Activitatea 3.7: Definitivarea, testarea, validarea şi împachetarea într-o soluţie „ready-to-use” a platformei integrate şi configurabile de prelucrare a textelor în limba română

1. Platforma de prelucrare a textelor TEPROLIN

Platforma de prelucrare a textelor TEPROLIN a fost îmbunătăţită după cum urmează:

1. *Am introdus dependenţe de tip graf între operaţiile de prelucrare a textelor*. Acest tip de a preciza ce operaţii trebuie rulate mai întâi pentru a se putea rula operaţia de prelucrare dorită e mult mai eficient decât tipul de rulare în secvenţă pe care se baza platforma. De exemplu, pentru a putea rula operaţia de adnotare cu etichete morfo-sintactice, nu mai sunt necesare operaţii precum silabificarea sau detecţia accentului. Modulul Python 3 în care sunt precizate aceste operaţii este `TeproAlgo.py` iar metoda se numeşte `_assignAlgorithmsToOperations()`.
2. *Am adăugat modulul de prelucrare a textelor UDPipe* (http://ufal.mff.cuni.cz/udpipe/1) ca alternativă la TTL şi NLPCube. Este foarte rapid (cea mai rapidă componentă de prelucrare din cele trei) şi are performanţe bune. A fost configurat ca modul implicit dacă se solicită operaţii precum adnotare cu etichete morfo-sintactice sau analiză cu relaţii de dependenţă sintactică.
3. *Modulul de inserare a diacriticelor* `DiacRestore.py` *a fost îmbunătăţit* pentru a detecta mai bine când un text este scris fără diacritice sau cu puţine diacritice şi a rula, astfel, automat pentru a insera diacriticele lipsă.
4. Am eliminat numărarea caracterelor din modulul de statistici pentru că frecvenţa caracterelor prelucrate de TEPROLIN putea creşte foarte mult când se prelucrau texte de zeci de milioane de cuvinte. Au rămas statisticile despre numărul de accesări la serviciu şi despre numărul de cuvinte prelucrate pe zi.

Păiş et al., (2020) descriu un studiu de caz în care TEPROLIN rulează pe mai multe fire de execuţie în RELATE şi adnotează corpusul legislativ din proiectul MARCELL. În total, au fost

adnotaţi aprox. 456 de milioane de tokeni, *ceea ce demonstrează că testarea şi validarea platformei s-au încheiat cu succes*.

2. Soluţia „ready-to-use” a platformei TEPROLIN

TEPROLIN se poate utiliza într-unul din următoarele patru moduri:

1. Pentru testare cu fraze scurte (pentru evaluarea performanţelor) cu efectuarea tuturor operaţiilor disponibile, se poate accesa link-ul https://relate.racai.ro/index.php?path=teprolin/complete şi se pot vizualiza adnotările făcute;
2. Pentru rularea unor operaţii la alegere pe fraze scurse, folosind algoritmii preferaţi, se poate accesa link-ul https://relate.racai.ro/index.php?path=teprolin/custom;
3. Pentru adnotarea corpusurilor cu mai mult de 1000 de cuvinte, se poate solicita acces la platforma RELATE care rulează TEPROLIN pe mai multe fire de execuţie;
4. Ca modul Python 3, clonând repository-ul https://github.com/racai-ai/TEPROLIN şi urmând indicaţiile din fişierul `README.md`. Recomandăm ca toate pachetele necesare să fie instalate într-un mediu dedicat Python 3 (eng. „virtual environment”), executând comenzile:
   a. `python3 -m venv /calea/către/mediul/dedicat/teprolin`
   b. `pip3 install -r requirements.txt`

2.1 Download şi acces public

Versiunea publică (şi finală) a platformei TEPROLIN se află pe GitHub, la adresa https://github.com/racai-ai/TEPROLIN. O versiune anterioară mai veche, care conţine software proprietar, se află la https://gitlab.com/raduion/teprolin. Pentru a avea acces la această versiune, trebuie să aveţi cont pe GitLab şi să solicitaţi accesul autorului platformei.

3. Referinţe

Păiş, V., Tufiş, D. şi Ion, R. (2020) A Processing Platform Relating Data and Tools for Romanian Language. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC), European Language Resources Association, Marseille, France, pages 81—88.