First release of the Data Marketplace services for cleaning, anonymizing and clustering data

This document contains a short description of the activities performed by the CEF Data Marketplace consortium on Activity 3 “Data Processing Tools”. The aim of this Activity is to provide advanced services to enable and facilitate the exchange of data through the Marketplace. The three services are: ***translation memory (TM) cleaner***, ***TM anonymizer*** and ***TM data matching***. The TM cleaner has the goal of removing wrong and dirty translation units (TUs) from large TMs provided by the seller. The TM anonymizer has the goal of identifying in the TM those TUs that can contain private information and mark them to be checked by the seller. The TM matching data service has the goal of retrieving a portion of the available data in the Marketplace that is most similar to an initial seed (query).

To evaluate these services and build the software, four steps have been performed:

- Identification of available tools and evaluation specifications
- Creation of evaluation data
- Tools’ evaluation and selection
- Creation of services and demos

Identification of available tools and evaluation specifications

For each service, the consortium identified a set of tools available within the consortium, previously tested on several languages and open source. Following this strategy, *(i)* four cleaning tools, *(ii)* two named entity recognition models for anonymization and *(iii)* two clustering algorithms were selected and evaluated.
The evaluation was a multifaceted step, where each tool was analyzed by different perspectives:

- Intrinsic evaluation: measures the performance of an NLP component on its defined subtask, usually against a defined standard (e.g. accuracy over a manually labelled test set).
- Extrinsic evaluation: focuses on the component’s contribution to the performance of a complete application, in our scenario the machine translation task.
- Processing time evaluation: measures the speed of performing a specific task on a benchmark.
- Analysis of technical requirements: identifies all the technical conditions under which a tool runs.

Creation of the Evaluation Data

To perform the intrinsic and extrinsic evaluations of the ***cleaning*** and ***clustering*** tools, manually annotated test sets were created. To this purpose, for each of the five chosen language pairs (En-Cs, En-De, En-It, En-Lv, De-It), 2,500 translation units (TUs) were extracted from the TAUS repository and manually annotated by professional translators. In the resulting gold standard data, each TU is annotated with information about whether (i) it is clean - i.e. the translation is correct and fully equivalent to its source (ii) it belongs to the legal domain. Then, three different test sets were extracted from the gold data, according to the features required to perform the evaluations described above.

The gold standard data created for the five language pairs are publicly released through the ELRC-SHARE repository, and are available at the following link:

https://www.elrc-share.eu/repository/browse/cef-data-marketplace-multilingual-benchmark-for-the-evaluation-of-cleaning-and-clustering-tools/365a8b821aa011eb913100155d02670611118e05e423402bb729137ecf6ac864/

The evaluation of the ***anonymization*** tools was carried out (i) against publicly available gold standards for named entity recognition, as well as (ii) *a posteriori*, i.e. asking professional translators to manually assess the precision of the tool on a sample of automatically identified entities.

Tools' evaluation and selection

Each tool was evaluated based on the specifications mentioned above and using the manually annotated test sets. For each service, the best performing tool was identified:

- TM Cleaner: BiCleaner
- TM Anonymizer: DeepPavlov NER MBert and Translated Anonymizer
- TM Matching Data: TAUS matching data

A presentation containing an extensive summary of the evaluation is attached to this document.

Creation of services and demos

Around each of the selected tools, a web service has been developed and exposed as a REST API (Application Programming Interface) with at least one endpoint. For the TM Cleaner and TM Anonymization tools, the endpoint works at the TU level, while for the TM Matching Data the

endpoint works at the corpus level. Each web service API provides specific parameters (e.g. the language pair of the TU/corpus), mandatory or optional, depending on the interfaced tools.
In addition to the web services, a simple web GUI (Graphical User Interface) has been developed for each service. The GUI consists of a web page where the user can interact with the service by uploading data and see the results of the processing steps. Such output can be optionally sent to an email address provided by the user in the GUI.
Each web service and GUI are provided in the form of a Docker container image with all the advantages that such virtualization offers.

The source codes of the three services are made available at the following git repositories:

- TM Cleaner: http://github.com/hlt-mt/TM-Cleaner
- TM Anonymizer: http://github.com/hlt-mt/TM-Anonymizer
- TM Matching Data: http://github.com/hlt-mt/TM-MatchingData

The services are also made available through the ELRC-SHARE repository, at the following links:

- TM Cleaner: https://elrc-share.eu/repository/browse/translation-memory-cleaner/ca1a1e581a0711eb913100155d026706d08b69de159c4530b7916526208ccb3e/

- TM Anonymizer: https://elrc-share.eu/repository/browse/translation-memory-anonymizer/9816335017e211eb913100155d0267068139202586bb42f586f777191ec6f5b2/

- TM Matching Data: https://elrc-share.eu/repository/browse/translation-memory-matching-data/5a48e00819c811eb913100155d0267064f9c9d81922f4b9da595c8b41f5e3196/

The demos of the services are available here:

- TM Cleaner: https://cef-datamarketplace-tmcleaning.translated.com/
- TM Anonymizer: https://cef-datamarketplace-tmanonymizer.translated.com/
- TM Matching Data: https://cef-datamarketplace-tmmatchingdata.translated.com/