RAPORTARE ŞTIINȚIFICĂ

FAZA DE EXECUȚIE NR. 3

CU TITLUL Dezvoltarea modulului de limba romana,testarea si evaluarea metodelor proiectate in etapele anterioare

Avizat,

Coordonator	**Agent Economic**
Denumire: Universitatea din Bucuresti	**Denumire:** Pluriva SRL
Reprezentant Legal: prof. dr. Mircea Dumitru	**Reprezentant Legal:**
Semnătură:	**Semnătură:**
Ștampilă:	**Ștampilă:**

Director Proiect: prof.dr. Liviu P. Dinu

Semnătură:

Responsabil de proiect:

Semnătură:

Raport stiintific
privind implementarea proiectului in perioada ianuarie – septembrie 2018

Proiect: *Sistem Inteligent de Generare Automată a Răspunsurilor (SIGAR).*

SIGAR: Project 53BG/2016, funded by Romanian National Authority for Scientific Research and Innovation, CNCS/CCCDI – UEFISCDI, PNCDI III (Programul 2 - Creşterea competitivităţii economiei româneşti prin cercetare, dezvoltare şi inovare. Transfer de cunoaştere la agentul economic "Bridge Grant")

Durata Proiectului: 1 octombrie 2016- 30 septembrie 2018

Director: Prof. univ. dr. Liviu P. Dinu
Contractor: Universitatea din Bucuresti
Etapa: 30 septembrie 2018 (unica)

Obiectivul principal

Principalul obiectiv al proiectului constă în implementarea unei soluții bazate pe învățare automată, care va parsa și analiza baza de date de tichete în vederea construirii unor răspunsuri automate la întrebări sau probleme frecvent întâlnite.

Obiectivul proiectului pentru anul 2018 este:

1. **Dezvoltarea modulului de limba romana,testarea si evaluarea metodelor proiectate in etapele anterioare.** Pentru îndeplinirea acestui obiectiv, au fost planificate următoarele activități:

Activitate 3.1***: Dezvoltarea de resurse specific proiectului pentru limba romana.***.
În scopul realizarii acestei activități următoarele acțiuni au fost planificate și realizate:

- analiza morfologica a datelor
- extragerea părților de vorbire pentru datele disponibile
- finalizarea adnotarilor pentru limba romana

Activitate 3.2: ***Activitati diseminare***. Principalele acțiuni desfășurate în cadrul acestei activități au fost participarea la conferinte, efectuarea unor stagii de cercetare, diseminarea rezultatelor in intalniri formale si informale, in organizarea unui seminar de cercetare, etc.

Activitate 3.3: ***Evaluare si testare module pentru limba romana***

În scopul realizarii acestei activităţi următoarele acţiuni au fost planificate şi realizate:

- analiza datelor fara diacritice
- detectarea entitatilor in romana
- evaluare, testare, ajustare

Activitate 3.4: ***Finalizarea modulelor,evaluarea, ajustarea rezultatelor.*** Principalele acţiuni desfăşurate în cadrul acestei activităţi au fost finalizarea resurselor pentru generarea raspunsurilor, testarea, evaluarea si ajustarea metodelor dezvoltate in etapele anterioare, intocmire rapoarte finale.

În cadrul acestei etape, principalul punct de lucru a fost determinat de optimizarea şi testarea metodelor prezentate anterior pe limba română, în particular pe limba română utilizată în cadrul datelor Pluriva. O caracteristică aparte a acestor date constă în faptul că este adesea întâlnită o terminologie specifică, determinată de limbajul business şi tehnic folosit de către utilizatorii sistemului, care nu este întâlnită în mod general în cadrul resurselor lingvistice publice disponibile pentru limba română. Mai mult decât atât, uneltele deja existente pentru limba română neavând terminologia învăţată în cadrul modelelor automate şi nici acces la resurse similare, prezintă o eficienţă foarte scăzută atunci când sunt aplicate în cazul de faţă. Procesul de curăţare a datelor din Etapa 2, ne-a permis să construim câteva resurse specifice integrate şi optimizate pentru procesarea datelor Pluriva. Printre acestea numărăm o metodă de adnotare a părţilor de vorbire care include şi entităţi nume proprii, modele de word embeddings specifice limbii române utilizate în datele Pluriva şi o unealtă de extragere a conţinutului esenţial din schimburile de mesaje dintre agenţi şi clienţi.

Rezultatele activitatii in cadrul proiectului in perioada 1 ianuarie-30 septembrie 2018

Activitatea ştiinţifică:

În anul 2018 activitatea ştiinţifică a membrilor proiectului a constat în special pe analiza tipului de discurs utilizat în datele Pluriva, analiza morfologică şi implementarea unui detector de părţi de vorbire customizat pentru aceste date, capabil de asemenea să indice şi prezenţa unor entităţi de tip nume propriu. În paralel au fost proiectaţi algoritmi care să identifice stări sau sentimente din cadrul textului, algoritmi care au un potenţial practic în situaţia în care un client nemulţumit îşi exprimă această părere în interacţiunile cu agenţii firmei.

În ultima parte a proiectului am dorit să ne concentrăm pe cercetarea abordărilor care s-ar dovedi fezabile pentru procesarea textelor în limba română şi mai exact pentru procesarea tipurilor de texte pe care agentul economic le stochează. Conform analizelor efectuate în cadrul activităţilor din Etapa 2, am constatat că aceste texte nu doar că prezintă un limbaj de specialitate, dar că acoperă şi un lexic care nu este standard şi care nu poate fi analizat utilizând resursele disponibile pentru limba română (DEX online, Romanian Treebank, POS taggers sau Named Entity Recognizers) din cauza lipsei adnotărilor pre-existente pentru acest tip de date.

Un part of speech (POS) tagger sau detector de părţi de vorbire reprezintă o unealtă esenţială pentru orice tip de analiză text. În cazul nostru am analizat uneltele şi resursele deja existente pentru limba română şi am concluzionat că acestea nu doar că prezintă o performanţă relativ scăzută pentru limba română, dar că pot fi aproape inutile pentru datele agentului economic care conţin un lexic specializat ce nu face parte decât pe jumătate din vocabularul limbii române aşa cum este el normat în DEX. Aşadar, am fost nevoiţi să găsim soluţii alternative pentru adnotarea automată a acronimelor, cuvintelor împrumutate din engleză sau a celor ne-normate. Am ales să utilizăm corpusul *Romanian UD treebank (RoRefTrees)* (Barbu Mititelu et al., 2016) adnotat deja cu părţi de vorbire şi arbori sintactici. Corpusul are o distribuţie neuniformă de genuri: literatură - 1818 propoziţii, legislativ - 1606 propoziţii, medical - 1210 propoziţii, traduceri din FrameNet - 1092 propoziţii, texte academice - 950 propoziţii, ştiri - 933 propoziţii, ştiinţific - 362 propoziţii, wikipedia - 251 propoziţii, altele - 1301 propoziţii. În total conţine 218 mii de cuvinte adnotate cu părţi de vorbire. Utilizăm ultima versiune a acestui corpus pentru a extrage seturi de reguli de adnotare a părţilor de vorbire. Ca mecanism de învăţare utilizăm Single Classification Ripple Down Rules (RDR) care s-au dovedit a fi eficiente *(Dat Quoc Nguyen et al., 2016)* pentru acest tip de problema, prin care înăţăm anumite structuri condiţionale de tip arbore de decizie care conduc la determinarea unei părţi de vorbire în funcţie de contextul în care se află un cuvânt. Avantajul acestei metode constă în simplitate şi capacitatea de a fi descriptivă şi a modifica regulile învăţate din corpus prin supervizare umană.

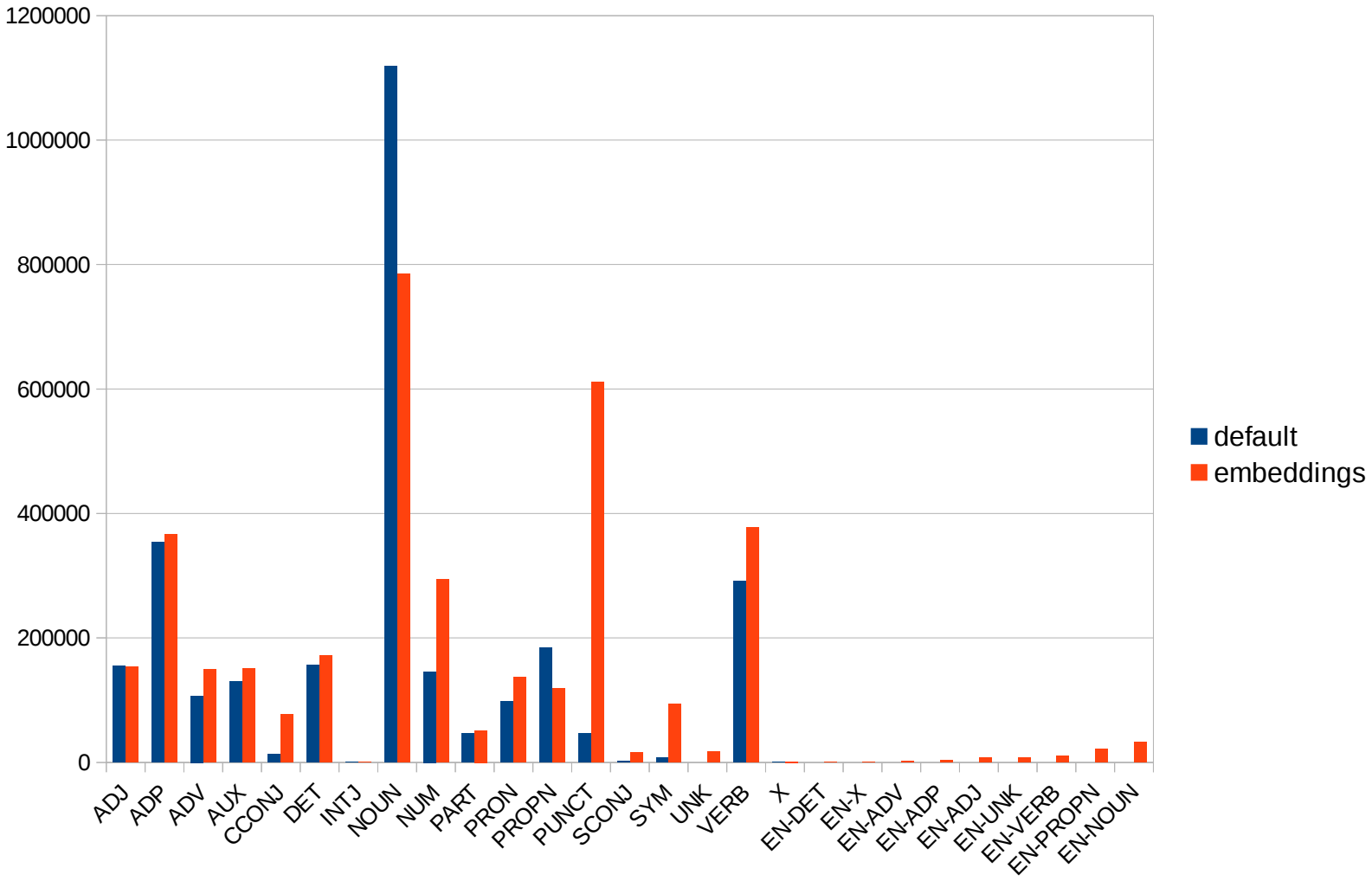
Câteva exemple de reguli învăţate:
- cuvântAnterior1 == "va" : conclusion = "VERB"
- cuvânt == "până" şi următorulCuvânt == "ce" : conclusion = "ADP"
- suffixUltimele3Litere == "şpe" : conclusion = "NUM"
- cuvântAnterior1 == "să" : conclusion = "VERB"
- cuvânt == "este" and următorulPOS == "CCONJ" : conclusion = "AUX"
- cuvântAnterior2 == "liceul" : conclusion = "PROPN"

În plus, regulile învăţate din corpus au fost verificate augmentate manual iar cuvintele care lipsesc în corpusul de antrenare au fost analizate separat astfel:

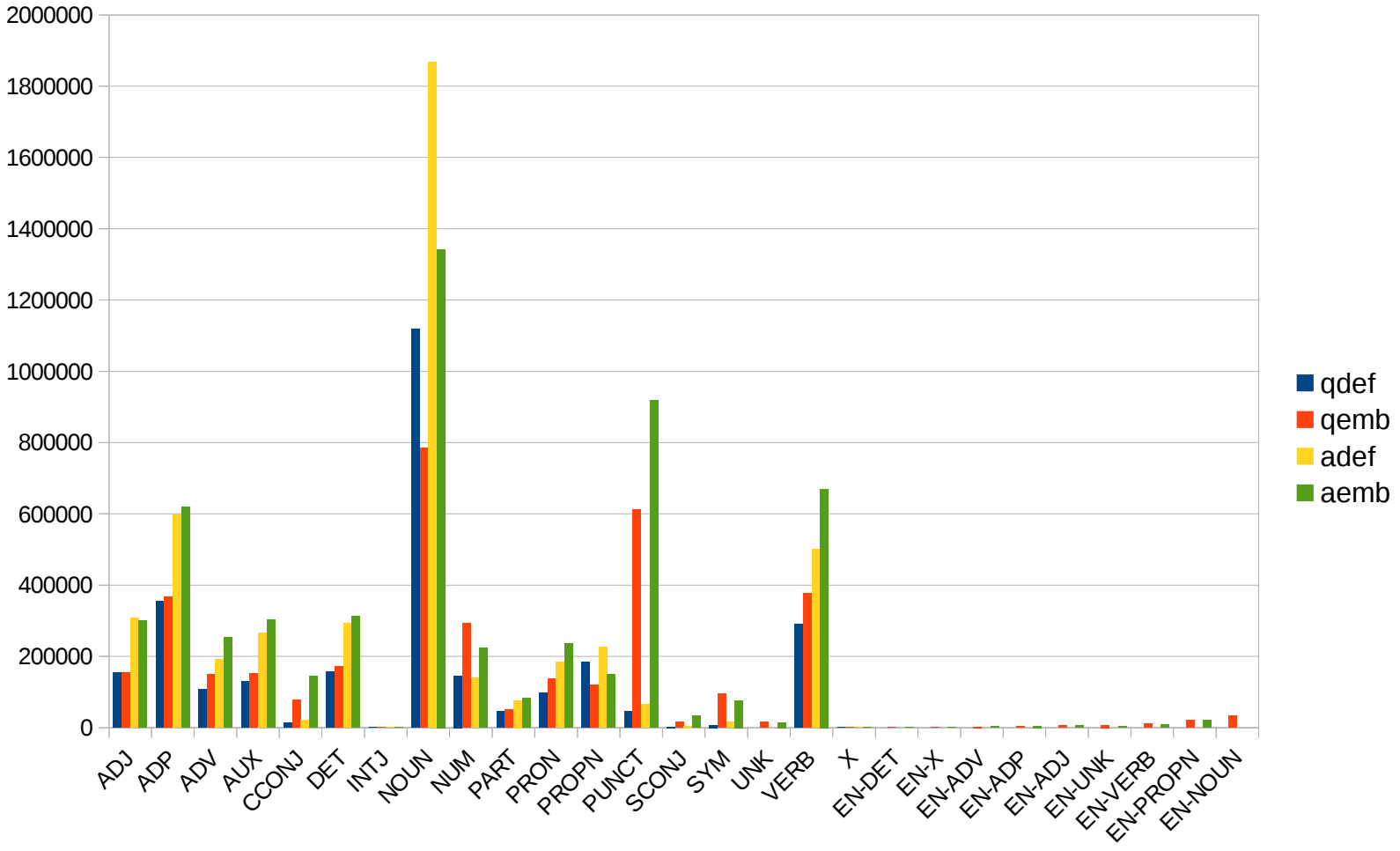
Pas 1. determină dacă un cuvânt lipseşte din dicţionarul extras din datele de antrenare sau pentru care nu se poate deduce partea de vorbire folosind una din regulile anterioare (de exemplu acronime: fac (factură), Ms (mulţumesc), task, sample, etc.)
Pas 2. verifică dacă există word embeddings antrenate care includ acel cuvânt
Pas 3. dacă nu există, partea de vorbire a cuvântului este UNK (necunoscut)
Pas 4. altfel, caută cele mai similare 10 cuvinte în funcţie de distanţa euclidiană dintre reprezentările vectoriale ale acestora şi a cuvântului dat

Pas 5. atribuie cuvântului partea de vorbire a celui mai apropiat vecin în raport cu distanţa euclidiană, care fie se găseşte în dicţionar sau pentru care se poate aplica una din regulile anterioare

Şi nu în ultimul rând, am dorit să identificăm dacă din cuvintele adnotate, avem de fapt cuvinte provenite din engleză pentru care nu am putea identifica partea de vorbire cu foarte multă acurateţe sau care ar putea rezulta ca o excepţie de la regulile de adnotare. Aşadar, am ales să adnotăm părţile de vorbire cu un prefix ”EN-”, dacă acel cuvânt face parte din synsets existente în WordNet. În plus, am implementat funcţii care să trateze automat lipsa diacriticelor din corpusul nostru, astfel că sunt executate căutări ale cuvintelor cu şi fără diacritice. În mod similar, am mai introdus câteva reguli de adnotare care să identifice cuvinte cu formă de titlu sau de nume propriu. În graficul de mai jos sunt reprezentate distribuţiile părţilor de vorbire adnotate pe corpusul de întrebări. Modelul de vectori de cuvinte folosit este de asemenea antrenat pe acelaşi corpus, care a fost pre-procesat şi curăţat înainte. Modelul de bază (în albastru, default) are tendinţa de a identifica mai multe substantive (NOUN), dar o analiză manuală mai atentă asupra adnotărilor generate a arătat că aceasta este tendinţa algoritmului de a adnota substantive când partea de vorbire este necunoscută, generând multe greşeli pentru cuvinte care nu apar în dicţionarul învăţat din corpus. Pe corpusul nostru, acurateţea predicţiilor este de 65% cu sistemul de bază, pe când acurateţea pe datele de antrenare este de 95.75%. Folosind structurile vectoriale de reprezentări ale cuvintelor, am putut introduce în adnotator informaţii suplimentare pe baza cărora deciziile sunt mult mai bune. După cum se observă şi în diagrama de mai jos, adnotatorul cu embeddings identifică un număr semnificativ mai mare de cuvinte din fiecare categorie, mai puţin categoria substantivelor. Are loc o redistribuire a părţilor de vorbire ceea ce face ca sistemul să fie mai robust şi cu un nivel de acurateţe mai mare (85.3%) conform unor evaluări manuale ale rezulatelor adnotate.



Distribuţiile comparative ale părţilor de vorbire pentru corpusul de întrebări (qdef şi qemb) şi pentru cel de răspunsuri (adef şi aemb) se regăsesc în următoarea diagramă:



Numărul de substantive greşit adnotate este cu mult mai mare în corpusul de răspunsuri (galben şi verde), comparativ cu cel de întrebări, ceea ce face şi redistribuirea lor mai amplă în cadrul celorlalte părţi de vorbire. Îm ambele corpora observăm şi o redistribuire a numelor proprii, numărul total al acestora scăzând după utilizareaa reprezentărilor vectoriale.
Un alt aspect demn de observat este faptul ca un numar de aproximativ 86000 de părţi de vorbire din corpusul de întrebări sunt parte din limba engleză, iar un număr comparabil de 82000 de părţi de vorbire din corpusul de răspunsuri au ca rezultat prefixarea cu EN-. Ceea ce ne indică faptul că acele cuvinte sunt greu de adnotat fără eroare şi că deşi ele sunt adnotate cu o parte de vorbire (DET, ADV, VERB etc.), cel mai probabil este rezultatul algoritmului de antrenare. În acest fel, putem filtra cuvintele străine sau le putem pune în containere separate pentru procesare.

Metodele de detectare a entităţilor necesită un corpus vast adnotat manual cu anumite tipuri de entităţi şi cu părţile de vorbire ale cuvintelor care sunt vecine cu acele entităţi. În cazul nostru, neavând la dispoziţie o resursă de asemenea magnitudine, am ales să ne folosim de tag-ul PROPN pentru a identifica numele proprii. În această categorie includem nume de firmă, nume de persoane, nume de oraş etc. Toate aceste nume se identifică cu uşurinţă în corpusul agentului economic deoarece aceste cuvinte sunt în mod majoritar scrise cu literă mare la început. În mod paradoxal, deşi schimburile de replici dintre agenţi şi clienţi sunt caracterizate de un număr semnificant de cuvinte nenormate şi lipsa diacriticelor, în cazul numelor proprii este păstrată

norma de scriere, mai exact 25.7% din cuvintele înregistrate în modelel antrenate sunt scrise cu literă mare, iar aproximativ 8% sunt cuvinte aflate la începutul unei propoziții, rămânând un procent de 17.7% de entități nume proprii pentru care putem deduce tag-ul doar pe baza ortografiei. Acestea au fost adăugate în dicționarul de adnotare folosit de către POS tagger.

Pentru finalizarea resurselor, am implementat o metodă de curățare și extragere a conținutului esențial din date. Această metodă poate fi folosită pentru pre-procesare și trimitere a conținutului esențial către agregator sau poate fi folosită pentru extragerea unui sumar care să fie livrat agentului care citește conținutul ticketului. De asemenea poate fi folosită pentru îmbunătățirea categorisirilor de date pe subiecte și pentru indexare și căutarea mai rapidă a elementelor duplicate.
Pentru a antrena un model de extragere de conținut esențial, am construit un corpus paralel din perechi de documente aliniate linie cu linie. Dintr-un document care conține o secvență de întrebări – răspunsuri, am eliminat liniile care nu conțin informații esențiale legate de subiectul abordat. Aceste informații pot cuprinde propoziții întregi, elemente de structura a mailurilor (header, formulări), tabele html, elemente de conversație. Corpusul inițial cuprinde 2000 de documente selectate aleatoriu cumulând în total peste 200,000 de cuvinte, iar în urma înlăturării elementelor care nu fac parte din conținutul esențial, rămânem cu un corpus filtrat care conține 73,000 de cuvinte. Utilizăm cele două componente aliniate pentru a eticheta liniile într-un format binar – linii care trebuie păstrate sau care trebuie scoase. Având liniile etichetate putem antrena un clasificator care să distingă între liniile care sunt irelevante și cele cu conținut esențial. Pentru asta am extras trăsături la nivel lexical folosind tf-idf și am antrenat un model de regresie logistică pe acele perechi de trăsături – etichete. Un dezavantaj major al acestei metode constă în faptul că nu ia în considerare propozițiile vecine în momentul adnotării. Pentru a trata această problemă, am utilizat probabiliățile liniilor vecine de a fi parte din conținut astfel:
- dacă clasificatorul afișează 1, atunci linia este cu conținut esențial
- dacă clasificatorul afișează -1 și probabilitatea cumulată a propoziției curente cu cea anterioare este mai mare de 0.2, atunci linia este cu conținut esențial
- dacă propoziția anterioară este -1 și probabilitatea cumulată a propoziției curente cu cea anterioare este mai mare 0.2, atunci linia anterioară este cu conținut esențial
- altfel, linia curentă nu este cu conținut esențial
Pragul de 0.2 a fost stabilit în urma mai multor procese de cross-validare, iar cumularea probabilităților liniilor de a fi sau nu parte din conținut esențial îmbunătățește considerabil calitatea rezultatelor.
În urma implementării, am decis să extragem conținutul esențial din întregul corpus și evaluarea rezultatelor după un criteriu manual și subiectiv. Astfel că, am inspectat un sample de 1000 de texte din care am extras conținut relevant pentru a observa ce tipuri de greșeli face algoritmul, mai exact am observat că modelul era antrenat într-un mod mult prea restrictiv, eliminând propoziții cu caracter esențial în situații în care acest lucru nu era necesar. Motivele pentru acest comportament rezultă din felul în care corpusul a fost adnotat și din faptul că trăsăturile cu tf-idf extrase nu au capacitatea de a generaliza pentru date cu varietate mare lexicală. Altfel spus, datele de antrenare și datele pentru care am executat filtrare, au lexic diferit fapt care duce la overfitting. Pentru a compensa asta, am antrenat structuri de reprezentare vectorială a propozițiilor folosind centroidul de word embeddings. Antrenarea clasificatorului sau regresorului s-a executat pe centroizii de reprezentări vectoriale iar rezultatele s-au dovedit

promiţătoare (F1 0.714, 0.73 acurateţe). Analizând manual adnotările făcute de clasificator, am putut observa că acestea erau mult prea permisive, adnotând drept conţinut fraze care ar fi făcut în mod normal parte din clasa celor eliminate. Aşadar am ales să agregăm modelul care execută predicţiile pe baza reprezentărilor tf-idf cu cel care utilizează reprezentările vectoriale într-un model combinat. Tabelul următor conţine un rezumat al rezultatelor obţinute în identificarea şi extragerea conţinutului esenţial.

Method	F_1	**Accuracy**
tf-idf classifier	0.746	0.890
emb classifier	0.714	0.873
tf-idf context proba	0.775	0.897
emb context proba	0.738	0.878
combined	0.774	0.893

Drept concluzii, putem vedea că utilizarea probabilităţilor contextuale ale liniilor îmbunătăţeşte acurateţea şi metrica F1 de evaluare, dar că această metrică nu este neapărat suficientă în evaluarea exactă a frazelor care pot fi eliminate sau păstrate. În urma procesului de sampling şi inspectare manuală a datelor adnotate, am putut observa diferenţe semnificative în momentul folosirii celor două tipuri de clasificatoare într-o formă combinată, astfel că forma combinată oferă un echilibru mai bun între clasificatorul prea restrictiv şi cel prea permisiv. Considerăm că acest rezultat este deosebit de important pentru agentul economic, permiţându-i acestuia să extragă conţinutul esenţial care ulterior poate fi folosit pentru adnotare, căutare sau procesare.

Activitati de diseminare

În aceasta etapa, membrii proiectului au reuşit să publice 9 lucrări (3 lucrari la conferinte categoria A, 2 la conferinte categoria B, 1 la conferinte tip C, 2 lucrari la workshop-uri asociate unor conferinte tip A si o lucrare la granita dintre ştiinţele umaniste şi cele reale) la unele din cele mai importante conferinţe dedicate procesării limbajului natural şi lingvisticii computaţionale (COLING, EMNLP, CICLING, LREC, workshopuri asociate EMNLP, etc), si o serie de alte articole au fost trimise spre publicare.

Articolul *Lexical Analysis and Content Extraction from Customer-Agent Interactions,* autori: Sergiu Nisioi, Anca Bucur and Liviu P. Dinu, în curs de publicare (W-NUT la EMNLP 2018) în care sunt prezentate problemele cu care ne confruntăm în cadrul analizei de text pe date cu continut specific, cum sunt cele de la Pluriva şi o metodă de eliminare a redundanţelor din texte cu scopul de a păstra elementele esenţiale din schimbul de replici dintre agenţi şi clienţi.

Articolul *Exploring Optimism and Pessimism in Twitter Using Deep Learning*, in curs de publicare la EMNLP 2018 este printre primele articole din literatura care abordeaza intr-o maniera computationala detectarea starii de optimism/pesimism din texte. Daca pentru analiza opiniilor si detectarea sentimentelor sunt peste 7000 de articole publicate, acest rezultat se numara printre primele articole (a fost considerat chiar primul articol din domeniu de catre unul dintre referentii anonimi) si consideram ca va fi extrem de util pentru o mare varietate de utilizatori. Un rezultat extrem de important a fost cel publicat in articolul *Ab Initio: Automatic*

Latin Proto-word Reconstruction (COLING 2018) in care reusim sa prezicem protocuvantul latin din care provin ccuvintele in limbile romanice moderne cu o acuratete superioara rezultatelor raportate anterior, devenind astfel state of art. Tehnicile computationale utilizate sunt folosite in premiera, si metoda folosita are in vedere inclusiv predictia cuvintelor din care au provenit cuvintele romanesti moderne.

Un articol la intersecţia dintre ştiinţele umaniste şi cele reale, care abordează aspecte filozofice, etice şi teoretice care ţin de impactul tehnologiei în contemporaneitate, este *Posthumanism, Technology, and Monstrous Life Forms*, autor: Anca Bucur, publicat la ISEA 2018. Aceste aspecte devin relevante din punct de vedere social şi economic, în momentul în care tehnologia capătă agentivitate şi un potenţial de înlocuire a forţei de muncă umană. Un alt articol publicat este *A Detailed Evaluation of Neural Sequence-to-Sequence Models for In-domain and Cross-domain Text Simplification,* autori: S Stajner, S Nisioi, LREC 2018, care expune problemele şi avantajele utilizării unei reţele neuronale de tipul celei implementate pentru generarea răspunsurilor, în cadrul unui task care are ca scop simplificarea conţinutului şi extragerea esenţialului din textele originale. În baza rezultatelor obţinute aici, am putut determina metoda optimă, conform opiniei noastre, de a esenţialul din datele avute la dispoziţie.

In articolul *Full Inflection Learning Using Deep Neural Networks* (CICLING 2018, in curs de publicare) au fost folosite retelele neuronale in scopul modelarii formelor flexionare ale substantivelor si verbelor din romana (pornind de la forma de dictionar), si am obtinut cele mai bune rezultate raportate pana acum.

De asemenea, membrii proiectului au participat la stagii de cercetare, au sustinut mai multe conferinte in tara si strainatate, au organizat a doua conferinta internationala RAAI 2018 (Recent Advances in Artificial Intelligence), a fost continuat în cadrul proiectului seminarul internaţional bilunar de lingvistică matematică şi computaţională „Solomon Marcus", pe parcursul căruia a continuat seria de prezentari începută in 2016. În cadrul prezentărilor participă cu regularitate atât membrii proiectului, cât şi masteranzi, doctoranzi sau postdoctoranzi din Departamentul de Informatică, din alte departamente ale Universităţii din Bucureşti, din cadrul Facultăţii de Automatică a Universităţii Politehnica din Bucureşti, membri ai partenerului Pluriva, dar şi alţi colegi din industrie interesaţi de tematica abordată. Considerăm că prin această activitate aducem un beneficiu proiectului prin activitatea de diseminare, si, în acelaşi timp realizam noi punţi de comunicare între mediul academic şi industrie.

A fost întreţinută pagina web a proiectului (http://nlp.unibuc.ro/projects/sigar.html), astfel încât toate rezultatele aşteptate la finalul acestei etape (pagina web, articol trimis spre publicare, raport) au fost realizate.

Activităţi administrative şi de organizare

Membrii echipei s-au întâlnit periodic pentru a prezenta şi discuta aspecte legate de organizare, de repartizarea temelor de cercetare, de colaborare între membrii echipei, de stabilirea conferinţelor şi jurnalelor în care se trimit spre publicare rezultatele cercetării şi de convenirea asupra unui calendar comun cu respectarea termenelor de atingere a obiectivelor proiectului şi de livrare a rezultatelor. Intre membrii proiectului si partenerul economic a fost o continua comunicare.

Au fost îndeplinite astfel toate activităţile administrative necesare bunei desfăşurări a proiectului.

Concluzii

- Obiectivele etapei au fost substanţial realizate.

Publicaţii şi conferinţe susţinute de către membrii proiectului în anul 2018:

1. Cornelia Caragea, Liviu P. Dinu, Bogdan Dumitru, **2018.** *Exploring Optimism and Pessimism in Twitter Using Deep Learning*. In Proc. **EMNLP 2018**, Brussels, Belgium, 2018 (to appear)
2. Sergiu Nisioi, Anca Bucur, Liviu P Dinu, **2018**. *Lexical Analysis and Content Extraction from Customer-Agent Interactions*. Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, pages 132–136 Brussels, Belgium, 2018.
3. Alina Ciobanu, Liviu P. Dinu, **2018**. *Ab Initio: Latin Proto-word Reconstruction*. In Proc.**COLING 2018 (main conference)**, p. 1604-1614, Santa Fe, USA, 2018 .
4. Alina Ciobanu, Liviu P. Dinu, **2018**. *Simulating Language Evolution: A Tool for Historical Linguistics*. In Proc.**COLING 2018 (demo section)**, pp68-72, Santa Fe, USA, 2018
5. Anca Bucur, **2018**. Posthumanism, Technology, and Monstrous Life Forms. In Proceedings of ISEA 2018 - 24th International Symposium on Electronic Art, p 313-316, Durban, South Africa, 2018.
6. Bogdan Dumitru, Alina Ciobanu, Liviu P Dinu, **2018**. *ALB at SemEval-2018 Task 10: A System for Capturing Discriminative Attributes*. In Proc **SemEval 2018 (SemEval@NAACL-HLT 2018)**, task 10, p. 963-967, New Orleans, USA, 2018
7. Maria Sulea, Bogdan Dumitru, Liviu P. Dinu. *Full Inflection Learning Using Deep Neural Networks***, CICLing 2018**, Hanoi, Vietnam, 2018
8. Liviu P Dinu, Ana Uban, **2018**. *Analyzing Stylistic Variation across Different Political Regimes*, **CICLing 2018**, Hanoi, Vietnam, 2018
9. Sanja Stajner, Sergiu Nisioi, **2018**. *A Detailed Evaluation of Neural Sequence-to-Sequence Models for In-domain and Cross-domain Text Simplification*. Proceedings of 11th edition of the Language Resources and Evaluation Conference (**LREC 2018**), p. 3026-3033, 7-12 May 2018, Miyazaki (Japan) .

Director Proiect,
Prof. dr. Liviu P. Dinu