Curated Multilingual Language Resources for CEF AT (CURLICAT): Overall View

Tamás Váradi
Research Institute for Linguistics, Budapest, Hungary
varadi.tamas@nytud.hu

Marko Tadić
University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb, Croatia
marko.tadic@ffzg.hr

Svetla Koeva
Institute of Bulgarian Language, Bulgarian Academy of Sciences, Sofia, Bulgaria
svetla@dcl.bas.bg

Maciej Ogrodniczuk
Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
maciej.ogrodniczuk@ipipan.waw.pl

Dan Tufiș
RACAI, Romanian Academy, Bucharest, Romania
tufis@racai.ro

Radovan Garabík
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia
garabik@kassiopeia.juls.savba.sk

Simon Krek, Andraž Repar
Institute Jozef Stefan, Ljubljana, Slovenia
simon.krek@ijs.si,
repar.andraz@gmail.com

Abstract

The work in progress on the CEF action CURLICAT is presented. The general aim of the action is to compile curated monolingual datasets in seven languages of the consortium in domains of relevance to European Digital Service Infrastructures (DSIs) in order to enhance the eTranslation services.

1 Introduction

The paper presents the work in progress on the CEF action Curated Multilingual Language Resources for CEF AT (CURLICAT, which runs from 2020-06-01 till 2022-11-30). The aim of the action is to compile monolingual curated datasets in seven languages of the consortium (Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak, Slovenian) in domains of relevance to European Digital Service Infrastructures (DSIs) with a view to enhancing the eTranslation automated translation system.

2 Datasets

The primary data come from national or reference corpora of the above languages and it is planned to cover domains of interest for CEF DSIs such as eHealth, Europeana or eGovernment. When completed, the corpus will contain at least at least 2 million sentences from each language, i.e. 14 million sentences, estimated to number at least 140 million words, from domains including culture, health, science and economy/finances. For each language, it is expected to produce corpora in each of the above mentioned four domains with at least 500 000 sentences and 5 million words. In case that legally non-binding data with a clear licence allowing free redistribution could not be found from the national corpora in the required quantities, additional data is included from other sources.

2.1 Annotation

Apart from corpora being domain classified, data are linguistically annotated including sentence splitting, tokenisation, lemmatisation, part-of-speech/morphosyntactic-descriptor tagging, dependency parsing and NERC. The annotation

follows the extended CoNLL-U Plus[1] format presented by Váradi et al. (2020). Additionally, terms from the most recent version of the IATE terminological database are identified and annotated so that the language models built with the help of these corpora could take into account not only single words but also multi-word expressions since these terms represent an additional layer of annotation in stand-off manner. With this additional annotation these corpora can serve as a valuable resource for terminological processing as well.

2.2 Intellectual Propery Rights Issues and Anonymisation

The data are technically and legally cleaned by either of two procedures: 1) inclusion of text samples published under permissive licences, or for which consent was obtained from the content producer, or 2) scrambling of the order of sentences. In this way these corpora will be useful for producing language models up to the level of a sentence, while they will not be useful for higher linguistic level language modelling, but even with this limitation we see these corpora as a valuable resource for MT training. The metadata will specify whether the texts were scrambled or not.

For legal reasons data will also be anonymised through replacement of named entities of the same kind and with similar phonological, morphological or graphemic structure (a process that is inherently language-dependent, but, e.g. for Romanian "Maria" becomes "_#PER#1_", while "Mariei" becomes "_#PER#1_ei"). To ensure a higher degree of privacy preservation, local pseudonymisation, as the process of compete replacement of named entities by one or more artificial identifiers, at document or sub-document level, is used.

During the course of the project, we will develop an anonymisation solution tailored to the specific needs of the CURLICAT corpus by leaning on existing European anonymisation initiatives (i.e. Multilingual Anonymisation for Public Administrations[2] (MAPA) project (Ajausks et al. 2020) which provided anonymisation support for all EU languages) and local solutions developed by the project partners. Specifically, Hungarian, Romanian, Bulgarian and Slovak plan to implement local solutions, while Slovenian, Croatian and Polish will use a solution based on the MAPA project. The approaches for all seven languages will be combined in a single user interface and made available via the European Language Grid[3] repository.

3 Conclusions

Since an important aspect of today's neural machine translation technology is the quality of the language model, the envisaged seven language corpora, although monolingual datasets in themselves, can be rightly expected to make an impact on the quality of the eTranslation system through the enhanced language models built with them. Since these corpora in seven languages cover systematically the same four domains, they could be regarded also as comparable corpora for these domains and thus be used for further processing, e.g. in parallel terminology extraction. Moreover, the action addresses the gap in MT technology, which crucially depends on the provision of domain specific quality language resources for the under-resourced languages.

Acknowledgements

The work reported here was supported by the European Commission in the CEF Telecom Programme (Action No: 2019-EU-IA-0034, Grant Agreement No: INEA/CEF/ICT/A2019/1926831) and the Polish Ministry of Science and Higher Education: research project 5103/CEF/2020/2, funds for 2020–2022).

References

Váradi, Tamás; Koeva, Svetla; Yamalov, Martin; Tadić, Marko; Sass, Bálint; Nitoń, Bartłomiej; Ogrodniczuk, Maciej; Pęzik, Piotr; Barbu Mititelu, Verginica; Ion, Radu; Irimia, Elena; Mitrofan, Maria; Păiș, Vasile; Tufiș, Dan; Garabík, Radovan; Krek, Simon; Repar, Andraž; Rihtar, Matjaž; and Brank, Janez. 2020. The MARCELL legislative corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC2020)*, pp. 3761–3768.

Ajausks, Ēriks; Arranz, Victoria; Bie, Laurent; Cerdà-i-Cucó, Aleix; Choukri, Khalid; Cuadros, Montse; Degroote, Hans; Estela, Amando; Etchegoyhen, Thierry; García-Martínez, Mercedes; García-Pablos, Aitor; Herranz, Manuel; Kohan, Alejandro; Melero, Maite; Rosner, Mike; Rozis, Roberts; Paroubek, Patrick; Vasiļevskis, Artūrs; Zweigenbaum, Pierre. 2020. The Multilingual Anonymisation Toolkit for Public Administrations (MAPA) Project. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT2020)*, pp. 471–472.

[1] https://universaldependencies.org/ext-format.html

[2] https://mapa-project.eu

[3] https://www.european-language-grid.eu