ACADEMIA ROMÂNĂ

SCOSAAR

REZUMATUL TEZEI DE ABILITARE

TITLUL Contributions to natural language processing with applications on the Romanian language

Domeniul de abilitare: *Calculatoare, tehnologia informaţiei şi ingineria sistemelor*

Autor: PĂIȘ VASILE-FLORIAN

Contents

Această teză de abilitare este organizată în 7 capitole. Capitolul 1 prezintă o introducere în domeniu și în specificul cercetării personale.. Capitolul 2 conține descrierile resurselor dezvoltate de mine sau cu participarea mea. Capitolul 3 descrie tehnologii și algoritmi creați de mine sau cu implicarea mea. Capitolul 4 este dedicat modelelor de inteligență artificială pentru limba română create de mine sau cu implicarea mea. Capitolul 5 descrie platforma RELATE. În Capitolul 6 prezint perspectivele de cercetare viitoare, pentru a concluziona apoi în Capitolul 7. La finalul tezei sunt incluse anexe cu liste de figuri, tabele, ecuații, algoritmi și acronime.

1. Introducere

Procesarea limbajului natural (pe scurt PLN sau NLP, din engleză „Natural Language Processing") este un domeniu cheie al cercetării în domeniul inteligenței artificiale (IA sau AI, din engleză „Artificial Intelligence"). Se ocupă de prelucrarea limbajului în diferitele sale forme: scris și vorbit. Datele aferente sunt reprezentate în computere sub formă de fișiere. Acestea includ fișiere text, fișiere audio, imagini cu text și fișiere video. Luând în considerare diferitele modalități de reprezentare, distingem între resurse și algoritmi care se ocupă cu date unimodale (fie text, imagine sau vorbire) sau date multimodale (combinații ale diferitelor reprezentări, cum ar fi un videoclip care conține text în imagini, vorbire asociată și subtitrări).). Pe măsură ce ne îndreptăm atenția către resursele multimodale mari, algoritmii de IA de „învățare profundă" de ultimă generație (SOTA, din engleză „state-of-the-art") necesită cantități mari de putere de calcul, de obicei sub forma mai multor unități de procesare grafică (GPU, din engleză Graphic Processing Unit) sau unități de procesare tensorială (TPU, din engleză Tensor Processing Unit), instalate în clustere de procesare. Algoritmii sunt orientați spre analiza limbajului sau transformarea modalității. Analiza limbajului include operațiuni precum extragerea informațiilor, clasificarea, regăsirea documentelor. Transformarea modalității include sarcini precum conversia vorbirii în text (recunoaștere automată a vorbirii - ASR, din engleză Automatic Speech Recognition), text în vorbire (TTS, din engleză Text-to-Speech), text în imagine, imagine în text (generare de descriere), text în video. Pe măsură ce AI generativă avansează, sunt abordate și alte sarcini, inclusiv generarea de modele 3D din text, generarea (sau optimizarea) circuitelor electronice, scrierea programelor pe baza descrierii problemei, generarea interfeței grafice pentru utilizatori (GUI, din engleză Graphical User Interface). Aplicabilitatea algoritmilor SOTA AI continuă să se extindă și sunt explorate noi aplicații.

2. Resurse de limbă

2.1. ROBIN Technical Acquisition Speech Corpus (RTASC)

Corpusul ROBIN Technical Acquisition Speech Corpus (RTASC) [3,4] a fost dezvoltat pentru a evalua și îmbunătăți interacțiunea om-mașină în contextul proiectului ROBIN. În cadrul acestui proiect a fost creat un agent de conversație, simulând un asistent de

magazin care încearcă să ajute un client la achiziționarea de echipamente electronice, precum laptop-uri. Este un corpus de vorbire citită, creat de 6 participanți. Fiecare participant a primit un număr de propoziții care trebuiau citite în fața unui microfon. Setul de date rezultat are 6,5 ore de vorbire în limba română, specific unui domeniu tehnic. Setul de date este echilibrat în funcție de gen (3 bărbați și 3 femei). Fișierele audio sunt în format WAV, cu o rată de eșantionare de 44,1 KHz. Corpusul are 3.786 de segmente audio, fiecare cu un fișier text corespunzător, aliniat. Înregistrările au fost create în platforma RELATE (vezi Capitolul 5).

Corpusul a fost utilizat pentru antrenarea unui sistem ASR, bazat pe arhitectura Deep Speech 2 [5]. Experimentele au arătat că includerea vorbirii specifice domeniului, din corpusul RTASC, a îmbunătățit capacitățile ASR cu 16,3% WER și 7,8% CER în comparație cu un sistem de bază antrenat similar (fără datele RTASC). Experimentele detaliate sunt prezentate în [4]. Rezultatele au arătat importanța includerii datelor specifice domeniului atunci când se antrenează sisteme ASR. Ele arată, de asemenea, lipsa unor astfel de date în seturile de date vocale existente pentru limba română, oferind astfel o direcție pentru viitoare proiecte de cercetare care vizează îmbunătățirea performanței ASR pentru limba română. Setul de date a fost publicat inițial în platforma Zenodo[1] și apoi a fost indexat în European Language Grid (ELG)[2].

2.2. Romanian Named Entity Recognition in the Legal Domain (LegalNERo)

Corpusul "Romanian Named Entity Recognition in the Legal Domain" (LegalNERo) [1,2] este prima resursă în limba română din domeniul juridic pentru recunoașterea entităților numite (NER, din engleză Named Entity Recognition). Resursele specifice domeniului sunt foarte importante pentru algoritmii NLP, deoarece acoperă anumite cuvinte, expresii și termeni specifici care nu sunt folosite în mod uzual (sau mai puțin frecvent utilizate) în corpusurile de domeniu general. Astfel de resurse sunt folosite pentru a evalua algoritmi de domeniu general și pentru a dezvolta sau ajusta algoritmi și modele specifice domeniului juridic. LegalNERo a fost adnotat manual cu 5 clase de entități: organizații, locații, persoane, expresii de timp și referințe legale. Referințele juridice au fost rafinate în continuare folosind tipuri de documente juridice cu granulație fină: Lege, Ordonanță, Publicare, Decret, Decizie, Tratat, Raport, Ordin, Regulament, Directivă, Ordonanță de urgență, Normă, Convenție, Cod și altele. În plus, entitățile de tip locație au fost legate la ontologia GeoNames, acolo unde aceste legături puteau fi stabilite automat. Corpusul a fost creat în platforma RELATE, folosind componenta de adnotare integrată BRAT [6]. Fișierele text au fost extrase din corpusul MARCELL-RO [7].

2.3. Romanian micro-blogging named entity recognition (MicroBloggingNERo)

MicroBloggingNERo [10,11] este un corpus de text în limba română colectat de pe platformele de social media de tip microblogging (mesaje scurte), cum ar fi Twitter/X. A fost adnotat manual pentru entitățile numite. Deoarece am dorit să permitem utilizarea acestui set

[1] https://zenodo.org/records/4626540
[2] https://live.european-language-grid.eu/catalogue/corpus/13123

de date pentru mai multe sarcini, am stabilit o schemă de adnotare care cuprinde 9 clase de entități. Am folosit cele 5 clase de entități (persoană, organizație, locație, expresie în timp și referință legală) din corpusul LegalNERo (vezi Secțiunea 2.2) și am adăugat 4 clase suplimentare legate de sănătate (tulburări, substanțe chimice, dispozitive medicale și părți anatomice). Clasele legate de sănătate au fost parțial inspirate de corpusul SiMoNERo [12,62]. Permițând ca setul de date să utilizeze clase similare cu alte seturi de date, am sperat să facem ca modelele rezultate să funcționeze mai bine și să aibă rezultate îmbunătățite în toate domeniile de cercetare asociate corpusurilor: social media, domeniul legal, domeniul biomedical.

Textul din rețelele sociale are caracteristici unice: emoticoane, emoji, hashtag-uri, text cu majuscule necorespunzătoare (uneori toate literele majuscule sunt folosite pentru a sublinia mesajele), text incorect din punct de vedere gramatical (oamenii nu folosesc verificatoare gramaticale sau scriu în grabă), abrevieri (uneori abrevieri mai puțin frecvente), lipsa semnelor diacritice. Emoticoanele sunt formate prin utilizarea semnelor de punctuație grupate împreună (cum ar fi „ :) ” sau „ :-) ”). Emoji-urile sunt caractere speciale UTF-8 care au pictograme asociate, cum ar fi fețele zâmbitoare. Atât emoticoanele, cât și emoji-urile sunt folosite pentru a transmite sentimentele utilizatorului sau pentru a reprezenta o parte a mesajului. De exemplu, în loc să spună „hai să bem o cafea” , utilizatorul poate folosi pur și simplu un emoji reprezentând o ceașcă de cafea.

2.4. USPDATRO: Underrepresented Speech Dataset from Romanian language Open Data

Setul de date USPDATRO [13] a fost creat pentru a aborda o problemă importantă prezentă în seturile de date uzuale de voce în limba română, aceea a categoriilor slab reprezentate de vorbitori: voci feminine, tineri, bătrâni. În mod obișnuit, seturile de date de vorbire au fost create prin înregistrarea vocii participanților la diferite proiecte (această abordare a fost adoptată și pentru setul de date RTASC, descris în Secțiunea 2.1). Pentru USPDATRO, am adoptat o abordare diferită. Am explorat platforme de partajare a conținutului multimedia (cum ar fi YouTube) pentru utilizarea conținutului deja creat. Aceasta reprezintă o metodă nouă care poate fi aplicată și altor limbi cu mai puține resurse, cu condiția să fie disponibil online suficient conținut. Ne-am concentrat eforturile de cercetare pe 5 platforme cunoscute de partajare a conținutului: YouTube, Vimeo, TikTok, SoundCloud și LinguaLibre. În plus, am vizat numai conținutul lansat în mod special sub licențe deschise, cum ar fi Creative Commons. Acest lucru ne-a determinat să folosim numai YouTube, Vimeo și SoundCloud datorită disponibilității conținutului deschis. Utilizatorii care își lansează conținutul sub o licență deschisă trebuie să selecteze manual licența Creative Commons, indicând astfel că sunt conștienți și sunt de acord să-și ofere conținutul sub licența specificată.

2.5. RoMEMEs

RoMEMEs [14] este un set de date de memeuri românești colectate din Internet. Acestea sunt imagini, adesea combinate cu text, care se răspândesc rapid în locații online, cum ar fi rețelele sociale, blogurile, site-urile web. De obicei, sunt menite să fie amuzante, dar pot fi folosite împotriva persoanelor (cum ar fi politicienii) sau a grupurilor (companii, organizații, partide politice). Am adunat doar memeuri care conțin text în limba română sau imagini specifice românești (cum ar fi persoane publice din România). Pentru fiecare imagine, am adăugat mai multe niveluri de adnotări, folosind adnotatori umani: textul asociat extras din imagine, complexitatea memei (fie imagine simplă, fie combinație de mai multe imagini), autenticitatea imaginii, polaritate, emoție, clasificare politică. Pentru autenticitatea imaginii, distingem între imaginile reale, imaginile false și imaginile false profunde („deep fakes"). Deoarece la momentul adnotării nu deținem această informație, am folosit bunul simț pentru a identifica falsurile profunde. Astfel, dacă o imagine părea a fi reală, fără nicio manipulare evidentă, dar în contradicție cu faptele cunoscute (de exemplu, un politician cunoscut avea un corp feminin), atunci am adnotat o astfel de imagine ca deep fake. Polaritatea a fost concentrată în jurul unui sentiment pozitiv, neutru sau negativ exprimat de memă. Pentru clasificarea emoțiilor am folosit emoțiile descrise de Parrot [15]: bucurie, dragoste, frică, furie, surpriză, tristețe. Atât pentru clasificarea polarității, cât și a emoțiilor, nu am luat în considerare partea amuzantă a memelor, ci ne-am concentrat pe mesajul propriu-zis transmis către persoana, organizația sau evenimentul descris.

2.6. CoRoLa frequency lists

Corpusul de referință pentru limba română contemporană (CoRoLa) [16] a fost construit ca proiect prioritar al Academiei Române. Conține atât texte scrise, cât și înregistrări orale, care urmăresc să acopere principalele stiluri de limbaj funcțional (juridic, științific, jurnalistic, imaginativ, memorii, administrativ), în patru domenii (arte și cultură, natură, societate, știință) și în 71 de subdomenii, ținând cont de drepturile de proprietate intelectuală (IPR, din engleză Intellectual Property Rights). Pe parcursul anilor de doctorat am fost implicat în crearea corpusului CoRoLa și a interfețelor de interogare [17]. Cu peste 1 miliard de cuvinte (scrise și vorbite), CoRoLa este unul dintre cele mai mari corpusuri de referință din lume, complet libere de IPR. Corpusul în sine este accesibil oricăror utilizatori, în scopuri de cercetare, utilizând interfețele sale de interogare online. Fișierele de date sunt accesibile partenerilor de proiect și entităților care au solicitat și au primit acces la date de cercetare. În urma tezei mele de doctorat, am fost implicat în crearea mai multor algoritmi care utilizează informații derivate din CoRoLa și modele generate. Deoarece statisticile derivate sunt utile pentru diferiți algoritmi de procesare a limbajului natural, am creat ulterior un set de date format din liste de frecvențe extrase din CoRoLa.

2.7. Resurse în limba română în format „linked data"

Tehnologiile de tip Web Semantic urmăresc să facă datele mai ușor de utilizat de către aplicații. Acestea combină tehnologii precum Resource Description Framework (RDF) și

Web Ontology Language (OWL) pentru a produce seturi de date bogate care pot fi încărcate și interogate mai ușor de aplicații generice. În plus, aceste tehnologii permit conectarea mai multor seturi de date într-o pseudo-bază de date mare. Interogările pot fi efectuate în seturi de date multiple, permițând astfel mai multe modalități de exploatare a datelor, care nu au fost luate în considerare de către creatorii individuali ai seturilor de date. RDF folosește conceptul de „triplet" (sub forma Subiect, Verb, Obiect) pentru reprezentarea datelor. Tripletele pot fi folosite pentru a conecta concepte în interiorul unui set de date sau între mai multe seturi de date. OWL este folosit pentru a descrie reguli cunoscute sub numele de relații semantice, exprimate și ca triplete.

Pentru a face resursele în limba română ușor de descoperit și utilizat, am migrat mai multe seturi de date existente către structuri specifice LLOD (din engleză Linguistic Linked Open Data). Un exemplu este oferirea setului de date LegalNERo în format LLOD, împreună cu formate mai tradiționale, așa cum este descris în Secțiunea 2.2. Am migrat următoarele resurse: WordNet românesc, liste de frecvențe CoRoLa (acestea au fost descrise în Secțiunea 2.6), corpusul RRT, corpusul SiMoNERo, corpusul PARSEME-Ro, setul de date de vorbire RTASC (descris în Secțiunea 2.1) și lexiconul RoLex . Procesul de conversie este detaliat în [19,21] și scripturile utilizate în timpul procesului de conversie, care pot fi aplicate și altor resurse, sunt disponibile în GitHub[3]. Resursele LLOD românești sunt disponibile online și pot fi descărcate gratuit[4].

2.8. Corpusul legislativ românesc MARCELL

"Multilingual Resources for CEF.AT in the legal domain (MARCELL)" a fost un proiect care viza îmbunătățirea traducerii automate folosind corpusul legislației naționale (legi, decrete, reglementări) din șapte țări: Bulgaria, Croația, Ungaria, Polonia, România, Slovacia și Slovenia. În acest scop, sub auspiciile Connecting Europe Facility - Automatic Translation (CEF.AT), am construit un corpus comparabil mare din cele șapte limbi, aliniate la domeniile de nivel tematic identificate de descriptorii EUROVOC. Pe lângă îmbunătățirea generală preconizată a sistemului de traducere automată (MT, din engleză Machine Translation) în cele șapte limbi în cauză, acțiunea era de așteptat să aibă un impact atât asupra e-justiției, cât și asupra infrastructurilor de servicii digitale (DSI, din engleză Digital Service Infrastructure), deoarece resursele se concentrează pe legislația națională, care are relevanță directă pentru DSI. Subcorpusul legislativ român (MARCELL-RO) [53, 7] al corpusului MARCELL mai mare conține 163.274 de fișiere, care reprezintă corpusul legislației naționale cuprinsă între 1881 și 2021. Acest corpus cuprinde în principal: hotărâri guvernamentale, ordine ministeriale, hotărâri, decrete și legi. Toate textele au fost obţinute prin crawling de pe portalul public legislativ român. Nu am făcut distincție între legile în vigoare și cele vechi, deoarece este dificil să facem acest lucru automat și nu există nici o resursă externă de folosit pentru a face distincția între ele. Textele au fost extrase din formatul HTML original și convertite în fișiere TXT, prin procese automate.

[3] https://github.com/racai-ai/RoLLOD
[4] https://www.racai.ro/p/llod/

2.9. Corpusul românesc CURLICAT

"Curated Multilingual Language Resources for CEF AT (CURLICAT)" [30] a fost un proiect care își propunea să adune seturi de date organizate în șapte limbi vizate de consorțiu (bulgară, croată, maghiară, poloneză, română, slovacă și slovenă) în domenii relevante pentru infrastructurile europene de servicii digitale (DSI) cu obiectivul declarat de a îmbunătăți Traducerea automată în cadrul CEF AT. Sursa principală de date au fost corpusurile naționale ale partenerilor consorțiului. Pentru limba română, aceasta înseamnă corpusul CoRoLa. Anonimizarea [31] a fost folosită pentru a elimina datele personale și sensibile din resursele lingvistice colectate. În plus, metadatele au fost armonizate între corpusurile monolingve produse. În cele din urmă, fiecare corpus specific limbii a fost îmbogățit cu termeni IATE. Acest ultim pas, împreună cu o distribuție echilibrată a domeniilor, contribuie la exploatarea resursei ca un corpus comparabil mare, util pentru traducerea automată.

Corpusul CURLICAT este disponibil în platforma ELRC-Share, incluzând subcorpusul de limba română[5]. Resurse adiționale, incluzând baze de date de terminologie (în formatele CSV și TBX) sunt disponibile de pe site-ul web al proiectului[6]. Curpusurile au fost indexate de European Language Grid[7]. Corpusul de limba română conține 26k documente, 3.56M propoziții și respectiv 95.10M tokeni.

3. Tehnologii de limbă și algoritmi

3.1. Recunoașterea automată a entităților

Cercetarea pe care am efectuat-o pentru teza mea de doctorat s-a concentrat pe sistemele NER generale. Se foloseau atunci reprezentări vectoriale statice asociate cuvintelor (în engleză „static word embeddings"). În următorii ani, cercetarea mea în NER s-a concentrat pe implementarea de noi tehnologii SOTA pentru limba română, precum și pe producerea de resurse și sisteme pentru NER specifice diferitelor domenii. Sistemele generale pot fi aplicate textului specific unui domeniu, dar performanța acestora este mai scăzută în comparație cu cea obținută pe textele generice pe care au fost dezvoltate, instruite și testate. Acest lucru se datorează prezenței cuvintelor, expresiilor și entităților numite (NE, din engleză Named Entities) specifice domeniului. O dată cu crearea setului de date LegalNERo (vezi Secțiunea 2.2) am explorat NE în domeniul juridic. Am extins acest lucru cu MicroBloggingNERo (vezi Secțiunea 2.3) unde am inclus entități legate de sănătate, specifice domeniului biomedical, inspirate din corpusul SiMoNERo (dezvoltat de colegii de la ICIA). NER este un pas important pentru alte sarcini NLP, cum ar fi extragerea informațiilor, regăsirea documentelor și anonimizarea. Anonimizarea este o altă sarcină care a constituit un

[5] https://elrc-share.eu/repository/browse/curlicat-romanian-corpus/8b6c8dca58ea11ed9c1a00155d026706fb03ef8b4c1847cfbe9cea869a82731e/
[6] https://curlicat-project.eu/deliverables.html
[7] https://live.european-language-grid.eu/catalogue/search/curlicat

obiect pentru cercetarea mea și este descrisă în Secțiunea 3.2. Acuratețea NER este un factor cheie în îmbunătățirea rezultatelor anonimizării, deoarece multe dintre persoanele, organizațiile sau entitățile de locație identificate trebuie anonimizate, împreună cu alte entități specifice domeniului.

3.2. Anonimizare

NER și anonimizarea sunt sarcini NLP asemănătoare, deși diferite. NER își propune să identifice toate NE prezente în text. Anonimizarea are ca scop înlocuirea doar a entităților relevante cu un șir de caractere nou, pentru a transforma textul în așa fel încât să fie imposibilă recuperarea informațiilor personale din textul rezultat. Astfel, anonimizarea este o sarcină mai grea în comparație cu NER. Entitățile identificate trebuie clasificate dacă sunt relevante pentru procesul de anonimizare. Acest lucru se face în conformitate cu legile și reglementările în vigoare aplicabile textului și poate fi dependent de domeniu (se pot aplica diferite reguli pentru textul general, textul din domeniul juridic și textul legat de sănătate). În plus, tipurile reale de entități care necesită anonimizare pot depinde de domeniul textului. În textul general, suntem preocupați de entități și expresii comune: persoane, locații specifice (cum ar fi adresele de domiciliu), date specifice (cum ar fi datele de naștere), organizații (în special în relația cu indivizii, în expresii precum „fondatorul Organizației"). În alte tipuri de text, clasele de entități suplimentare pot deveni relevante, cum ar fi numerele personale de identificare, telefoanele, adresele de e-mail, identificatorii de documente, conturile bancare.

3.3. Procesarea vorbirii

Vorbirea joacă un rol cheie în comunicarea inter-umană. Din perspectiva computerului, vorbirea poate fi folosită de utilizatori umani pentru a interacționa cu aplicațiile. Facem distincție între două sarcini principale: recunoașterea automată a vorbirii (ASR) și sinteza text-to-speech (TTS). Alte sarcini, cum ar fi identificarea cuvintelor cheie sau analiza sentimentelor bazată pe vorbire, sunt, de asemenea, posibile folosind înregistrările vocale. Interacțiunea cu chatboții sau roboții conversaționali este uneori simplificată prin utilizarea vocii. În proiectul ROBIN a fost implementat un agent conversațional, utilizabil într-un robot de asistență. Acesta permitea unui utilizator uman să interacționeze cu robotul și să ceară sfaturi cu privire la achiziția de calculatoare și alte echipamente tehnice aferente, în cadrul unui magazin specializat. În [35] am dezvoltat un sistem ASR folosind arhitectura DeepSpeech2. Experimentele au continuat în [36] cu diferiți parametri și date suplimentare. În [37] am folosit sistemul ASR descris în [35,36] bazat pe DeepSpeech2 pentru a construi un sistem modular de traducere din vorbire în vorbire (în engleză „speech-to-speech"). În [38] am continuat experimentele noastre ASR cu modele Wav2Vec [44] pre-antrenate. Modelul Wav2Vec este îmbunătățit în continuare în [41] prin includerea unui strat de inhibiție lateral (acest lucru va fi descris în Secțiunea 3.4). În [39] ne-am preocupat de impactul diferitelor componente legate de vorbire în contextul interfețelor om-mașină. În [40], ne concentrăm asupra vorbirii românești slab reprezentate, folosind setul de date USPDATRO (descris în

Secțiunea 2.4) pentru evaluare. În plus, antrenăm un nou model ASR bazat pe arhitectura OpenAI Whisper.

3.4. Inhibarea laterală

În biologie, inhibarea laterală este o formă de comportament inhibitor neuronal, referindu-se la capacitatea neuronilor excitați de a reduce activitatea vecinilor lor [48]. În creierul uman, procesul de inhibiție laterală este întâlnit în zonele senzoriale, cum ar fi regiunile creierului implicate în vedere. Se știe că sporește contrastul și claritatea informațiilor percepute care sunt livrate creierului. La momentul în care investigam inhibiția laterală, nu exista nici o modelare a acestui proces în ceea ce privește sarcinile NLP. Astfel, în [27] am implementat un nou strat de rețea neuronală artificială inspirat de mecanismul biologic al inhibării laterale și l-am aplicat la recunoașterea complexă a entităților numite, rezultând îmbunătățiri în mai multe limbi și texte mixte (conținând cuvinte din mai multe limbi în aceeași propoziție). Pentru sistemul NER dezvoltat în [27], am folosit modelul XLM-RoBERTa ca punct de plecare. Apoi am adăugat stratul de inhibare laterală la finalul XLM-RoBERTa, urmat de un strat liniar și un strat de clasificare. În [29] am aplicat arhitectura NER îmbunătățită cu inhibare laterală la NER în limba română în domeniul biomedical. În [28] am aplicat același sistem pentru NER într-un corpus de mesaje Twitter, cu scopul de a detecta mențiunile de boli. În [50] am extins un sistem pentru identificarea expresiei cu mai multe cuvinte în text prin încorporarea stratului de inhibiție laterală. În [41] am explorat aplicarea inhibării laterale pentru recunoașterea vorbirii.

3.5. Adnotarea pe bază de terminologie

În [51], am dezvoltat un sistem automat de extracție a termenilor ca un ansamblu de 5 algoritmi. Dintre aceștia, 2 algoritmi au fost implementați de noi și descriși în lucrare. Primul algoritm folosește modele despre modul în care se formează termenii pe baza unui set de antrenament. Aceasta include caracteristici precum: simboluri permise într-un termen; cuvinte stop permise la începutul, mijlocul sau sfârșitul termenului; separatori între termeni; sufixe; cuvinte de context. După învățarea caracteristicii, mai multe cuvinte sunt selectate pe baza statisticilor de colocare. Al doilea algoritm folosește reprezentări de încorporare a cuvintelor, obținute din vectori generali și specifici domeniului, mediate peste cuvintele care formează expresii specifice. Presupunerea făcută este că termenii sunt mai aproape de setul de termeni de antrenament decât de cuvintele din domeniul general.

Corpusul legislativ MARCELL [53, 7] a fost îmbogățit automat cu termenii IATE[8] și EuroVoc[9]. Terminologia interactivă pentru Europa (IATE) este baza de date terminologică a UE și este utilizată pentru diseminarea și gestionarea terminologiei specifice UE. Unul dintre obiectivele sale principale este de a facilita sarcina traducătorilor care lucrează pentru UE. În prezent, are peste 8 milioane de termeni și folosește tezaurul EUROVOC ca sistem de clasificare a domeniilor. Având în vedere dimensiunea bazei de date terminologice și

[8] https://iate.europa.eu/home
[9] https://eur-lex.europa.eu/browse/eurovoc.html

numărul mare de documente din corpusul MARCELL, accentul nostru a fost pe dezvoltarea unui sistem de adnotare eficient în timp, așa cum este descris în [7], folosind potrivirea aproximativă a șirurilor. Corpusul a fost lansat în format CoNLL-U Plus, cu termenii adăugați în ultimele două coloane ale fișierului. O abordare similară a fost utilizată în [30] pentru adnotarea părții românești a corpusului CURLICAT cu termeni IATE (spre deosebire de MARCELL, EuroVoc nu a fost folosit pentru CURLICAT). Detaliile implementării sunt date în [54].

4. Modele pentru limba română

Aplicațiile moderne de inteligență artificială, bazate pe rețele neuronale artificiale, se bazează pe modele de limbaj pentru reprezentarea intrării lor. În cazul NLP, intrarea este reprezentată de cuvinte (sau tokeni), astfel încât modelul de limbaj este denumit reprezentare vectorială a cuvintelor. Aceasta poate fi o reprezentare statică (fiecare cuvânt are o singură reprezentare asociată) sau contextuală (fiecare cuvânt are mai multe reprezentări, cea folosită este generată în funcție de context). Pentru vorbire, modelul asociază o reprezentare mostrelor audio de intrare. Pentru conținutul multimodal, modelele pot crea reprezentări pentru diferite tipuri de conținut (de exemplu, imagine și text). Modelele sunt antrenate folosind o rețea neuronală. Reprezentările obținute prin utilizarea modelelor sunt apoi utilizate în rețele neuronale mai complexe pentru a efectua sarcini NLP reale. Modelele sunt fie de uz general, reutilizabile între sarcini (fie direct, fie prin reglare fină - „fine-tuning"), fie specifice sarcinii. În zilele noastre, se obișnuiește să se folosească un model de bază pre-antrenat pe cantități mari de date și apoi să fie ajustat pe anumite limbi, domenii și sarcini. În [29] am antrenat reprezentarea vectorială de cuvinte biomedicale pe corpusul BioRo. În scopul proiectului MARCELL, în [7] am pregătit modele de clasificare pentru categoriile EuroVoc. În [55] ne-am continuat cercetările privind clasificarea EuroVoc și modelele bazate pe BERT ajustate pentru această sarcină. Modelele sunt disponibile pentru 22 de limbi. Modelele de recunoaștere a vorbirii au fost antrenate folosind diferite arhitecturi. În [35,36,37] ne-am ocupat de arhitecturile DeepSpeech2. În [38,41] am folosit modelele pre-antrenate Wav2Vec 2.0 și le-am reglat fin pe limba română. În [40] am ajustat mai multe modele bazate pe arhitectura OpenAI Whisper. Au fost dezvoltate mai multe modele NER specifice domeniului pentru limba română și sunt disponibile în platforma RELATE[10] pentru descărcare sau interogare. În [60], am antrenat modele de limba română pentru mai multe kituri de procesare a limbajului de bază SOTA (BLARK, din engleză Basic Language Annotation Resource Kit).

5. Platforma RELATE

Lucrul la platforma RELATE [69,70,71] a început în timpul studiilor mele de doctorat. A continuat să fie dezvoltată în ultimii ani și se află încă în dezvoltare activă la momentul scrierii acestei teze. Acesta integrează rezultatele din mai multe proiecte de cercetare.

[10] https://relate.racai.ro/index.php?path=ner/demo

Instrumentele și resursele integrate sau puse la dispoziție prin intermediul platformei RELATE provin atât din eforturile de dezvoltare la ICIA, cât și din partea partenerilor din diferitele proiecte de cercetare.

5.1. Arhitectura platformei

Platformele de tehnologie lingvistică (LT) oferă servicii și resurse pentru procesarea limbajului scris sau vorbit. Platformele moderne LT folosesc metode AI pentru a implementa funcționalitățile specifice. Platformele ar trebui să fie construite astfel încât să fie interoperabile și să interacționeze între ele pentru a crea sinergii către un ecosistem LT productiv [72]. Având în vedere acest lucru, platforma RELATE a fost dezvoltată pentru a utiliza formate standardizate, atât pentru intrare, cât și pentru ieșire, permițând schimbul de date între platforme. În plus, funcționalitatea internă poate fi expusă sub formă de servicii web care permit integrarea în alte sisteme. Funcțiile disponibile sunt furnizate de module dezvoltate în mai multe proiecte naționale și internaționale, atât interne, cât și de către instituțiile partenere. Încă de la început, ne-am propus să folosim formate de fișiere standardizate și ușor de utilizat, combinate cu API-uri web, permițând astfel integrarea cu alte sisteme (după cum se discută în [69]), în funcție de necesități. Integrarea componentelor se realizează fie direct, prin consumarea API-urilor furnizate de pe serverul unui partener, fie prin intermediul containerelor Docker, găzduite pe unul sau mai multe servere asociate platformei. Astfel, într-un mod mai simplu, urmează filozofia din spatele European Language Grid.

5.2. Creare de corpusuri

RELATE permite crearea manuală a corpusurilor în interiorul platformei. Oferă funcții precum: adnotare text (utilă pentru o parte de vorbire, expresii cu mai multe cuvinte, corpusuri NER), clasificare a conținutului (utilă pentru clasificarea polarității, clasificarea emoțiilor), înregistrarea vorbirii, transcrierea imaginilor din PDF, metadate create la nivel de document. Atunci când se creează un corpus nou, platforma oferă posibilitatea de a activa diferite caracteristici pentru acel corpus.

5.3. Analiza de corpus

După crearea corpusului (sau încărcarea unui corpus existent), platforma RELATE poate fi utilizată pentru adnotarea și efectuarea automată de analize a fișierelor din corpus. Acest lucru se realizează folosind un mecanism bazat pe sarcini. Utilizatorul poate programa orice număr de sarcini pentru adnotare, clasificare, traducere, recunoaștere a vorbirii, traducere a vorbirii, sinteză a vorbirii. După finalizarea sarcinii, utilizatorul poate programa o sarcină specială pentru calcularea statisticilor pe fișierele rezultate. Sarcinile disponibile depind în primul rând de tipul de corpus (așa cum este descris în secțiunea anterioară). În plus,

dacă rulați o anumită instanță a platformei RELATE (cum ar fi instanța unui adnotator), este posibil ca anumite module să nu fie disponibile, indiferent de tipul de corpus. Interfața grafică cu utilizatorul (GUI) permite setarea diferiților parametri aferenți sarcinii. Aceștia depind de tipul sarcinii.

5.4. Traducere automată în platforma RELATE

Proiectul "CEF Automated Translation toolkit for the Rotating Presidency of the Council of the EU" a avut drept scop să facă platforma eTranslation a Comisiei Europene utilă pentru utilizatorii din țările membre EU, extinzând facilitățile platformei cu un set de mecanisme de traducere automată specifice domeniului președenției UE. Sistemele de traducere Română-Engleză și Engleză-Română au fost îmbunătățite prin dezvoltarea de sisteme calitative de traducere pentru domeniile aferente președenției UE și diferitelor DSI. Pentru limba română, Institutul de Cercetare pentru Inteligența Artificială "Mihai Drăgănescu" a fost implicat și a contribuit la dezvoltarea sistemului de traducere (Ro-En și En-Ro), care este o componentă a sistemului mai mare pentru Președenția Consiliului UE. Actuala platformă MT permite utilizatorilor să traducă documente întregi și site-uri web locale, inclusiv sisteme securizate de traducere automată pentru toate limbile oficiale ale UE. Folosind API-ul TILDE Machine Translation, componenta de traducere textuală pentru Ro-En și En-Ro, a fost integrată în platforma RELATE. Astfel încât utilizatorii să poată traduce documente direct în platformă și, de asemenea, să analizeze documentul rezultat folosind funcționalitățile platformei. Figura 25 prezintă interfața de traducere a textului pentru un singur document. Textul tradus rezultat poate fi descărcat sau transmis ulterior prin platformă pentru analiză. A doua opțiune este disponibilă doar pentru traducerea engleză-română. În acest caz, documentul în limba română rezultat poate deveni intrarea pentru oricare dintre componentele de procesare a limbajului natural integrate în platforma RELATE.

5.5. Catalog de resurse

În cadrul proiectului European Language Equality (ELE), a fost construită o listă de resurse și tehnologii lingvistice pentru toate limbile europene. Scopul principal a fost acela de a evalua dezvoltarea diferitelor domenii ale cercetării NLP în ceea ce privește limbile europene. Un rezultat cheie a fost o planificare strategică pentru atingerea egalității lingvistice digitale în Europa. Aceasta este detaliată în cartea „European language equality: a strategic agenda for digital language equality" [79], care conține un capitol despre statutul fiecărei limbi, inclusiv limba română [80]. Lista de resurse a fost integrată în European Language Grid (ELG) pentru a contribui la descoperirea și reutilizarea resurselor și instrumentelor lingvistice existente. Cu toate acestea, în timpul investigației noastre, a devenit evident că în România nu există o platformă care să indexeze resursele disponibile în limba română. Prin urmare, am extins platforma RELATE adăugând o caracteristică de catalog de

resurse. Inițial, lista ELE a fost importată în platforma RELATE. Apoi, resurse suplimentare dezvoltate în cadrul institutului nostru sau utilizate în diferite proiecte au fost importate în platforma RELATE.

5.6. Date de tip „linguistic linked data” în platforma RELATE

În Secțiunea 2.7 au fost descrise o serie de resurse în limba română care sunt disponibile în format de date legate (din engleză „linked data”). În plus, platforma RELATE a fost extinsă pentru a include suport pentru unele caracteristici aferente datelor legate de tip lingvistic („linguistic linked data”). Aceasta include: găzduirea resurselor de date conectate în depozit, generarea directă a datelor RDF, interacțiunea directă cu seturile de date RDF folosind un server integrat. Acesta poate fi descărcat sau punctul final poate fi folosit pentru interogări automate.

5.7. Dezvoltarea platformei

Dezvoltarea platformei RELATE se face open source, conform principiilor „științei deschise”. Atât codul, cât și resursele sunt publicate online. Dezvoltarea codului are loc în principal într-un depozit GitHub dedicat[11]. Este urmată o abordare modulară, bazată pe componente, fiecare componentă fiind într-un director dedicat. Fiecare componentă are un fișier descriptor bazat pe json care configurează modul în care este expusă în interfața GUI web. Descriptorul permite specificarea componentelor grafice (cum ar fi intrările de meniu, eticheta, pictograma, submeniurile, ordinea meniului) și scripturile pentru gestionarea cererilor web specifice, fie pentru pagini web complete, fie pentru transferul de date API.

6. Perspective de cercetare viitoare

6.1. Procesarea limbajului natural

Prin introducerea modelelor de limbaj generativ de ultimă generație (cum ar fi ChatGPT, GPT-4o, Llama-3), multe sarcini de procesare a limbajului natural (cum ar fi rezumarea, răspunsul la întrebări, extragerea informațiilor) se apropie de nivelul uman. Cu toate acestea, modelele mari de limbă (LLM, din engleză Large language Models) generative sunt uneori predispuse la halucinații sau idei preconcepute (din engleză „biases”), care decurg adesea din datele utilizate pentru antrenament. În acest context, sunt încă necesare cercetări în domeniul ancorării rezultatelor în realitate și al reducerii ideilor preconcepute. Acest lucru este necesar atât la nivel internațional, multilingv, cât și la nivel monolingv, având în vedere că limba română este mai puțin reprezentată în LLM-uri comparativ cu engleza. Formarea unui LLM se bazează pe existența unor volume mari de text pentru o anumită limbă. Acest lucru este o provocare pentru scenariile cu resurse reduse, cum ar fi limbile cu un număr

[11] https://github.com/racai-ai/relate

redus de vorbitori sau dialectele utilizate în anumite zone. Avansarea cercetării în aceste domenii necesită noi tipuri de modele lingvistice sau combinarea LLM-urilor cu alte mecanisme, cum ar fi regăsirea informațiilor ghidată de rețele neuronale. Textele combinate (folosind cuvinte scrise de obicei în două limbi) reprezintă provocări suplimentare pentru aplicarea LLM-urilor. De exemplu, varietatea limbii române vorbite în Republica Moldova este adesea amestecată cu cuvinte rusești atunci când sunt rostite. Textul recunoscut de un sistem ASR în acest caz va fi uneori dificil de procesat direct pentru un LLM. În afară de textul scris corect (cum ar fi cărțile), procesarea limbajului natural se preocupă de alte domenii de utilizare a textului, cum ar fi rețelele sociale. În acest caz, oamenii folosesc diferite moduri de a-și exprima sentimentele, cum ar fi emoticoane, emoji, text scris cu majuscule și hashtag-uri. Având în vedere legile și reglementările actuale privind confidențialitatea personală, cum ar fi GDPR, anonimizarea este un domeniu cheie de cercetare pentru multe organizații care se ocupă cu volume mari de text.

Oarecum legată de anonimizare, dar și de protejarea proprietății intelectuale, este problema protejării conținutului LLM-urilor. Atunci când mai multe entități (de exemplu, mai multe spitale, organizații de stat sau întreprinderi private) sunt implicate în formarea unui LLM, problema deținerii datelor și a accesului la acestea este primordială. Instruirea distribuită a LLM permite ca datele să rămână la furnizorul de date și doar ponderile sau informații numerice similare să fie schimbate.

6.2. Trustworthy AI

Sistemele AI pot provoca perturbări și pot crea provocări etice, legale și societale. AI poate sprijini chiar și crearea de arme cibernetice utilizabile în războiul cibernetic. Mai mult, AI poate facilita sau permite noi acte criminale („AI-Crime”), cum ar fi frauda prin utilizarea platformelor de social media sau manipularea pieței. Pentru o adoptare cu succes a AI, atât persoanele fizice, cât și întreprinderile trebuie să aibă încredere în sistemele AI. Guvernele și factorii de decizie din întreaga lume încep să investigheze măsuri de politică care ar putea aborda riscurile potențiale asociate cu AI. La nivel european, Grupul de experți la nivel înalt pentru inteligența artificială (AI HLEG - High Level Expert Group on Artificial Intelligence) și-a publicat Orientările de etică pentru IA de încredere. Noțiunea de IA de încredere (TAI - Trustworthy AI) este descrisă ca având trei componente, care ar trebui luate în considerare pe parcursul întregului ciclu de viață al sistemului (dezvoltare, implementare și utilizare): (1) legal - sistemul trebuie să respecte toate legile și reglementările aplicabile; (2) etic - sistemul trebuie să adere la principii și valori etice; (3) robust - sistemul nu trebuie să provoace vătămări neintenționate. Sistemele AI ar trebui să urmeze o abordare centrată pe om, în care drepturile individuale sunt întotdeauna respectate. AI ar trebui doar să completeze și să sporească abilitățile umane, în timp ce utilizatorii ar trebui să rămână în control deplin asupra acțiunilor lor. Prin urmare, sistemele AI nu ar trebui să înșele sau să constrângă (eventual prin manipulare) utilizatorii umani.

6.3. Procesarea vorbirii

Interfețele om-mașină necesită procesarea vorbirii fie pentru recunoașterea intrărilor utilizatorului (recunoaștere automată a vorbirii - ASR), fie pentru producerea de răspunsuri folosind vorbirea umană (sinteză text-to-speech - TTS). Arhitecturile SOTA actuale pentru ASR produc rezultate bune cu condiția ca suficiente date de vorbire să fie disponibile pentru antrenament. Cu toate acestea, acest lucru este o provocare pentru limbi sau scenarii cu resurse reduse. Având în vedere limba română, există provocări în ceea ce privește tipurile de vorbire slab reprezentate (cum ar fi copiii, persoanele în vârstă), zgomotul prezent în audio (cum ar fi înregistrările în medii zgomotoase), mai multe persoane care vorbesc în același timp, amestecarea de cuvinte în diferite limbi (de exemplu în Republica Moldova oamenii îmbină adesea limba română cu cuvinte sau expresii în limba rusă). Experimentele inițiale efectuate în cadrul proiectului „Underrepresented speech dataset from open data: case study on the Romanian language (USPDATRO)", sub supravegherea mea în calitate de director de proiect, au indicat că vorbirea slab reprezentată în limba română este încă o provocare pentru sistemele ASR. Am descoperit că, într-o oarecare măsură, platformele existente de rețele sociale care găzduiesc videoclipuri (cum ar fi YouTube, Vimeo) pot fi exploatate pentru a aduna discursuri slab reprezentate lansate sub licențe deschise (de exemplu, Creative Commons). Cu toate acestea, acest lucru necesită transcrierea manuală și segmentarea videoclipurilor. Cercetările viitoare trebuie să se concentreze pe identificarea de noi surse pentru vorbirea slab reprezentată, de preferință care necesită mai puțină muncă manuală, dar și pe arhitecturi ASR îmbunătățite care sunt mai robuste la variațiile de vorbire.

6.4. Procesare multimodală

În lumea de astăzi dominată de interacțiunile cu rețelele sociale și platformele de partajare a conținutului multimedia, conținutul este adesea multimodal. Acesta include: video (combinând imagini în mișcare cu audio), video cu subtitrări (imagini + audio + text, uneori în mai multe limbi), imagini cu subtitrări, memeuri (imagini cu mesaje suprapuse), videoclipuri scurte care prezintă o imagine cu descriere audio. Gestionarea corectă a unui astfel de context necesită combinarea lanțurilor individuale de procesare (pentru imagine, audio și text) într-o singur mecanism sincronizat. În mod alternativ, sunt necesare modele neuronale end-to-end capabile să gestioneze mai multe modalități simultan.

6.5. Combatere deep fakes

Conținutul de înaltă calitate generat de AI devine foarte ușor de obținut, datorită proliferării modelelor generative, și poate fi văzut pe diverse site-uri și platforme online. Inițial, termenul „deep fake" se referea la imagini sau videoclipuri generate de înaltă calitate,

care erau extrem de greu de distins de cele reale. Prin extensie, același termen poate fi aplicat conținutului text de înaltă calitate generat. În unele cazuri, modelele AI sunt folosite pentru a genera conținut multimodal, combinând text și imagini sau videoclipuri. Mai mult, textul poate fi sintetizat folosind sisteme TTS pentru a produce o voce asemănătoare cu cea a unui vorbitor uman. Chiar mai mult, tehnologiile AI pot fi folosite pentru a înlocui fața unei persoane reale cu cea a altcuiva sau vocea unei persoane reale cu cea a altei persoane reale.

În scenariile de război hibride de astăzi, generarea de știri false și propagarea rapidă prin diverse canale online joacă un rol cheie. Situația este agravată de utilizarea modelelor lingvistice mari (LLM) generative de următoarea generație, care permit crearea unor volume mari de articole de știri false și mesaje sau postări pe rețelele sociale într-un interval de timp foarte scurt. În plus, generatoarele text-to-image și alte modele de manipulare a imaginii bazate pe inteligență artificială produc fotografii false profunde de înaltă calitate, adesea imposibil de distins de cele reale. Acest lucru are potențialul de a copleși serviciile de verificare a faptelor și chiar de a compromite modele care utilizează informații noi generate de voluntari despre evoluție subiectelor de știri, atunci când suficiente dintre aceste rapoarte sunt generate de roboți. Atât Strategia de război hibrid a NATO, cât și Cadrul comun al UE pentru abordarea amenințărilor hibride menționează necesitatea unei apărări colective în fața unei campanii hibride[12]. În urma agresiunii ruse din Ucraina, combaterea campaniilor de dezinformare a devenit esențială pentru România și Republica Moldova, precum și în regiunea extinsă a Mării Negre.

6.6. Combatere știri false

Știrile false reprezintă fenomenul de răspândire a informațiilor false, de obicei ca parte a campaniilor de dezinformare. Acest lucru este de obicei relevant pentru știrile politice și economice, dar poate viza și alte domenii. Trebuie făcută o distincție între știrile false și falsurile profunde („deep fakes”). Definiția știrilor false descrie informațiile false transmise ca fiind reale. Falsurile profunde descriu conținutul generat de AI. De obicei, conținutul fals profund este, de asemenea format din știri false, dar nu este întotdeauna cazul (conținutul generat de inteligență artificială poate conține și informații adevărate, mai ales luând în considerare LLM-urile). Imaginile false profunde sunt de obicei mai predispuse să fie neadevărate în comparație cu conținutul text sau audio. Un prompt care cere unui LLM să furnizeze un raport bazat pe un anumit eveniment poate avea ca rezultat conținut real, în funcție de cunoștințele disponibile în timpul antrenamentului de model și de detaliile furnizate în prompt. Cu toate acestea, o solicitare similară prin care se cere unui generator de imagini să producă o imagine va avea mai probabil ca rezultat o imagine falsă (o imagine care nu corespunde unui anumit loc sau eveniment adevărat).

[12] https://nmiotc.nato.int/wp-content/uploads/2020/02/Building-a-Comprehensive-Approach-to-Countering-Hybrid-Threats-in-the-Black-Sea-and-Mediterranean-Regions-by-Chris-Kremidas-Courtney.pdf

6.7. Platforme pentru resurse și tehnologii de limbă

Platforma RELATE (descrisă anterior în Capitolul 5) este un prim pas către o platformă de tehnologii lingvistice pentru limba română. Aceasta integrează resurse, instrumente și modele dezvoltate fie la ICIA, fie de către parteneri în diferite proiecte. Cu toate acestea, trebuie extinsă (sau înlocuită cu o platformă îmbunătățită) pentru a lua în considerare noile tehnici care apar în sfera tehnologiilor lingvistice. Acestea includ: integrarea cu alte platforme europene sau internaționale, LLM-uri care oferă doar funcționalități de procesare bazate pe cloud, instruire și/sau execuție descentralizată, procesare îmbunătățită a seturilor mari de date.

6.8. Îmbunătățirea arhitecturilor de rețele neuronale artificiale

În Secțiunea 3.4, am descris abordarea mea pentru implementarea unui mecanism de inhibare laterală în rețelele neuronale artificiale. Acest lucru a oferit deja rezultate îmbunătățite în mai multe sarcini, în special atunci când se iau în considerare scenarii cu resurse reduse. Totuși, această activitate va continua, mai întâi prin explorarea sarcinilor suplimentare care ar putea beneficia de pe urma abordării bazate pe inhibarea laterală și apoi prin explorarea altor mecanisme. În afară de abordările inspirate din punct de vedere biologic, consider că ar trebui explorate și alte variații ale implementărilor rețelelor neuronale. De exemplu, algoritmii de auto-organizare ar putea permite rețelei să schimbe modul în care sunt conectate straturile sau modul în care neuronii sunt plasați în interiorul unui strat. Combinarea mai multor modele pre-antrenate pentru a forma o nouă rețea neuronală este o altă abordare care a arătat rezultate promițătoare. De exemplu, modelele recente pentru răspunsuri la întrebări pe bază de vedere (VQA - „Visual Question Answering") au combinat un model de limbaj pre-antrenat, un model de vedere pre-antrenat și o rețea intermediară nou antrenată care conectează celelalte două modele.

7. Concluzii

Această teză de abilitare a acoperit cercetările mele recente (capitolele 2-5), ulterior susținerii tezei mele de doctorat. Accentul meu principal a fost pe resursele, tehnologiile, modelele și platformele specifice prelucrării limbajului natural. Am început de asemenea să lucrez la resurse și instrumente multimodale și la noi arhitecturi de rețele neuronale artificiale. În plus, am avut un interes pentru predare, fiind implicat în activități de mentorat și tutorat pentru liceeni și studenți la nivel de licență și master. Am reușit să-i implic în activitățile mele de cercetare, inclusiv în crearea de resurse și instrumente lingvistice. Acest lucru este demonstrat de mai multe lucrări care au co-autori studenți de liceu și facultate (de exemplu, vezi referințele [1], [2], [8], [9], [10], [11], [14]) . Am fost și membru în comisia consultativă pentru doctoranzi. După abilitare, intenționez să continui să îndrum tinerii studenți către o carieră de cercetare, și obținerea doctoratului.