards improved machine translation literacy in the scholarly community. Emerald Group Publishing (2019)
3. Koehn, P.: Neural machine translation. Cambridge University Press (2020)
4. Kreutzer, J., Bastings, J., Riezler, S.: Joey NMT: A minimalist NMT toolkit for novices. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. pp. 109–114 (2019)
5. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics (2002)
6. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015)