MINISTERUL EDUCAȚIEI NAȚIONALE

UEFISCDI

Raport științific și tehnic pentru proiectul ROBIN
ROBIN-Dialog

REZUMAT

În cursul anului doi al proiectului ROBIN-Dialog au fost urmărite și realizate toate obiectivele incluse în Agenda Comuna și Planul de realizare pe anul 2019 și anume:

1) Transcrierea fonetică a cuvintelor din lexiconul validat
2) Alinierea cu semnalul vocal corespunzător;
3) crearea înregistrărilor vocale pentru cuvinte pentru care nu există înregistrări în corpus
4) Sistemele de antrenare ASR și TTS vor fi alimentate cu rezultatele activității precedente. Sistemele ASR si TTS vor fi testate și evaluate.
5) Implementarea și descrierea prototipului generic de sistem de dialog cooperant.

În cele ce urmează vor fi prezentate activitățile desfășutare pentru realizarea acestor obiective cu precizarea că toate sarcinile asumate au fost îndeplinite integral.

Capitolul 5.

Descrierea științifică și tehnică

Autori: Dan Tufiș, Radu Ion, Elena Irimia, Vasile Păiș, Maria Mitrofan (Carp), Verginica Mititelu, Eric Curea, Valentin Badea, George Cioroiu – Institutul de Cercetări pentru Inteligență Artificială ”Mihai Drăgănescu”

Mihai Dascălu, Ștefan Trăușan-Matu, Dragoș Corlătescu – Universitatea ”Politehnica” București

5.1 Transcrierea fonetică a cuvintelor din lexiconul validat

În etapa anterioară, descriam procedura de construcție a unui lexicon pe baza descrierilor micro-lumilor țintă. Vorbeam despre extragerea tuturor lemelor din aceste descrieri într-o listă inițială și menționam două strategii de extindere a acestei liste cu cuvinte similare sau aflate în relație semantică: utilizarea reprezentărilor vectoriale învățate automat (cunoscute și ca “word embeddings”) pentru a identifica cuvinte similare și utilizarea wordnetului românesc (RoWordnet) pentru a extrage hiperonime și sinonime. De asemenea, menționam utilizarea resursei interne tbl.wordfom.ro (peste 1.150.000 de intrări) pentru a genera familia de cuvinte a fiecărei leme și pentru a completa lexiconul ROBIN cu etichete morfo-sintactice. La sfârșitul etapei 2018, lexiconul (validat la nivel de lemă și etichetă morfosintactică) conținea 99150 înregistrări de forma: *<formă ocurentă>tab<lemă>tab<etichetă morfo-sintactică>*.

În cadrul acestei etape, ne-am propus, pe de o parte, (1) să rafinăm prin metode semantice lexiconul pentru a ne asigura ca nu există intrări care să fie în afara universului de discurs și pe de alta parte (2) să completăm lexiconul cu informația de silabificare, accent și transcriere fonetică, utile în aplicațiile de ASR și TTS. Rafinarea lexiconului a avut loc într-o cercetare descrisă în (Irimia et al., 2019), care își propunea să prezinte crearea lexiconului în contextul mai larg al proiectului ROBIN și să evalueze utilitatea vectorilor semantici și a wordnetului românesc în extinderea lexiconului. În acest context, am experimentat cu dezambiguizarea semantică în contextul micro-lumilor a cuvintelor din lexicon

(prin asocierea de sensuri din wordnet) înainte de expandarea lexiconului pe relațiile de hiperonimie și sinonimie. Astfel, am reușit să restrângem lista de leme din lexiconul expandat la 1827, iar lista de intrări din lexicon (conținând variantele morfologice ale acestor leme) s-a redus de la 99150 intrări la 27559 intrări. Considerăm că versiunea rafinată a lexiconului este mai utilă sistemului de dialog ROBIN, deoarece elimină ambiguități semantice și evită supraîncărcarea acestuia cu informație inutilă.

Completarea lexiconului cu informațiile de silabificare, accent și transcriere fonetică s-a făcut în două etape (descrise în raportul în extenso), intrările lexiconului având forma:

<formă>tab<lemă>tab<etichetă_morfo-sintactică>tab<silabificare>tab<accent>tab<transcriere_fonetică>

Exemplu:

cercetăm cerceta Vmip1p cer.ce.tăm cercet'ăm [tS e r tS e t @ m]

Împărțirea în silabe este marcată prin simbolul ".", accentul este marcat printr-un apostrof în poziție anterioară vocalei accentuate iar transcrierea fonetică este afișată între paranteze drepte. Alfabetul folosit pentru transcrierea fonetică este SAMPA[1].

Corectarea erorilor de silabificare, accent și transcriere fonetică (de așteptat în urma generării automate) a avut loc în ordinea silabificare -> accent -> transcriere fonetică. Etapizarea corecturii permite automatizarea anumitor pași, deoarece în unele cazuri transcrierea fonetică este dependentă, în mod determinist, de silabificare și accent (v. regulile de mai jos).

Corectarea silabificării s-a făcut integral manual, concentrându-ne pe: cuvintele care nu se găsesc în RoSyllabiDict; cuvinte care conțin silabe mai lungi de patru litere; cuvinte care conțin secvențe de vocale + semivocale, etc. În corectarea manuală a accentului, am acordat atenție specială cuvintelor cu două variante de accent (omonimii), forme care pot fi substantive sau verbe (ex.: „data"), forme verbale diferite ale aceleiași leme (ex.: „atribui"), forme verbale diferite pentru leme identice (ex. „alungi", cu lemele „alunga" și „alungi"). Raportul în extenso detaliază principalele probleme și rezolvarea lor.

Bibliografie

[Barbu, 2008] Barbu, Ana-Maria. "Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries." LREC (2008)

[Irimia et al., 2019] Irimia, E., Mitrofan, M., & Mititelu, V. B. (2019). Evaluating the Wordnet and CoRoLa-based Word Embedding Vectors for Romanian as Resources in the Task of Microworlds Lexicon Expansion. In Wordnet Conference (p. 176).

[Toma et al., 2017] Toma, Ștefan-Adrian, et al. "MaRePhoR—An open access machine-readable phonetic dictionary for Romanian." 2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). IEEE. (2017)

Obiectivul a fost integral îndeplinit, s-au transcris fonetic toate intrările lexiconului construit, s-au completat intrările cu informație suplimentară, necesară sistemelor de prelucrare a vorbirii (silabificare, marcarea accentului).

5.2 Aliniere text-voce și încărcarea în componenta de vorbire a corpusului CoRoLa

Având în vedere că înregistrările realizate în format .WAV sunt redări fidele ale fișierelor text, a fost realizată alinierea automată a acestora, utilizând un proces similar celui utilizat la alinierea fișierelor text-voce din componenta audio a corpusului CoRoLa. Acest proces, descris în (Boroș et al., 2018), constă în realizarea unui format intermediar de text fără semne de punctuație sau alte elemente în

[1] https://www.phon.ucl.ac.uk/home/sampa/romanian.htm

afară de cuvintele pronunțate, cu toate literele mici și urmat de alinierea propriu-zisă a cuvintelor cu sunetele înregistrate. Fișierele realizate în urma aplicării acestei proceduri sunt salvate cu extensia ”.lab”. Pentru alinierea propriu-zisă, a fost utilizată unealta software HTK (Young et al., 2002), fiind produse la final fișiere .phs conținând momentul de start și momentul de sfârșit asociat fiecărui fonem prezent în text. Pe baza acestora putându-se reconstitui momentele de început și sfârșit asociate fiecărui cuvânt din text.

În urma alinierii, fișierele au fost indexate în componenta audio a corpusului CoRoLa, fiind apoi disponibile pentru căutare în interfața acestuia, disponibilă online la adresa: http://89.38.230.23/corola_sound_search/. În urma realizării unei interogări, cuvântul găsit poate fi ascultat în fiecare fișier audio (pe baza alinierilor realizate) sau întreaga frază poate fi ascultată.

Referințe

[Boros et al., 2018] T. Boros, S.D. Dumitrescu, V. Pais, Tools and resources for Romanian text-to-speech and speech-to-text applications, In *Proceedings of the International Conference on Human-Computer Interaction - RoCHI 2018*, pp 46-53.

[Young et al., 2002] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. 2002. The HTK book. Cambridge university engineering department 3 (2002), 175.

Obiectivul a fost integral îndeplinit, toate înregistrările noi au fost aliniate cu transcrierile lor. În plus, ele au fost indezate și introduce in componenta de vorbire a corpusului de referință al limbii române CoRoLa.

5.3 Crearea înregistrărilor vocale pentru cuvinte care nu au înregistrări în corpus

Au fost identificate în lexiconul construit anul trecut aproape 300 de cuvinte pentru care nu există inregistrari vocale in corpusul CoRoLa. Din corpusul textual s-au extras propoziții ce conțineau cuvintele respective și s-au inregistrat propozițiile alese în rostiri de către doi bărbați și două femei. Au fost construite fișierele txt pentru care au fost create fișierele wav corespunzătoare cu frecvența de eșantionare 48khz și 44khz. Echipamentul de înregistrare a fost reprezentat de casti cu microfon (Huawei P20 Pro). Microfonul avea atat functie de "noise reduction" cat si de "echo cancellation". O parte din înregistrari au fost efectuate în camera izolată fonic și restul în camere obișnuite. Programul software care s-a folosit a fost Audacity, cu setarile default. Fișierele cu extensia .lab au fost pregătite corespunzător (v. secțiunea 5.1.2). În final, noile înregistrări au fost încărcate în secțiunea de voce a corpusului public CoRoLa.

Obiectivul a fost integral indeplinit, s-au realizat înregistrările de foarte bună calitate ale frazelor context pentru cuvintele țintă.

5.4 Sistemele de antrenare ASR și TTS vor fi alimentate cu noile date. Testarea și evaluarea sistemelor ASR si TTS.

Pentru antrenarea modulului de TTS au fost urmati pasii de la https://github.com/tiberiu44/TTS-Cube/blob/master/TRAINING.md. In bazele de date de care dispunea RACAI, existau deja fisiere pregatite pentru acest sistem. Cele mai multe fisiere, erau pentru vocea Ancai, motiv pentru care calitatea ei este cea mai buna. Calitatea vocii este bună, se poate înțelege clar mesajul transmis. Principalul dezavantaj al sistemului este timpul mare de prelucrare a textului. Chiar și pe echipamente hardware puternice, timpul de sinteza depinde de lungimea textului, astfel incat pentru o propozitie mai lungă poate ajunge la câteva secunde.

Pentru antrenarea modulului ASR S-a pornit de la toate bazele de date de care dispunea ICIA. Fisierele .txt care contineau caractere invalide (cifre, accente din alte limbi) au fost eliminate impreuna cu inregistrarile. S-a obținut astfel o serie de înregistrări cu transcrierea aferentă, transcriere care este curată și poate fi folosită pentru antrenat. Am dezvoltat un ASR in Kaldi cu ajutorul

resurselor din Corola. Demo-ul este disponibil la adresa http://relate.racai.ro/index.php?path=robin/asr.

Există limitare pentru fisiere input să fie doar .wav, Dimensiunea maximă nu reprezintă o limită dar serverul va da timeout cu un fișier mai lung de 10 minute pentru că durează mult procesarea. (>10 min). Inregistrările sunt din proiectul Robin iar dictionarul folosit la antrenare este 3-gram pruned cu rescoring pe 4 gram. Este folosit Kaldi cu modele GMM-HMM.
Performanta modelului este WER = 30%.

Demo:

Text rostit:

ÎN ROMÂNIA DOAR OPTSPREZECE LA SUTĂ DIN POPULAȚIE CONȘTIENTIZEAZĂ IMPORTANȚA SALVĂRII UNEI VIEȚI OMENEȘTI PRIN DONAREA DE ORGANE IAR UN PROCENT FOARTE MIC ESTE ÎNREGISTRAT ÎN RÂNDUL TINERILOR ÎNTRE ȘAISPREZECE ȘI DOUĂZECI ȘI CINCI DE ANI TRANSMITE CORESPONDENTUL RADIO ROMÂNIA ACTUALITĂȚI IOAN SUCIU CITÂND DATELE PREZENTATE ASTĂZI LA ARAD

Predicție:

ÎN ROMÂNIA DOAR OPTSPREZECE LA SUTĂ DIN POPULAȚIE CONȘTIENTIZEAZĂ IMPORTANȚA SALVĂRII UNEI VIEȚI OMENEȘTI PRIN DONAREA DE ORGANE IAR UN PROCENT FOARTE MIC ESTE ÎNREGISTRATĂ ÎN RÂNDUL TINERILOR ÎNTRE ȘAISPREZECE ȘI DOUĂZECI ȘI CINCI DE ANI TRANSMITE CORESPONDENTUL RADIO ROMÂNIA ACTUALITĂȚI IOAN SUCIU CITÂND DATELE PREZENTATE ASTĂZI LA ARAT

Un alt experiment de ASR a fost realizat în grupul de la UPB folosind alte metode (DeepSearch) decât cele utilizate la ICIA (GMM-HMM). Prezentarea și evaluarea acestui experiment sunt furnizate în continuare.

Setul de date SWARA
Seturile de date pentru limba română sunt limitate, având un număr mic de vorbitori. De obicei, se pot găsi înregistrări audio specifice pentru fragmente mici de text, care au fost create pentru antrenarea unor modele adaptate unor nevoie foarte specifice. De asemenea, nu există seturi de date mai dificile, cu zgomot de fundal sau vorbitori simultan. În contextul antrenării modelului DeepSpeech, aceasta este o limitare pentru obținerea unei performanțe bune în ceea ce privește WER.

În ciuda faptului că seturile de date în limba română sunt limitate ca dimensiune, corpusul SWARA (Stan et al., 2017) s-a dovedit a fi un bun candidat pentru antrenarea rețelei neurale de la DeepSpeech. Acesta conține aproape 21 de ore de conținut vorbit, înregistrat în condiții de studio de către 17 vorbitori, bărbați și femei. Aceste statistici arată o varietate bună, cu toate că rețeaua nu ar putea învăța să ignore zgomotul de fundal și alte probleme din lumea reală. De asemenea, setul de date este curățat, aliniat și conține transcrierile fonetice ale tuturor cuvintelor care apar în el. Acest lucru este foarte util, pentru că permite testarea unor metode variate, printre care și modele bazate pe HMM sau DeepSpeech.

Antrenare modelului DeepSpeech
Primul pas în antrenarea arhitecturii DeepSearch (Amodei et al., 2016) este segmentarea corpusului SWARA. O intrare a setului de date inițial conține transcrierea unei propoziții și un fișier .wav cu înregistrarea unui vorbitor care rostește propoziția curentă. Colecția acestor intrări conține aproape 20000 de elemente. Durata medie a unei înregistrări este de 30 secunde, iar propozițiile transcrise au în medie 10 cuvinte. Aceste intrări au fost împărțite în partiții de aproximativ 60%

Rezultate
Așa cum a fost menționat anterior, rețeaua DeepSpeech fost antrenată pe setul de date SWARA, modificând hiper-parametrii pentru a obține cel mai bun model în ceea ce privește WER. În timpul

experimentelor, dimensiunea celulelor a fost variată de la 512 la 2048, pentru a micșora modelul, păstrând același număr de straturi. WER a variat între 58% și 63% pentru diferite dimensiuni ale rețelei folosind modelul de limbă predefinit (vezi Tabelul 5.1). Deși WER nu a fost considerabil mai bun de la 1024 la 2048, antrenarea modelului din urmă dura aproape 4 ore. Prin comparație, procesul dura numai 1.5 ore pe modelele mai mici. Aceleași experimente au fost efectuate și cu modelul de limbă pe română KenLM (Heafield, 2011), iar WER s-a îmbunătățit substanțial cu aproximativ 20%, variind între 39% și 42%. Toate rezultatele din Tabelul 5.1 au fost obținute folosind o rată de antrenare de 0.0001, 20 de epoci, iar dimensiunea unui batch fiind 24. De asemenea, au fost făcute experimente și cu rate de antrenare mai mari, dar rezultatele au fost mai slabe. De asemenea, numărul de epoci a fost variat între 15 și 20, dar rezultatele au fost similare.

Tabelul 5.1. WER pentru diferite configurații ale modelului (rata de antrenare 0.0001 și 20 de epoci).

Dimensiunea celulei	**WER folosind modelul predefinit**	**WER cu modelul în română**
2048	58 %	39%
1024	60 %	40%
512	63%	42%

Tabelul 5.2 conține exemple de propoziții din setul de test care au fost recunoscute folosind modelul de dimensiune 1024 și modelul de limba română. Din pricina faptului că există diferențe între formele cuvintelor, iar unele elemente au fost recunoscute drept pauze în loc de foneme, WER este influențat. Acesta este calculat drept numărul de înlocuiri, adăugări și ștergeri de cuvinte, împărțit la numărul de cuvinte din propoziția corectă. În acest caz, rezultatul a avut mai multe cuvinte, influențând WER. Acest lucru arată că un model specific de limbă poate îmbunătăți recunoașterea și compensa cu un corpus mai mic de antrenare. O altă metrică măsurată a fost eroarea la nivel de caracter (eng. „Character Error Rate" – CER), care a fost în jur de 15% pentru variantele cu model specific de limbă.

Tabelul 5.2. Exemple de propoziții din setul de test folosind modelul cu dimensiunea 1024.

Propoziția originală	**Propoziția recunoscută**
"erorile ne-au costat"	"erori le ne au costat"
"apreciază punctualitatea"	"aprecia să punctualitate"
"directorii nu organizaseră niciun concurs"	"direct primul ma nica să rând cum concurs"
"a precizat antrenorul sorin cârțu"	"apreciat an trenul ori încă u"

Referințe

[Stan et al., 2017] Stan, A., Dinescu, F., Ţiple, C., Meza, Ş., Orza, B., Chirilă, M., & Giurgiu, M. (2017). The SWARA speech corpus: A large parallel Romanian read speech dataset. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)* (pp. 1-6): IEEE.

[Amodei et al., 2016] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., & Chen, G. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *33rd Int. Conf. on Machine Learning* (pp. 173–182). New York, NY, USA: JMLR: W&CP volume 48.

[Heafield, 2011] Heafield K. KenLM: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation* (pp. 187-197): Association for Computational Linguistics.

Obiectivul a fost integral îndeplinit, modulele de TTS si ASR au fost antrenate, testate și evaluate.

5.5 Implementarea prototipului generic de sistem de dialog cooperant

Sistemul de dialog cooperant dezvoltat pentru proiectul ROBIN-Dialog este un sistem de dialog care funcționează pentru o „microlume" definită de utilizator. Așa cum am descris în rapoartele anterioare, o microlume este o mulțime de concepte împreună cu relațiile care se stabilesc între ele despre care poate fi vorba într-un schimb de interacțiuni dintre utilizator și robot.

Sistemul de dialog a fost implementat în Java 1.8 și conține următoarele componente:

- **Definiția unei microlumi** care se face într-un fișier text dar care, datorită nivelurilor de abstractizare oferite, se poate face și în clase derivate, direct în Java;
- **Analiza interogării în limba română** care se face automat cu RELATE (Păiș et al., 2019) și în urma căreia sistemul de dialog recunoaște concepte și predicate prezente în cerere. De exemplu, folosind dialogul de mai sus, în întrebarea „În ce sală se ține laboratorul de robotică?", analizorul de limbaj natural extrage variabila „sala" fiind cea căreia trebuie să i se găsească o valoare în microlumea dată și „laboratorul de robotică" care este un eveniment (laborator în cazul de față) ancorat în microlumea dată (adică definit în microlumea dată de către inginerul de cunoștințe);
- **Algoritmul de unificare** care caută cel mai apropiat predicat din universul de discurs (microlume) care poate răspunde cererii utilizatorului prin legarea uneia sau mai multor variabile lăsate în suspensie. Pentru exemplificare, folosind de asemenea dialogul de mai sus, întrebarea „Unde se află sala în care se ține laboratorul de robotică?" se traduce automat în predicatul `ține(laboratorul de robotică, S, P)`, predicat care unifică cu `ține(laboratorul de robotică, sala 412, Popa Eugenia)` prin atribuirile `S = sala 412` și `P = Popa Eugenia`;
- **Managerul de dialog** care este bucla (infinită) de I/O pentru sistemul de dialog: așteaptă întrebări de la utilizator, cere informații suplimentare dacă este necesar și verbalizează informațiile solicitate de utilizator sau clasifică așteptările utilizatorului pentru sistemul de planificare al robotului (cel care decide ce acțiuni poate efectua Pepper).

Sistemul de dialog ROBIN-Dialog Este disponibil la https://gitlab.com/raduion/robindialog și este scris în Java 1.8. Fiecare clasă care implementează componente ale sistemului are o versiune abstractă (clasă sau interfață) care poate fi rescrisă dacă implementările existente nu sunt suficiente pentru un scop sau altul și, de asemenea, este comentată în stilul JavaDoc pentru a facilita utilizarea.

Pachetul `ro.racai.robin.dialog` conține clasele care implementează **managerul de dialog** și **algoritmul de unificare**[2]. Așa cum am mai menționat, din întrebarea utilizatorului, analizorul de limbaj natural extrage un predicat pe care algoritmul de unificare încearcă să-l „potrivească" cât mai bine cu un predicat din universul de discurs. Predicatele au un număr variabil de argumente iar fiecare argument are un tip. De exemplu, folosind scenariul de orientare a studenților în facultate, există argumente de tip „sală" sau de tip „curs". Pentru a facilita potrivirea, fiecare concept are o mulțime de sinonime[3] cum ar fi de exemplu „sală, cameră, încăpere" sau „curs, laborator, seminar". Algoritmul de unificare va putea potrivi un predicat cu un număr mai mic de argumente (sub-specificat) cu un predicat din microlume care are mai multe argumente, cu condiția ca variabilele legate (cele care sunt instanțiate) să fie identice și de același tip.

Pachetul `ro.racai.robin.mw` conține clasele care construiesc **universul de discurs al sistemului de dialog** dintr-o microlume. În implementarea actuală, o microlume se construiește automat dintr-un fișier `.mw` în care inginerul de cunoștințe descrie conceptele și predicatele care sunt „adevărate" (există) în microlumea respectivă.

Pachetul `ro.racai.robin.nlp` conține clasele **analizorului de interogări în limba română**, algoritm care extrage un predicat cu argumente parțial instanțiate din întrebarea utilizatorului. Acesta folosește platforma RELATE pentru a preprocesa întrebarea (segmentare lexicală, adnotare cu etichete morfo-sintactice, lematizare și analiza cu relații de dependență sintactică).

---

[2] https://gitlab.com/raduion/robindialog/blob/master/src/main/java/ro/racai/robin/dialog/RDUniverse.java

[3] Termen folosit aici într-un sens mai larg. Poate fi vorba și de hiperonime sau hiponime.

Un exemplu de fișier `.mw` se află în GitLab[4] și definește o microlume care facilitează orientarea unui student în facultate (în cazul nostru, în Facultatea de Automatică și Calculatoare a Universității POLITEHNICA din București). Inginerul de cunoștințe poate defini următoarele:

- **Concepte**: entități *despre care poate fi vorba* în dialogul dintre om și mașină;
- **Obiecte** referite de fiecare concept: instanțe ale fiecărui concept;
- **Obiecte de diverse tipuri** despre care poate fi vorba;
- **Predicate**: relații care se stabilesc între diverse instanțe de concepte sau obiecte de diverse tipuri și care au valoarea de adevăr „adevărat" în această microlume.

În raportul în extenso sunt exemplificate toate aceste entități înclusiv sunt indicate scheme noi de dialog.

Concluzii

Prototipul avansat al sistemului ROBIN-Dialog se află la https://gitlab.com/raduion/robindialog. Clasele sunt documentate (în limba engleză) cu comentarii în stilul JavaDoc. În stadiul actual, ROBIN-Dialog poate răspunde la întrebări factuale și poate continua conversația pe marginea subiectului început, cu vocabularul constrâns al microlumii active.

Referințe bibliografice

[Păiș et al., 2019] Păiș Vasile, Tufiș Dan și Ion Radu. Integration of Romanian NLP tools into the RELATE platform. În International Conference on Linguistic Resources and Tools for Natural Language Processing. Noiembrie 2019.

spaCy și RASA

O soluție alternativă pentru implementarea dialogului om-robot în limba română a fost investigate de grupul de cercetare de la UPB. Ea se bazează pe instrumente open-source spaCy și RASA. Prezentăm mai jos această abordare.

spaCy

spaCy este o bibliotecă în Python gândită să ușureze munca dezvoltatorilor de aplicații în care aceștia au nevoie de procesare de limbaj natural. Este ușor de instalat, asemănător cu orice altă bibliotecă de Python, este ușor de folosit datorită documentației clare. Din punct de vedere al performanțelor măsurate pe probleme specifice domeniului de procesare a limbajului natural, spaCy obține rezultate extrem de bune. Un alt beneficiu al acestei biblioteci îl constituie interfața unificată pentru 51 de limbi, printre care se numără și engleza, franceza, româna și germana. Acest lucru simplifică munca dezvoltatorilor care nu sunt nevoiți să învețe diferite moduri de a lucra cu biblioteca în funcție de limbă. spaCy oferă modele preantrenate de dimensiuni diferite (mic, mare) pentru limbile cu un grad de utilizare mai mare cum ar fi engleza și franceza. Pentru celelalte limbi suportul este minimal și modelul trebuie antrenat de către programatorii care vor să-l folosească pentru limbile respective. În acest caz se află și limba română. După antrenarea pe limba română (v. amănunte în raportul detaliat) s-au obținut următoarele rezultate:

- Part-of-speech tag accuracy (Tags_acc):97.288%;
- Unlabeled Attachment Score (UAS): 88.589%;
- Labeled Attachment Score (LAS): 81.172%;
- Named entity accuracy - Precision (ents_p): 75.514%;
- Named entity accuracy – Recall (ents_r): 78.102%;
- Named entity accuracy – F score (ents_f): 76.786%.

Este util să menționăm că la adresa http://relate.racai.ro este disponibil un lant de prelucrări pentru limba română, gata antrenat cu mai multe facilități și performanțe mai bune decât spaCy.

[4] https://gitlab.com/raduion/robindialog/blob/master/src/main/resources/precis.mw

Acesul este liber pentru prelucrări mono-fișier și pe bază de credențiale (user&password) pentru prelucrări masive de texte.

RASA

Rasa (https://rasa.com/, 2019) este o soluție open source de unelte de învățare automată și procesare de limbaj natural, ce are scopul de facilita implementarea de agenți inteligenți care să fie capabili de a purta conversații fluente și coerente cu utilizatorii. Rasa este distribuit sub forma unei versiuni gratuite ce poate fi modificată dupa propriile nevoi și scopuri, dar se poate opta și pentru o versiune enterprise ce oferă capabilități sporite. Pentru a face posibilă integrarea bibliotecii RASA în limba română este de dorit utilizarea unui model spaCy pe limba română, lucrul dezvoltat în paralel în cadrul acestui proiect. De asemenea, este nevoie de un corpus de antrenare în care să fie adnotate atât intenții cât și entități. Pentru acest proiect a fost generat un corpus, manual, disponibil în anexă. Odată având aceste componente, pentru instalare este nevoie să se urmeze doar pașii de instalare prezenți pe site-ul RASA care explică integrarea unei noi limbi.

Referințe

[Bouchard , 2007] Bouchard, G. Efficient bounds for the softmax function, applications to inference in hybrid models.

[Forney, 1973] Forney, G. D. The viterbi algorithm. *Proceedings of the IEEE, 61*(3), 268-278.

[Gal and Ghahramani , 2016] Gal, Y., & Ghahramani, Z. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems* (pp. 1019-1027).

[Kingma and Ba, 2014] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Nair and Hinton, 2010] Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 807-814).

[Shore and Johnson, 1980] Shore, J., & Johnson, R. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on information theory, 26*(1), 26-37.

5.6 Diseminare

Dan Tufiş, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, Radu Ion, and George Cioroiu (2019). Making Pepper Understand and Respond in Romanian. 22ndINTERNATIONAL CONFERENCE ON CONTROL SYSTEMS AND COMPUTER SCIENCE, Bucuresti, 28-30 mai 2019

Elena Irimia, Maria Mitrofan, Verginica Barbu Mititelu (2019). Evaluating the Wordnet and CoRoLa-based Word Embedding Vectors for Romanian as Resources in the Task of Microworlds Lexicon Expansion. The 10th Global WordNet Conference, Wroclaw, Polonia, 22-27 iulie 2019

Maria Mitrofan, Verginica Barbu Mititelu, Grigorina Mitrofan (2019). MoNERo: a Biomedical Gold Standard Corpus for the Romanian Language, The 18th BioNLP Workshop and Shared Task, Florenta, Italia, 1 August 2019

Panaite, M., Ruseti, S., Dascalu, M., Balyan, R., McNamara, D. S., & Trausan-Matu, S. (2019). Automated Scoring of Self-explanations Using Recurrent Neural Networks. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou & J. Schneider (Eds.), *14th European Conference on Technology Enhanced Learning (EC-TEL 2019)* (pp. 659–663). Delft, Netherlands: Springer.

Panaite, M., Ruseti, S., Dascalu, M., & Trausan-Matu, S. (2019). Towards a Deep Speech Model for Romanian Language. In *4th Int. Workshop on Design and Spontaneity in Computer-Supported Collaborative Learning (DS-CSCL-2019), in conjunction with the 22nd Int. Conf. on Control Systems and Computer Science (CSCS22)* (pp. 416–419). Bucharest, Romania: IEEE.

Boboc, I. G., Dascalu, M., & Trausan-Matu, S. (2019). Image Style Transfer using Text Descriptions. In *International Conference on Human-Computer Interaction (RoCHI2019)* (pp. 22–29). Bucharest, Romania: MatrixRom.

Nenciu, B., Ruseti, S., & Dascalu, M. (2018). Extracting Actions from Romanian Instructions for IoT Devices. In V. Pais, D. Gifu, D. Trandabat, D. Cristea & D. Tufis (Eds.), *13th Int. Conf. on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2018)* (pp. 168–176). Iasi, Romania.

Obiectivul a fost integral realizat. Toate lucrările menționează cu multumiri, finanțarea cercetărilor de către proiectul ROBIN-Dialog. Site-ul proiectului ROBIN-Dialog a fost actualizat cu rapoartele integrale și lucrările științifice realizate.

Toate obiectivele asumate de proiectul component ROBIN-Dialog au fost îndeplinite complet.
Servicii de cercetare şi tehnologice oferite de Institutul de Cercetări pentru Inteligență Artificială (https://erris.gov.ro/RACAI-ICIA): Interogare corpus de referință al limbii române scrisă și vorbită corola.racai.ro; TTL http://ws.racai.ro/ttlws.wsdl, Modular Language Processing for Lightweight Applications (MPLA) cu prelucrări pentru mai mult de 40 de limbi; sistemul ROBIN-dialog de prelucrare a dialogurilor în micro-lumile țintă; lexicon extins pentru aplicații de prelucrare a vorbirii;
Locuri susținute de acest proiect: 6 cercetători cu vechime în ICIA (D. Tufiș, V. Mititelu, E. Irimia, M. Mitrofan, R. Ion, E. Curea) plus 2 tineri cercetători angajați pe proiect (G. Cioroiu, V. Badea).
CEC-uri: sumele alocate cec-urilor nu au fost valorificate, în principal pentru că nu s-au identificat oportunități conforme cu reglementările de acordare.