MICE

Joachim Van den Bogaert,
Heidi Depraetere, Tom Vanallemeersch,
Frederic Everaert, Koen Van Winckel
CrossLang
Kerkstraat 106
9050 Gentbrugge
Belgium
{first.lastname}@crosslang.com

Artūrs Vasiļevskis, Valters Sics
Tilde IT
Naugarduko g. 100
LT-03160 Vilnius
Latvia
{first.lastname}@tilde.lt

Katri Tammsaar, Ingmar Vali,
Tambet Artma, Piret Saartee,
Laura Katariina Teder
RIK
Lubja 4
19081, Tallinn
Estonia
{first.lastname}@rik.ee

Johan Haelterman, David Bienfait
Bureau for Standardisation
Rue Joseph II 40 PO box 6
1000, Brussels
Belgium
{first.lastname}@nbn.be

Abstract

The MICE project (2018 – 2020) will deliver a middleware layer for improving the output quality of the eTranslation system of the European Commission's Connecting Europe Facility, through additional services, such as domain adaptation and named-entity recognition. It will also deliver a user portal, allowing for human post-editing.

1 Objectives

The MICE project (Middleware for Customised eTranslation), which is funded by the CEF Telecom programme (Connecting Europe Facility) and runs from October 2018 to September 2020, delivers a middleware layer for the improvement of the eTranslation machine translation (MT) system. The latter is developed by the Directorate-General for Translation (DGT), supports all 24 official EU languages, and is provided by the CEF Automated Translation building block of the Directorate-General for Communications Networks, Content and Technology (DG CNECT), as a service to digital service infrastructures (DSIs) of the European Commission (EC) and to public administrations of Member States. The project consortium of MICE consists of two companies, CrossLang (coordinator) and Tilde, and two public organisations, NBN (Bureau for Standardisation, Belgium) and RIK (Centre of Registers and Information Systems, Estonia).

The middleware layer consists of the following services:

- domain adaptation;
- terminology resolution;
- named-entity recognition;
- document filtering;
- normalisation.

MICE will also provide a human and automated post-editing (PE) environment for CEF eTranslation output. This will help users to dynamically improve the MT output and aggregate data for further system improvement.

The tests in the project involve four languages, i.e. English, Dutch, French and Estonian, and two domains, i.e. standards and e-Business/e-Land[1] register information, in two countries (Belgium and Estonia). Domain-specific neural MT systems will be made available by the project consortium.

---

[1] e-Land registers are electronic land registries that register the ownership of land and property.
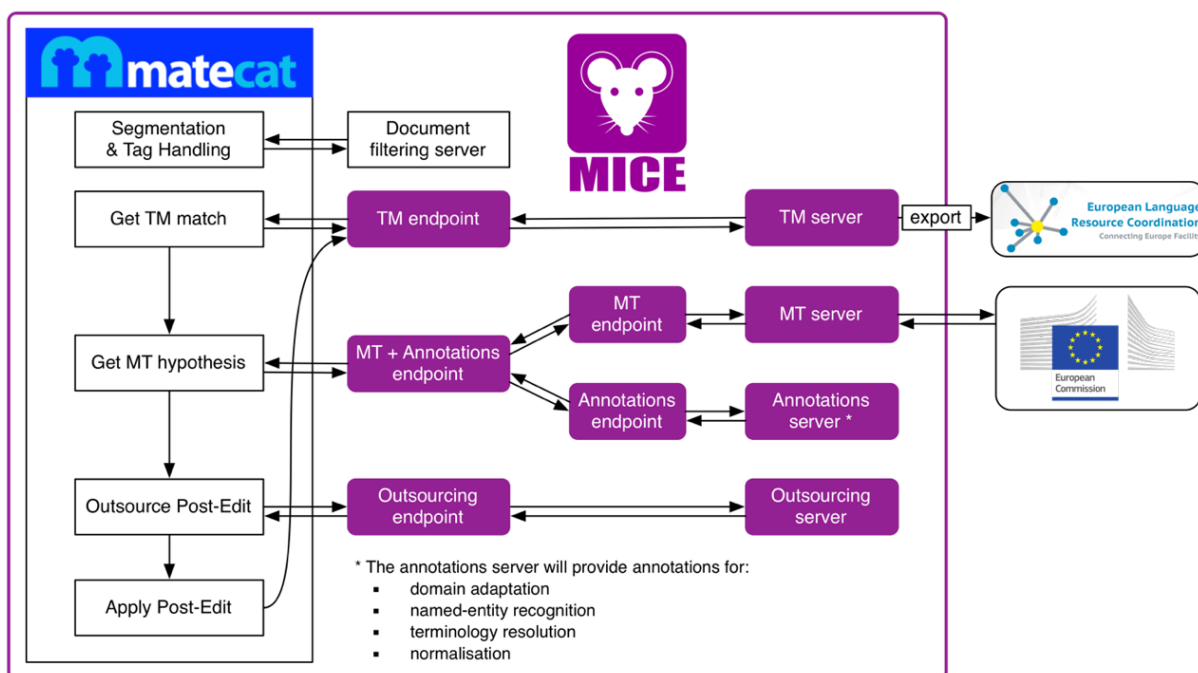
Figure 1: MICE architecture

2 Architecture

MICE will expose its middleware layer for customisation through APIs and a user portal, in order to increase its impact and usability. Tasks will be performed in real-time or offline, depending on user preference. The input consists of text snippets (messages in plain text of maximally 5,000 characters) or full-text documents (Microsoft Office, open document formats, etc.). The security level will be compliant with the Electronic Simple European Networked Services (e-SENS).

The MICE project will create a reference implementation for the automated translation of standards and e-Business/e-Land register information in Belgium and Estonia, and be extensible to allow for future add-ons of MT-related services, such as automated domain detection or combination of MT systems.

The architecture (outlined in Figure 1) will incorporate the open source MateCat[2] computer aided translation (CAT) tool to provide user portal functionality (user interface and configuration management).

Existing integrations for translation memory (TM) leveraging, MT and post-editing outsourcing will be abstracted into service endpoints that can be connected to external services. For example:

- The TM endpoint will allow for connecting MateCat to any TM server. Exports can be added to the European Language Resource Coordination repository.
- The MT + annotations endpoint will allow for connectivity with the CEF eTranslation service (or any other third-party MT service) while integrating advanced features, such as named-entity recognition.
- The outsourcing endpoint will allow for dispatching post-editing jobs to third-party providers.

While developing the MICE solution, existing interfaces will be upgraded to allow for advanced MT/TM features:

- The existing MT interface will be extended to encode annotations information, instructing MT engines to apply a distinct treatment to named entities, detected terms, etc.
- The TM interface will be adapted to allow for the storage of post-editing triplets. Currently, most TMs only store the source and target sentence. By storing also the MT hypothesis, instead of overwriting it with the human translation, MICE will contribute to the development of corpora for the training of quality estimation and automated post-editing systems.

Acknowledgement

---

[2] See https://github.com/matecat/MateCat