Roxana Patraș, Loredana Cuzmici
(editors)

TOOLS, METHODS, AND SOLUTIONS FOR THE EXPLORATION OF ROMANIAN CORPORA

INSTITUTUL EUROPEAN
2024

Table of Contents

INTRODUCTORY NOTES

RELATE: a modern processing platform for Romanian language

Vasile PĂIȘ, Radu ION,
Andrei-Marius AVRAM,
Maria MITROFAN, Dan TUFIȘ

1. Introduction

Language Technology (LT) platforms provide services for the analysis and production of written or spoken language, making use of artificial intelligence (AI) methods to do so. Furthermore, these platforms have been developed either to showcase new technologies or to function as powerful tools for processing large amounts of data, in the form of offline corpora or online requests.

One of the principal goals of the 1st International Workshop on Language Technology Platforms (IWLTP 2020) was to address fragmentation in the language technology landscape. As the workshop organizers noted, instead of competing with one another, platforms should be designed to be interoperable and to interact with each other to create synergies towards a productive LT ecosystem (Rehm *et al.* 2020a). Interoperability is usually achieved by providing input and output in standardized formats, specific to the tasks at hand. Interaction between systems involves making the functionalities of one system available for use by another system.

In this paper, we present the RELATE platform (Păiș *et al.* 2020), a modular state-of-the-art platform that is used for processing the Romanian language developed at the Research Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy (for short RACAI in English, or ICIA in the Romanian language). Integrating resources and technologies developed in our institute as well as those developed by partner institutions, RELATE is actively being used in multiple national and international research projects. From the beginning, it was

designed to use standardized file formats, thus ensuring interoperability with other language processing systems. Internal functions are available as JSON REST web services, thus allowing for interaction with other systems. In the Representational State Transfer (REST) architectural style, data and functionality are considered resources and are accessed using Uniform Resource Identifiers (URIs).

This paper is organized as follows: section 2 presents related work in the field of LT platforms; section 3 presents the history of how the RELATE platform developed; section 4 introduces RELATE's architecture; section 5 describes the platform's components; section 6 presents several scenarios in which the platform might be further used; and we draw conclusions in section 7.

2. Related work

META-SHARE[1] (Federmann *et al.* 2012), CLARIN[2] and ELRC-Share[3] (European Language Resource Coordination Share) are publicly available European websites for research and development in the field of natural language processing, allowing access to language resources. Both ELRC-Share and META-SHARE offer advanced search facilities through which one can easily find various language tools and corpora (e.g. text corpora, annotated corpora, audio corpora) in any (European and other) language. Complex processing pipelines, like NLP-Cube (Boroș *et al.* 2018a) and TTL are able to perform tokenization, lemmatization, POS tagging, chunking and dependency parsing, but they require programming knowledge in order to integrate these functions into an application. A more recently created web service, TEPROLIN, integrates multiple tools — including NLP-Cube, TTL and solutions for named entity recognition (Păiș 2019) and biomedical named entity recognition — into an easy-to-use pipeline. However, this web service still requires users to have programming knowledge in order to engage in web service communication.

GATE (Cunningham *et al.* 2002) and TextFlows (Perovšek *et al.* 2016) aim to make the composition of the language processing chains more user-friendly. The

[1] http://www.meta-share.org.
[2] https://www.clarin.eu.
[3] https://elrc-share.eu/.

graphical interfaces thus allow text-processing widgets to be dragged and dropped into a graphical processing workflow. However, text processing output is not enhanced with specialized visualization tools that would allow access to the computational resources used for annotation. Moreover, they only focus on purely textual content, while multimodal content (such as a combination of text and audio) is not handled within the platforms.

CoBiLiRo (Cristea *et al.* 2020) is a storage platform for multimodal (text and audio) corpora. It was developed in the context of the RETEROM[4] project for storing Romanian speech corpora in a way that was suitable for the project's purposes. It allows a large number of metadata fields to be defined, but it does not enable any complex language processing tasks.

RELATE specifically serves to carry out automatic text processing, with annotations at multiple levels, as well as annotation visualization and expansion into the corresponding linguistic computational resources. The platform focuses on the interactive user experience, through web-based interfaces. Unlike other platforms, such as the one designed by Che *et al.* (2010), RELATE is not primarily focused on exposing APIs, even though API access to platform functionalities is available. Text-processing APIs are used internally and are accessible by platform components. Most of these APIs can also be invoked externally.

RELATE makes use of a common internal format, comprising multiple files (text, standoff metadata, CoNLL-U Plus[5] annotations). The processing workflow is built by adding individual tasks in the graphical user interface. Each task is able to both use and also produce data according to the common internal format. For this reason, no workflow editor, such as the one used in TextFlows (Perovšek *et al.* 2016), is currently available.

Coleman *et al.* (2020) describe a platform for integrating multiple Machine Translation (MT) models. In RELATE, we provide those MT capabilities that are related to the Romanian language (currently supporting Romanian-English and English-Romanian translation only). Furthermore, in RELATE, translated content

[4] https://racai.ro/p/reterom/.
[5] https://universaldependencies.org/ext-format.html.

can be used as input for further language processing tasks. Rebai *et al.* (2020) describe a platform that integrates a voice assistant for improving efficiency and productivity in business. In the case of RELATE, we only integrate existing Romanian (and English) Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) systems while still focusing on the textual components of a corpus. Thus, ASR output can be used as input to further language processing tasks.

Rehm *et al.* (2020b) acknowledge that a large number of AI platforms are currently being developed, both nationally—supported through state funding programmes—and internationally, supported by the European Union. The authors further recognize the enormous fragmentation of the European AI and LT landscapes and believe that modern platforms should facilitate exchange information, data and services, in order to enable interoperability. We agree with this assessment and, in the RELATE platform, we therefore sought to make use of standardized formats as well as decoupling functionalities into components, which could be invoked externally if needed. In addition, similar to the AI4EU[6] and European Language Grid (ELG)[7] platforms, RELATE is able to integrate Docker containers for language processing tools. Such integration is not required, however, and only a small number of the available components are therefore integrated as containers.

3. Evolution of the RELATE platform

The RELATE platform was developed in the context of a number of national and international research projects and evolved according to the needs of the activities for which it was ultimately used. This section describes its evolution throughout these projects, from the first implementation to the current version.

RELATE platform development (Păiș *et al.* 2019) started in the context of the national RETEROM project, which began in 2018. One of the project's main goals, linked to the sub-project TEPROLIN, was to develop and integrate state-of-the-art technologies for Romanian natural language processing, such as those devoted to tokenization, part-of-speech tagging, dependency parsing, phonetic annotations, and

[6] https://www.ai4eu.eu/.
[7] https://www.european-language-grid.eu/.

named entity recognition. The primary result was the TEPROLIN web service (Ion 2018) which allows invocation that uses a raw text document and produces different levels of annotation (based on specified parameters) encoded using a custom JSON format.[8]

The first implementation of the RELATE platform allowed for the management (upload, download, editing) of large corpora and parallel processing using the TEPROLIN service. For processing purposes, we implemented a task management engine, which allowed documents to be distributed across any number of TEPROLIN processes that begin on the same server or over the network. As a result, the processing speed can be increased when needed, because the number of processing nodes can be adjusted dynamically. Furthermore, the platform is able to convert the custom JSON encoding associated with TEPROLIN into a more standard CoNLL-U format[9] (also used by the Universal Dependencies project[10]) or its extension CoNLL-U Plus,[11] when appropriate.

To allow human users to manually explore annotation results, we developed graphical representations, such as visualization in different column-based and JSON formats, tree-based representations for dependency parsing, and the highlighting of recognized named entities. In order to improve the user experience, we also included different interface elements that allow the Representative Corpus of Contemporary Romanian Language (CoRoLa) to be queried (Tufiș *et al.* 2019b). RELATE includes links to the main query interface of CoRoLa's text component, which are accessible through the KorAP corpus analysis platform (Banski *et al.* 2012), and to the speech component, which allows users to search audio files (Boroș *et al.* 2018b) and listen to words being pronounced by speakers of Romanian. Different visualizations also make use of word embeddings (Bojanowski *et al.* 2017) computed on the CoRoLa corpus (Păiș and Tufiș 2018) to suggest words that appear in similar contexts.

We integrated the Romanian WordNet (Tufiș and Mititelu 2015) into the RELATE platform to allow words to be searched online. Furthermore, we exploited

[8] http://relate.racai.ro/?path=teprolin/doc_dev.
[9] https://universaldependencies.org/format.html.
[10] https://universaldependencies.org/.
[11] https://universaldependencies.org/ext-format.html.

the alignment between the Romanian and English (Miller 1995) wordnets in order to provide aligned queries based on *synset* identifiers (in a WordNet, synsets represent sets of synonyms that share a common meaning).

A textual translation component—which was developed by TILDE with the involvement of ICIA, in the context of the project entitled "CEF Automated Translation toolkit for the Rotating Presidency of the Council of the EU"—was integrated into RELATE through the TILDE Machine Translation API.[12] This component allows users to translate documents directly within the platform. After a successful translation from English to Romanian, the resulting document can be analyzed using the platform's text analysis functionality.

As part of the "Multilingual Resources for CEF.AT in the legal domain" project (MARCELL),[13] a terminology annotation tool (Coman *et al.* 2019a) was developed, and we subsequently integrated it into the RELATE platform to facilitate the identification of terms available in the EuroVoc[14] and IATE[15] (Interactive Terminology for Europe) terminological databases. This annotation tool can be used following a lemmatization and part-of-speech tagging operation.

The ASR system (Avram *et al.* 2020b) that resulted from the national project ROBIN[16] and was initially developed for human-robot interaction (Ion *et al.* 2020; Tufiș *et al.* 2019a) was also integrated into the RELATE platform. This system allows text to be extracted from speech recorded directly in the platform or from uploaded audio files and then to be analyzed with the platform's processing tools. Several corrections on the recognized text (such as basic comma restoration and truecasing) can be performed before the actual analysis is carried out (Avram *et al.* 2020a).

We further exploited the availability of ASR and text translation features in order to provide speech-to-speech translation functionality for Romanian-English and English-Romanian. This brought together different platform modules and required the integration of additional text-to-speech components. To synthesize the

[12] https://www.tilde.com/developers/machine-translation-api.
[13] https://marcell-project.eu/.
[14] https://eur-lex.europa.eu/browse/eurovoc.html.
[15] https://iate.europa.eu/home.
[16] http://aimas.cs.pub.ro/robin/en/.