RAPORT ȘTIINȚIFIC proiect complex ReTeRom, etapa III - noiembrie 2020

Proiectul 2: TEPROLIN

Activitatea 3.6: Analiza erorilor sistemelor ASR și TTS antrenate în proiectele 3 și 4 pe corpusul bimodal agregat în proiectul 1, adnotat și corectat

ASR /RAV

ICIA: Verginica Mititelu, Elena Irimia
UPB: Horia Cucu, Lucian Georgescu, Cristian Manolache

Completarea vocabularului sistemului de RAV – propuneri de corectare a corpusurilor de pe platforma CoBiLiRo

În faza actuală a proiectului, echipa UPB a îmbunătățit performanțele sistemului de RAV pe de o parte prin antrenarea pe noi seturi de date create și accesibile prin proiectul CoBiLiRo, iar pe de altă parte prin recrearea modelului de limbă prin adăugarea de texte noi. (Pentru date despre cantitatea de texte noi folosite, despre tipurile de informații existente în aceste corpusuri folosite la antrenare, despre noile performanțe ale sistemului de RAV, a se vedea raportul fazei întocmit de echipa UPB.)
Evaluarea la nivel de cuvânt arată, așa cum era de așteptat, o îmbunătățire a tuturor criteriilor de analiză: cuvinte recunoscute greșit, cuvinte inserate eronat, cuvinte netranscrise.
Evaluarea rezultatelor sistemului de RAV asupra corpusului din platforma CoBiLiRo a arătat că, în afară de cuvintele obișnuite, se produc o serie de erori în cazul a două tipuri de cuvinte: nume de entități și cuvinte cu cratimă. Toate acestea sunt cuvinte inexistente in lexiconul folosit de sistemul de RAV (engl. „out of vocabulary word"). Pentru fiecare categorie am adoptat o metodă de lucru considerată adecvată.

- Nume de entități: am extras din diverse surse liste separate de nume de persoane, nume de locuri, nume de organizații/firme/etc. cu scopul de a îmbogăți lexiconul sistemului de ASR. Listele create sunt disponibile la adresa https://relate.racai.ro/resources/ro_namelists_20201013.zip.
- Cuvinte cu cratimă: folosind un lexicon intern ICIA, am validat o parte dintre cuvintele cu cratimă, adică acelea corecte, dar absente din lexiconul ASR.
- Cuvinte obișnuite: s-a folosit aceeași procedură ca la cuvintele cu cratimă.

N.B. Nici una dintre aceste abordări nu rezolvă complet problema evaluării rezultatelor sistemului de RAV, atâta timp cât textele din corpusul CoBiLiRo conțin greșeli de scriere, iar rezultatele sistemului sunt comparate cu forma scrisă a corpusului. Pentru aceasta, am folosit o măsură de similaritate a tuturor tipurilor de cuvinte (Levenshtein distance) pentru a indica forma cea mai apropiată din listele de nume și lexiconul folosite. Aceste aproximări au fost validate manual și s-au creat liste de corecturi propuse pentru îmbunătățirea transcrierilor existente în corpusul din platforma CoBiLiRo.

O altă observație rezultată din lucrul cu datele este nevoia de uniformizarea textelor din corpus în ceea ce privește tipul de litere cu diacritice folosit, i.e. utilizarea literelor ș și ț conform standardului actual.

În ceea ce privește alte tipuri de erori pe care le-am identificat printre rezultatele sistemului de RAV, unele au caracter general și considerăm că țin de procedura de evaluare, în sensul că o

relaxare a restricțiilor impuse în evaluare ar arăta o îmbunătățire a performanțelor. Este vorba aici despre posibilitatea de a scrie unele pronume clitice cu sau fără cratimă, distincția dintre cele două ținând uneori doar de durata mai scurtă, respectiv mai lungă a unui sunet: ex.: *te-aștept* versus *te aștept.* Întrucât nu se poate stabili o limită de durată dincolo de care să se scrie fără cratimă, considerăm că recunoașterea oricăreia dintre variante trebuie acceptată ca fiind corectă. Similar, doar cu un sunet în plus, sunt alte clitice, care pot pierde o vocală la o rostire mai rapidă: ex.: *mă aștept* versus *m-aștept.*
În aceeași situație se află inițiala în vocala *î* a cuvintelor precedate de anumite cuvinte funcționale: ex: *la început* versus *la-nceput*, *a început* versus *a-nceput*, *să înceapă* versus *să-nceapă* etc.

Denumire activitate: Analiza erorilor sistemelor ASR și TTS antrenate în proiectele 3 și 4 pe corpusul bimodal agregat în proiectul 1, adnotat și corectat

Autori: ICIA: Irimia Elena, Verginica Mititelu UTCN: Adriana Stan, Beáta LŐRINCZ

REZUMATUL ETAPEI

Această activitate are rolul de a evalua îmbunătățirile pe care lexiconul construit de noi le aduce aplicațiilor de TTS dezvoltate la UTCN, prin comparare cu rezultatele obținute cu resurse folosite anterior, și apoi prin analiza a diferite scenarii de evaluare care folosesc informația din lexicon integral sau parțial (atât ca număr de intrări, selectate aleatoriu sau sistematic, cât și ca tipuri de informație utilizată).

Rezultatele activității: Raport asupra evaluării lexiconului îmbogățit cu transcriere fonetică în scenarii de predicție concurențială a informației lexicale, utile în aplicațiile de TTS dezvoltate la UTCN.

Introducere. În rapoartele anterioare aferente activităților 1.8 și 2.7 descriam în detaliu lexiconul cu informație extinsă, dezvoltat și validat în cadrul proiectului 2, TEPROLIN; pentru fiecare intrare din lexicon, informația asociată reprezintă: lema (forma de dicționar a cuvântului), eticheta morfo-sintactică în format MSD (Erjavec, 2004), împărțirea în silabe, marcarea accentului (printr-un apostrof) și transcrierea fonetică a formei ocurență. Am subliniat de asemenea importanța acestui lexicon în aplicațiile pentru recunoaștere automată și pentru sinteza vorbirii, susținând că este esențial ca o astfel de resursă sa fie de bună calitate. În raportul aferent activității 2.7 am descris în detaliu procesul complex de validare prin care lexiconul ReTeRom a trecut până în acest moment. Activitatea de validare a constat atât în corectare manuală intrare cu intrare, cât și în corectare automată, prin implementarea ca expresii regulate a diverse tipuri de reguli, bazate pe cunoaștere lingvistică.

În această etapă, împreună cu echipa parteneră UTCN, am putut testa și dovedi utilitatea lexiconului în cadrul unei aplicații de sinteză a vorbirii. Pentru a simplifca procesul de evaluare, am restrâns contextul evaluat la nivel textual, deoarece calitatea vorbirii sintetizate este greu de evaluat în mod obiectiv, cuantificabil. Astfel, rețeaua neuronală folosită pentru evaluarea lexiconului are ca scop predicția (în format text) concurentă a silabificării, accentului și transcrierii fonetice pornind doar de la forma (ortografică a) cuvântului, în lipsa unui context de utilizare.

În Tabelul 1 puteți vizualiza câteva perechi de date de intrare/ieșire ale rețelei: punctul marchează limitele silabelor iar accentul este marcat prin apostrof urmând vocala accentuată.

Intrare	**Ieșire**

abandonarăți	a . b a n . d o . n a' . r @ ts i 0
basculantei	b a s . k u . l a' n . t e j
Ciclopul	tS i . k l o' . p u l

Tabelul 1. Exemple de date de intrare și ieșire pentru rețeaua neuronală

În familia arhitecturilor neuronale utilizate pentru a învăța funcții secvență-la-secvență (en. Sequence-to-sequence), rețelele neuronale convoluționale (CNN, Gehring et al. 2017) și rețelele bazate pe atenție (Vaswani et al., 2017) sunt cele care au demonstrat cea mai bună acuratețe în probleme de procesare a limbajului natural. De aceea, într-o primă etapă a acestui studiu, s-a realizat o evaluare CNN vs transformer (vezi (Stan, 2020) pentru detalii despre implementarea acestora), în cadrul căreia rezultatele rețelei transformer au fost evident mai bune. Arhitectura rețelei are o structură de tip encoder-decoder și este prezentată în figura 1.

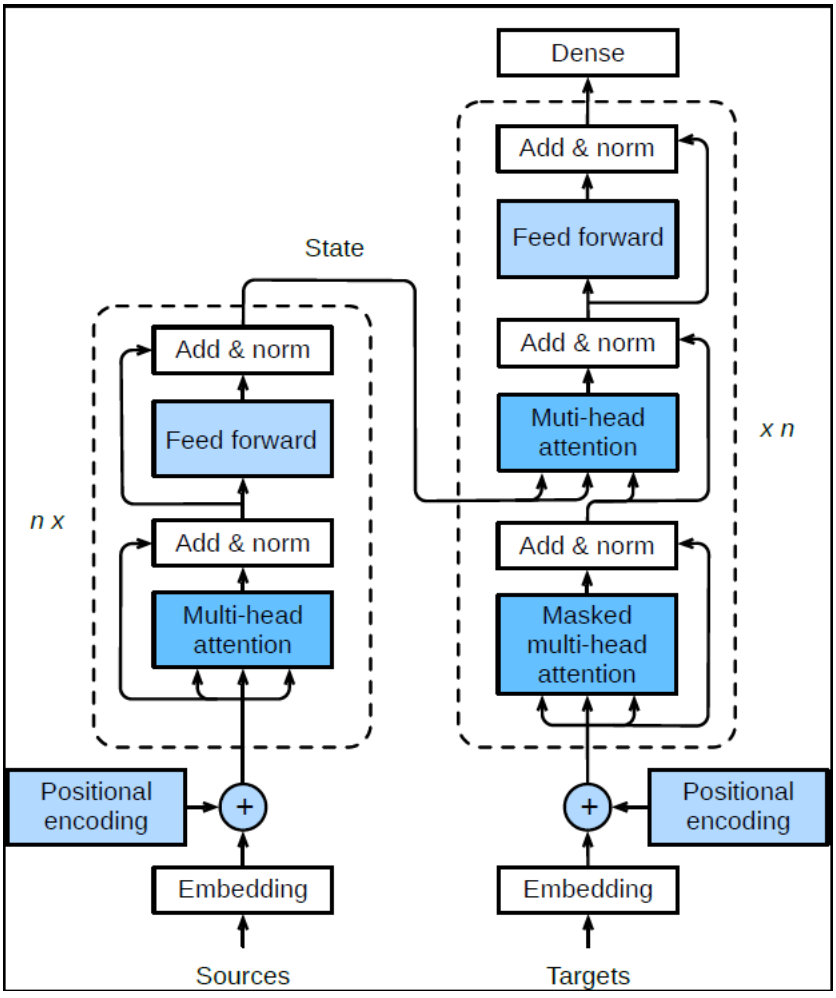

Figura 1. Arhitectura rețelei transformer folosite în experimentele noastre

Pentru selectarea parametrilor rețelei transformer ne-am bazat pe rezultatele din (Stan, 2020), rezultând o rețea cu 3 unități encoder, 4 unități decoder, 4 centre de atenție, dimensiunea stratului ascuns de 1.024 și dimensiunea vectorului embedding[1] de 128. Ponderile vectorului embedding sunt inițializate aleatoriu înainte de antrenare. Dimensiunea lotului (eng. „batch")

[1] reprezentare codificată numeric a șirului sursă/țintă

a fost stabilită la 512 iar pentru actualizarea ponderilor s-a folosit optimizatorul Adam cu o rată inițială de învățare de 0,0002. După 50 epoci, rata de învățare a fost redusă cu un factor de 0,2 iar procesul de antrenare a fost limitat de un criteriu de oprire bazat pe valoarea funcției de cost după 5 epoci.

În procesul de evaluare, setul de date a fost divizat în loturi de de 70%-10%-20%, dedicate proceselor de antrenare, validare și, respectiv, testare. Seturile de antrenare și validare variază în funcție de natura experimentelor, dar setul de testare este fix (aproximativ 80.000 de intrări).

În prealabil, ne-a interesat să investigăm utilitatea folosirii lexiconului ReTeRom în comparație cu folosirea concurentă a resurselor MaRePhor, RoSyllabiDict și a accentului din DEX (pe care și lexiconul nostru se bazează, extinzându-le), resurse utilizate până în acest moment pentru predicție concurentă de informație lexicală în aplicații de TTS la UTCN. Ulterior, am conceput scenarii de evaluare mai complexe, care să ne ajute să găsim modalități mai eficiente de folosire a resursei dezvoltate, după cum se va vedea în continuare.

Astfel, ne-a interesat îmbunătățirea performanței sistemului de predicție atunci cât setul de date crește gradat și identificarea momentului de platou al acestei îmbunătățiri. În acest scop, au fost selectate partiții aleatoare ale datelor pornind de la un minim de 5.000 de intrări.

Am urmărit de asemenea îmbunătățirea performanței în funcție de cantitatea de date atunci când selecția unui set redus de date se face asigurându-ne ca dispunem de întreaga acoperire a lexiconului la nivel de leme. Cum limba română este o limbă cu morfologie bogată, dar arareori o resursă de tip lexicon sau corpus cuprinde toate variantele morfologice ale unei leme, am vrut să vedem dacă performanța scade drastic atunci când doar lema, sau doar o formă (care nu este întotdeauna egală cu lema) sau două forme asociate fiecărei leme se regăsesc în datele de antrenare. Ne-am rezumat la a reduce numărul de forme doar pentru cuvintele conținut cu morfologie bogată: adjective, verbe si substantive, iar pentru restul tipurilor de cuvinte, numărul formelor asociate s-a păstrat intact. Astfel au fost derivate trei subset-uri, *1-FORM* (30.150 intrări, cu o singură formă pentru adjective, verbe și substantive) și *2-FORMS* (55.185 intrări, două forme pentru fiecare adjectiv, verb și substantiv) și *LEMMA* (35.890 intrări, cu o singură formă pentru adjective, verbe și substantive, care este întotdeauna identică cu forma dicționar, sau lema). În procesul de selectare a formelor din subset-urile *1-FORM* și *2-FORMS*, ne-am asigurat că reprezentăm în mod balansat fiecare trăsătură morfologică asociată acestor părți de vorbire (gen, număr, caz, mod, timp, persoană, etc.) pentru a încerca să păstrăm diversitatea terminațiilor morfologice existentă în limba română. Desigur, într-un scenariu real, nu se poate controla păstrarea integrală a acestei diversități (adică nu ne putem asigura că toate terminațiile sunt reprezentate într-un corpus sau într-un lexicon extras dintr-un corpus), dar o varietate ne-exhaustivă a formelor este naturală, pe când prezența exclusivă a formelor de dicționar este improbabilă.

În plus, ne-a interesat să investigăm potențialul utilizării părții de vorbire (primul simbol din eticheta morfosintactică) și a etichetei morfo-sintactice (MSD) ca trăsături suplimentare de intrare. Teoretic, această informație ar putea ajuta sistemul sa diferențieze între omografele care nu sunt omofone. În practică, în lexiconul nostru, omografele care au rostire diferită în funcție de POS-ul sau MSD-ul asociat reprezintă aproximativ 2000 de intrări și nu ne așteptăm ca rezolvarea acestei probleme de predicție să aibă un impact mare asupra ratei de erorare. Eventuala rezolvare reprezintă în schimb un plus de acuitate lingvistică adus sistemului. De asemenea, presupunem că rețeaua poate învăța să asocieze anumite terminații morfologice (și pronunțiile lor, care sunt specifice atunci când sunetele reprezintă terminații) cu anumite părți de vorbire/etichete morfosintactice, fapt ce ar putea compensa pentru lipsa tuturor formelor

asociate cuvintelor conținut. Vom verifica această presupunere în scenariile de evaluare bazate pe subset-urile LEMMA, 1-FORM și 2-FORMS.

În toate aceste scenarii de evaluare, output-ul rețelei transformer a fost, așa cum am prezentat în Tabelul 1, o combinație a celor trei sarcini de predicție lexicală, iar inputul a fost fie forma ortografică a cuvântului, fie forma împreună cu partea de vorbire, fie forma împreună cu eticheta morfosintactică asociată.

Drept metrici de evaluare, am folosit *word error rate* (WER, ro. "rata de eroare la nivel de cuvânt") și *character error rate* (CER, ro. „rata de eroare la nivel de caracter"), evaluate pe șiruri de caractere care includ marcajele de silabificare și accent. CER a fost măsurată folosind distanța Levenshtein (Levenshtein, 1966) între secvența de caractere prezisă și cea țintă. În plus, ne-am dorit să evaluăm erorile introduse de fiecare sarcină de predicție, motiv pentru care am calculat WER și CER eliminând din predicția concurentă informația furnizată de una sau mai multe dintre sarcini.

Evaluarea prealabilă, în care comparam resursa noastră cu cele folosite anterior, ne arată că, cu lexiconul ReTeRom, putem obține o rată a erorii **WER de 3,08** și o rată **CER de 1,08** pentru predicția concurențială, atunci când folosim toată informația disponibilă (inclusiv eticheta morfosintactică), ceea ce reprezintă o reducere importantă a ratei erorii de la 10,47 WER și 3,3 CER, obținute cu resursele utilizate anterior la UTCN.

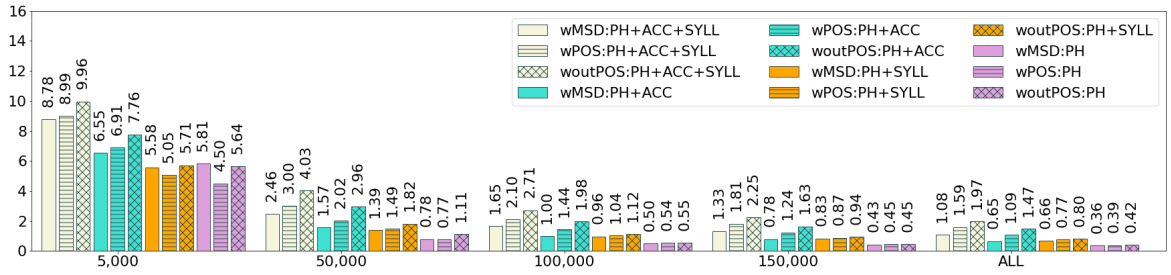

Figura 2. a

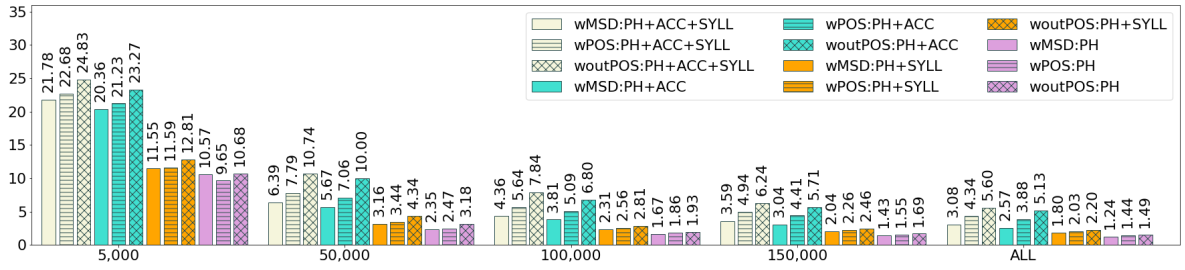

Figura 2. b

Fig 2. (a) CER și (b) WER) pentru diferite dimensiuni ale datelor de antrenare selectate în mod aleator, evaluate pentru: predicție lexicală completă (PH+ACC+SYL), eliminând din predicție silabificare (PH+ACC), eliminând din predicție silabificarea și accentul (PH). Rezultatele indică ratele de eroare cu POS (wPOS), cu MSD (wMSD) și fără nici o informație suplimentară în afară de forma cuvântului (woutPOS).

În Figura 2, unde prezentăm rezultatele scenariului de evaluare 1 combinat cu scenariul de evaluare 3, culorile corespund combinației de informație menținută în predicție iar simbolul de haşurare discriminează între cele două tipuri de input (cu sau fără informație morfosintactică). Se poate observa că după 100.000 de intrări, creșterea acurateții atinge un platou, dar totuși

există o îmbunătățire relativă de 24% începând de la 100.000 intrări și până la setul de date complet, în termeni de WER, pentru predicția întregii secvențe, atunci când se folosește informația morfosintactică. Din punct de vedere al contribuției sarcinilor de predicție la rata erorii, accentul are cea mai mare influență, lucru de așteptat pentru limba română unde accentul nu depinde de reguli predefinite (în imagine se pod vedea scăderile bruște ale ratelor de eroare de la evaluarea PH+ACC+SYLL (culoarea crem) la evaluarea PH+SYLL (culoarea cărămiziu, atunci când nu se evaluează acentul)). Un rezultat similar a fost prezentat în (Stan și Giurgiu, 2018). Așa cum ne așteptam, în cele mai multe etape de evaluarea, informația despre partea de vorbire și eticheta morfosintactică crește performanța modelului.

	LEMMA		**1-FORM**		**2-FORMS**		**All Forms**	
	WER	CER	WER	CER	WER	CER	WER	CER
woutPOS								
PH+ACC+SYLL	46.53%	19.03%	13.30%	5.00%	10.20%	3.86%	5.60	1.97
PH+ACC	45.02%	16.01%	12.17%	3.57%	9.28%	2.64%	5.13	1.47
PH+SYLL	22.31%	7.29%	5.47%	2.37%	4.15%	1.83%	2.20	0.80
PH	16.53%	5.30%	3.91%	1.36%	2.86%	0.86%	1.49	0.42
wPOS								
PH+ACC+SYLL	52.55%	20.69%	11.54%	4.49%	8.39%	3.24%	4.34	1.59
PH+ACC	50.87%	18.18%	10.42%	2.93%	7.49%	2.08%	3.88	1.09
PH+SYLL	27.82%	8.92%	5.71%	2.55%	4.21%	1.81%	2.03	0.77
PH	20.87%	6.50%	4.21%	1.50%	3.07%	0.96%	1.44	0.39
wMSD								
PH+ACC+SYLL	47.57%	18.38%	11.07%	4.29%	8.18%	3.16%	3.08	1.08
PH+ACC	46.35%	15.62%	9.86%	2.82%	7.21%	1.95%	2.57	0.65
PH+SYLL	24.01%	7.26%	5.46%	2.42%	4.13%	1.83%	1.80	0.66
PH	20.51%	6.90%	4.04%	1.46%	2.98%	0.95%	1.24	0.36

Tabelul 2. CER și WER pentru diferite dimensiuni ale datelor de antrenare selectate în modalitatea descrisă pentru scenariul 2, cu subset-urile *1-FORM* (30.150 intrări), *2-FORMS* (55.185 intrări) și *LEMMA* (35.890 intrări) evaluate pentru: predicție lexicală completă (PH+ACC+SYL), eliminând din predicție silabificare (PH+ACC), eliminând din predicție silabificarea și accentul (PH). Rezultatele indică ratele de eroare cu (wPOS) sau fără (woutPOS) etichete morfosintactice adăugate la intrarea în transformer; pentru comparație, pe ultima coloană reluăm rezultatele pentru lexiconul complet reprezentate în Figura 2.

În Tabelul 2 se poate observa o diferență foarte mare între ratele de eroare folosind subsetul de antrenare *LEMMA* și cele folosind subsetul de antrenare *1-FORM*. De asemenea, E(rror)R(ates)-urile subsetului LEMMA sunt de două ori mai mari decât ER-urile obținute cu

subestul de 5000 de intrări selectate aleator. Ne explicăm scăderea substanțială a ratelor de erorare în cazul subsetului *1-FORM* față de subsetul *LEMMA*, deoarece, așa cum spuneam atunci când am descris procesul de selecție a subset-urilor am avut grijă să ne asigurăm că păstrăm diversitatea formelor morfologice, chiar dacă am selectat o singură formă pentru fiecare lemă, în subsetul *1-FORM*. În contrast, subset-ul LEMMA conține doar formele de dicționar și nu învață nici o terminație morfologică. Un subset de intrări selectate aleator de 7 ori mai mic dublează probabilitatea rețelei de a învăța proprietățile terminațiilor morfologice.

În continuare, în cazul subsetului 2-FORMS, observăm că rata de eroare continuă să scadă semnificativ. Dacă comparăm performanțele acestor subseturi cu cele ale căror intrări sunt selectate aleatoriu, observăm că: 1-FORM (30.150 intrări), pentru scenariul wMSD, are o CER=4,29% pentru predicție concurentă, comparabilă cu CER=4,36% a subsetului de 100.000 de intrări în scenariul wMSD; 2-FORM (55.185 intrări), pentru scenariul wMSD, are o CER=3,16%, aproape de CER=3,08 obținută atunci când se folosește întregul set de date. Pentru setul de date cu

Aceste performanțe demonstrează că o selecție strategică a intrărilor reduce foarte mult munca de corectură manuală într-un scenariu de construire incrementală - cu adnotare automată și corectare manuală pe subset-uri de date – a unui lexicon.

Deși se observă și aici, în general, o imbunătățire a performanței legată de adăugarea informației POS și MSD, acest lucru nu este valabil pentru subsetul *LEMMA*. Aici, informația POS și cea MSD nu poate compensa absența variantelor morfologice, pentru că nu are ocazia sa învețe nici o asociere între terminații morfologice și etichete POS ale cuvintelor conținut (care reprezintă majoritatea covârșitoare a lexicului unei limbi), pe care să o aplice apoi cuvintelor noi (de exemplu: 1. la transcrierea fonetică a terminațiilor, pentru care există reguli de transcriere specială a sunetului "i" final, și a sunetelor "ce/ci/ge/gi/che/chi/ghe/ghi" la sfârșit de cuvânt; 2. la nivel de silabificare, rețeaua nu are ocazia să învețe despărțirea în silabe a terminațiilor morfologice). Dimpotrivă, faptul că ratele de eroare wPOS și wMSD sunt în acest caz mai mari decât ratele de eroare woutPOS ne face să presupunem că restricția de a prezice informație lexicală coerentă cu partea de vorbire a cuvântului (adică, de exemplu, pentru un verb, rețeaua decide în funcție de ce a învățat doar despe verbele din datele de antrenare) este de fapt nefericită, iar o predicție bazată doar pe forma morfologică este de preferat.

La nivel de accent, în fișierul de antrenare din scenariul *LEMMA*, 18.981 din 35.890 intrări (mai mult de jumătate) au accent pe ultima silabă (spre deosebire de scenariile *1-FORM* (6.329 din 30.150) și *2-FORMS* (11.966 din 55.185)). Dintre acestea, verbe care au accent pe ultima silabă sunt 6.421 din 7.158. Asta înseamnă că rețeaua transformer învață că probabilitatea ca accentul unui cuvânt să fie pe ultima silabă, în general, este mare, iar pentru verbe, aceasta este foarte mare. Pentru variantele morfologice din datele de test, acest comportament al rețelei este dezavantajos: de exemplu, pentru *"*a mânca*",* accentul cade pe ultima silabă doar pentru 20 din cele 30 de forme morfologice ale conjugării.

Procentul de cuvinte cu accent pe ultima silabă din lista de erori aferentă scenariului wPOS (24.125 intrări din 42.510, dintre care pentru 10.835 eroarea provine doar de la acest accent pe ultima silabă) este chiar mai mare decât cel din lista de erori aferentă scenariului wout POS (19.366 intrări din 37.806, dintre care pentru 10.183 eroarea provine doar de la accentul pe ultima silabă). Analizând în detaliu erorile specifice apărute cu subset-ul LEMMA wPOS vs LEMMA woutPOS, am observat că, din cele 12.085 de erori obținute cu LEMMA wPOS care nu se găsesc în lista de erori obținute cu LEMMA woutPOS, 7.365 de erori sunt asociate verbelor. În timp ce numărul erorilor distincte asociat celorlalte părți de vorbire rămâne relativ

același, numărul de erori asociate verbelor crește foarte mult (de mai mult de 5 ori), când trecem de la scenariul woutPOS la cel wPOS. Aceste rezultate demonstrează că preponderența cuvintelor cu accent pe ultima silabă din setul LEMMA produce un procent mare de erori în scenariile wPOS și wMSD. Așa cum spuneam anterior, restricția cuvântului la clasa părții de vorbire căruia îi aparține - în cazul acesta faptul că forțăm rețeaua să trateze verbele din datele de test asemănător verbelor din datele de antrenare - conduce la erori care ar fi putut fi evitate dacă rețeaua se uita doar la forma cuvântului.

Concluzii

Am arătat că folosirea resursei noi dezvoltate în ReTeRom aduce, în mod evident, îmbunătățiri unui sistem de TTS, în comparație cu resusele utilizate anterior. Am presupus și demonstrat că, atunci când ne interesează să evaluăm influența pe care cantitatea de date o are asupra ratelor de eroare, selectarea datelor în mod strategic și nu aleatoriu este esențială. De exemplu, în procesul de corectare manuală a unui lexicon de tipul celui pe care l-am construit noi, se poate reduce foarte mult timpul de corectare dacă selectăm un set esențial de intrări pe care să le corectăm și pe care să antrenăm instrumentul de adontare automată pentru a adnota intrările rămase, care ne așteptăm să fie mai corecte dacă am ales un set de anternare potrivit. În cazul nostru, vezi diferența între setul de antrenare *1-FORM*, setul de antrenare *LEMMA* și setul de antrenare selectat aleatoriu, de dimensiuni similare, dar producând efecte foarte diferite asupra rețelei la antrenare. Recomandăm deci o selecție a datelor conform scenariului 2 de evaluare atunci când este importantă reducerea set-ului de date de antrenare. În plus, după cum s-a observat, prin acest tip de selecție putem observa și o descreștere a ratelor de eroare atunci când sunt adăugate informații morfo-sintactice.

Bibliografie

Erjavec, Tomaz. "MULTEXT-East Morphosyntactic Specifications: Version 3.0." Supported By EU Projects Multext-East, Concede And TELRI (2004)

Gehring, J., M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, “Convolutional sequence to sequence learning,” CoRR, vol. abs/1705.03122, 2017.

V. Levenshtein, “Binary Codes Capable of Correcting Deletions, Insertions and Reversals,” Soviet Physics Doklady, vol. 10, p. 707, 1966.

Stan, A. and M. Giurgiu, “A Comparison Between Traditional Machine Learning Approaches And Deep Neural Networks For Text Processing In Romanian,” in Proceedings of the 13th International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR), 2018.

Stan, Adriana. “RECOApy: Data recording, pre-processing and phonetic transcription for end-to-end speech-based applications,” in Proceedings of Interspeech, Shanghai, China, 2020.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008.