MICE: Adapting MT through Middleware

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

Abstract

The MICE project (2018–2020) will deliver a middleware layer for improving the output quality of the eTranslation system of EC'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 Customer 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 DG Translation, supports all 24 official EU languages, and is provided by the CEF Automated Translation building block of DG CNECT as a service to Digital Service Infrastructures (DSIs) of the 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 also provides a human and automated post-editing (PE) environment for CEF eTranslation output, based on MateCat.[1] This will help users to dynamically enhance the MT output and aggregate data for further system improvement.

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

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 MICE architecture, which is shown in Figure 1, is compliant with the eDelivery building block.

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. It will be extensible in order to allow for future add-ons of MT-related services, such as automated domain detection or combination of MT systems.

3 Test results

In both use cases (NBN, RIK), similar tests for domain adaptation were performed. We will illustrate them based on the NBN use case.

[1] http://www.matecat.com

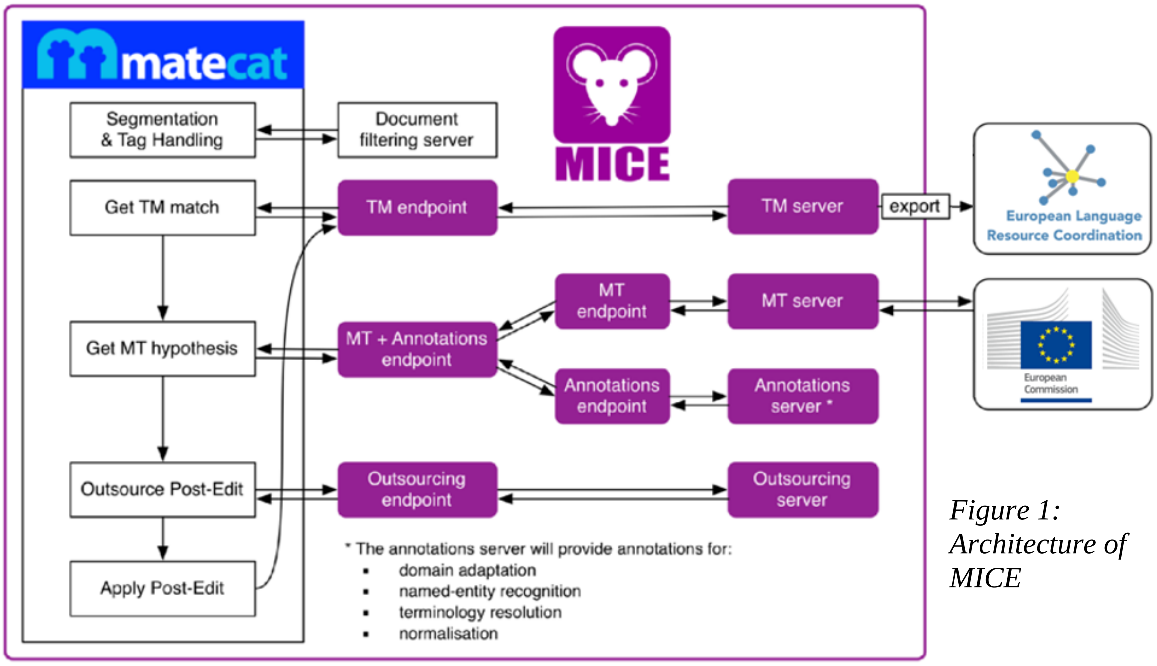

Figure 1: Architecture of MICE

In-domain English–French and English–Dutch MT systems were built by extracting and aligning sentences from PDF files of NBN. In addition, parallel synthetic data were created by backtranslating text from in-domain monolingual target language data, and other parallel text was obtained by scraping the NBN website and aligning web pages and sentence pairs in them. Overfitting on the domain was avoided by enriching the training data with text from other domains.

Further domain adaptation involved a preprocessing step before training the MT systems. This step consisted of named-entity recognition and detection of numbers, URLs, email addresses and technical entities. Sentences were normalized by replacing the detected parts with placeholders. When applying the MT system to new sentences, the placeholders in the output were replaced by a copy of the source text (e.g. person names) or a localized version (in case of numbers). Another type of domain adaptation consisted of enriching MT training data with bilingual term lists automatically extracted using TermCalc (Vanallemeersch and Kockaert 2010).

Domain-adapted MT systems were trained with OpenNMT (Klein et al. 2017) and evaluated using a manually cleaned in-domain test set, as well as using out-of-domain test sets (legal domain) in order to prevent a risk of overfitting. The BLEU scores shown in Table 1 and 2 indicate that MT systems with in-domain training data perform better than systems without, and that the use of bilingual terms further improves the results. The scores for the legal domain, on the other hand, are more or less constant.

Test Train	In-domain	Legal 1	Legal 2
Various domains	23.2	37.9	46.2
+ In-domain	35.7	35.8	44.6
+ Bilingual terms	35.9	37.1	45.1

Table 1 BLEU scores English–Dutch

Test Train	In-domain	Legal 1	Legal 2
Various domains	29.1	34.8	40.1
+ In-domain	49.5	34.9	39.5
+ Bilingual terms	51.3	35.3	40.6

Table 2 BLEU scores English–French

Acknowledgement

MICE is funded by the EC's CEF Telecom programme (project 2017-EU-IA-0169).

References

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ACL 2017*, *System Demonstrations*, Vancouver, Canada 67–72.

Vanallemeersch, Tom and Hendrik Kockaert. 2010. Automated Detection of Inconsistent Phraseology Translation. *Southern African Linguistics and Applied Language Studies*, 28(3):283–290.