# Developing Machine Translation Engines for Multilingual Participatory Spaces

**Pintu Lohar, Guodong Xie, Andy Way**
ADAPT Centre,
Dublin City University,
Dublin, Ireland
firstname.lastname@adaptcentre.ie

## Abstract

It is often a challenging task to build machine translation (MT) engines for a specific domain due to the lack of parallel data in that area. In this project, we develop a range of MT systems for 6 European languages (English, German, Italian, French, Polish and Irish) in all directions and in two domains (environment and economics).

## 1 Project Description

This work is part of a larger project called "EU-ComMeet"[1] on developing participatory spaces using a multi-stage, multi-level, multi-mode, multi-lingual, dynamic deliberative approach (M4D2). The goal of this project is to integrate together automated moderation and automated translation to allow multilingual, multi-national participation in deliberative democratic forums. Our main contribution is to facilitate a multilingual deliberative space (MDS) via MT. Users will be able to communicate with each other via MT while speaking or writing in their own languages.

## 2 MT for Participatory Space

MT is a process that automatically translates text from one language to another. It is usually ideal to train MT models using domain-specific parallel corpora (e.g. a corpus of biomedical domain (Névéol et al., 2018) for medical texts). However, to the best of our knowledge, no such data belonging particularly to the economics and environment domains is available. Accordingly, we decided to use the Europarl corpus (Koehn, 2005) because it is (i) a good-quality corpus, (ii) large enough for MT training, and (iii) mixed domain, so that a significant number text pairs belonging to many major domains such as science, environment, economics, politics etc can be found in this corpus. For each language pair, we built four translation models: (i) two baseline models, and (ii) two domain-adapted models in both translation directions. The baseline models are built using the whole Europarl corpus and tuned on the benchmark news development data set[2] provided by the organisers of WMT 2021. In contrast, the domain-adapted models are built using the same Europarl corpus but tuned on in-domain development data extracted from the news data by applying domain-specific key terms. Both models are tested on these datasets to compare their system performance on domain-specific test sets. In order to compile the in-domain development and test data, we form two lists of domain-specific key terms: one for 'environment' and another for 'economics'. A total of 150 environment and 201 economics key terms are used, including some of the following example terms: (i) **Environment:** *sustainability*, *pollution*, *climate* etc. (ii) **Economics:** *inflation*, *employment*, *privatization* etc. These key terms are then used to extract only those text pairs from the *news* data that contain at least one of these key terms in order to form the in-domain development and test datasets. Table 1 shows the data statistics and its domain-wise distribution. We provide a brief description on two types of translation models in Table 2. The MT models are built using OpenNMT (Klein et al., 2017) with transformer architecture (Vaswani et al., 2017). The translation outputs are



[1] https://www.eucommeet.eu/

[2] http://data.statmt.org/wmt21/translation-task/dev.tgz

| Domain | Europarl | News |
|---|---|---|
| Train (multidomain) | 1,957,832 | // |
| dev+test (news domain) | // | 38,647 |
| Environment domain | // | 1,145 |
| Economics domain | // | 4,014 |

**Table 1:** Training and domain-wise data distribution

| Model | Tuned on | Tested on |
|---|---|---|
| Baseline | News domain | Environment +Economics |
| Domain-adapted | Environment +Economics | Environment +Economics |

**Table 2:** Baseline and domain-adapted models

evaluated using BLEU (Papineni et al., 2002). We

| Test domain | Model | BLEU |
|---|---|---|
| Economics | Baseline | 21.67 |
| | Domain-adapted | **22.95** |
| Environment | Baseline | 21.56 |
| | Domain-adapted | **23.20** |

**Table 3:** BLEU scores for English–German

depict the results for English–German in Table 3 which shows that the domain-adapted models outperform the baselines in both domains. All these improvements are statistically significant as verified using MultEval (Clark et al., 2011).

## 3 Architecture of the MT system

Now that the systems have been built, multilingual discussions involving people speaking different languages from different countries in different citizens' assemblies will take place. Our MT engines will be used to translate among different participants through the project platform. Participants will be in different locations across the 5 countries. In making the MT engines accessible, we will need to bear in mind three closely related criteria: reliability, speed and security. To address this problem, we adopt a two-layer architecture and security verification, as shown in Figure 1. The first layer (the web server) handles access verification, and translation requests from different devices in multiple locations are sent to the translation GPU servers in the second layer. To speed up the translation response, the two-layer server groups (in the green rectangles) are deployed in different countries so that the translation requests will be processed locally. We expose our MT service through the web server which creates an HTTP REST server interface in the web server. To enhance the security of the MT system, we adopt the JSON Web Token (JWT)[3] to verify user access.

[3] https://jwt.io/introduction

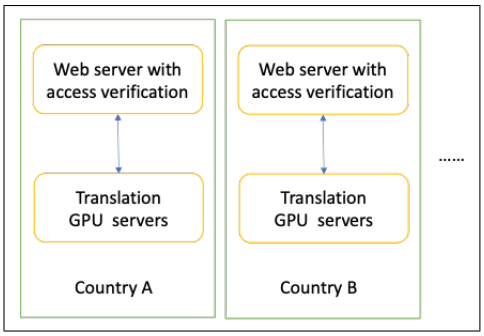


**Figure 1:** Architecture of MT system

## Acknowledgments

This work was funded by the European Commission under H2020-EU.3.6. - SOCIETAL CHALLENGES - Europe In A Changing World - Inclusive, Innovative And Reflective Societies, grant agreement ID: 959234.

## References

Clark, Jonathan H., Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies:*, pages 176–181, Portland, Oregon, USA.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 67–72, Vancouver, Canada.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, September 13-15.

Névéol, Aurélie, Antonio Jimeno Yepes, Mariana Neves, and Karin Verspoor. 2018. Parallel corpora for the biomedical domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 296–291, Miyazaki, Japan.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, Long Beach, USA.