Universitatea din Bucureşti
Facultatea de Matematica şi Informatică
Şcoala Doctorală de Informatică

Teză de Doctorat

Computational Aspects in Natural Language Processing

Aspecte Computaţionale în Procesarea Limbajului Natural

Rezumat

Autor: Octavia Maria Şulea
Conducător Ştiinţific: Prof. dr. Liviu P. Dinu

2019

Cuprins

1. Introducere

Inteligenţa Artificială (IA), ca domeniu de cercetare al informaticii, intră în al 63-lea an la momentul finalizării acestei teze, cu celebrul Test Turing (Turing 1950) apropiindu-se de cel de-al 70-lea an şi odată cu aceasta şi de obsolescenţa sa (Schmelzer 2018). Pe măsură ce mijloacele de comunicare în masă devin din ce în ce mai conştiente de faptul că datele generate de maşini trec drept sau chiar întrec oameni în tot mai multe contexte (Merchant (2015), Davies (2016), Borowiec and Lien (2016), Nieva (2018), Weiner (2018)), conţinutul generat de computer devine din ce în ce mai greu de detectat chiar şi cu ajutorul calculatorului, după cum demonstrează conferinţele de IA de top precum cea de-a treizeci şi şasea Conferinţă Internaţională despre Învăţarea Automată (ICML 2019) care organizează ateliere pentru detectarea lor [1], sau workshop-uri importante precum PAN care demarează sarcini publice pentru diferenţierea între conţinutul generat de maşină şi cel uman [2], şi ne arată cercetătorii din domeniu care devin din ce în ce mai îngrijoraţi de publicarea modelelor lor neuronale deja antrenate, întrucât acestea ar putea fi folosite în scopuri malefice, de exemplu pentru propagandă (Radford et al. n.d., Zellers et al. 2019).

Iniţial un sub-domeniu al IA, Învăţarea Automată (*Machine Learning* - ML) s-a detaşat lent de obiectivul original de a imita inteligenţa umană şi a trecut la rezolvarea sarcinilor specifice cu metode statistice. A devenit repede evident pentru cercetătorii din IA care se concentrau pe învăţare automată, informaţi fiind de diferite teorii ale minţii (*theory of mind*), că a învăţa automat din date anterioare o ipoteză generalizantă după care să se facă predicţii asupra datelor viitoare - scopul principal al unui model de ML supervizat - nu este singura componentă a inteligenţei umane. Astfel, domeniul ML s-a maturizat în a considera învăţarea automată ca o modalitate de soluţionare a unor sarcini specifice, mai degrabă decât în a căuta inteligenţa artificială generală (*Artificial General Intelligence* - AGI) (Deutsch 2012). Cu toate acestea, în ultimii ani, cercetarea din cadrul ML a făcut un cerc complet cu lucrări precum Graves et al. (2014) care încorporează maşina Turing în reţele neuronale recurente (RNN) şi Gross et al. (2017) care compară reţelele adversare generative (*Generative Adversarial Neural Networks* - GAN) cu Testul Turing. Comunitatea în mare a văzut o trecere de la învăţare automată supervizată peste modele proiectate cu caracteristici specifice fiecărei sarcini la modele de reţele neuronale adânci (*Deep Neural Networks* - DNN) semi-supervizate şi complet nesupervizate capabile să transfere abilităţile învăţate de la o sarcină la alta (Weiss et al. 2016, Tan et al. 2018) sau să generalizeze doar din câteva exemple (Ren et al. 2018), apropiindu-ne astfel aparent de obiectivul final de AGI.

[1] Synthetic Realities: Deep învăţarea pentru detectarea falsurilor AudioVisual: `https://sites.google.com/view/audiovisualfakes-icml2019/`

[2] A se vedea sarcina publică de diferenţiere între boţi şi oameni la workshop-ul PAN @ CLEF 2019: `https://pan.webis.de/clef19/pan19-web/author-profiling.html`

Un domeniu de cercetare al informaticii care se intersectează cu învăţarea automată, dar este considerat la fel de vechi ca IA datorită cerinţei testului Turing ca interacţiunea om-maşină să se facă prin text, este Procesarea Limbajului Natural (*Natural Language Processing* - NLP). Obiectivul principal al NLP, confundat în ultimii ani cu obiectivul general al IA de a imita inteligenţa umană (Metz 2015), este de a face comunicarea între computer şi utilizatorii umani tipici, spre deosebire de utilizatorii experţi (adică programatorii), cât se poate de *naturală* (Deangelis 2014). Mai exact, NLP se străduieşte să găsească un punct de întâlnire între computer, operat prin *limbaje formale* (adică limbaje de programare) şi individul uman, cooperând cu alte persoane prin *limbaj natural*. Atunci când presupunem că acesta este obiectivul real al NLP şi îl observăm în contextul supra-abundenţei de date textuale de pe Internet, devine mai uşor să înţelegem definiţia şi soluţia multor sarcini şi subdomenii din cadrul NLP de-a lungul deceniilor: Traducerea Automată, Extragere Informaţiei, Regăsirea Informaţiei, Question Answering, Clasificarea Textului, etc. O problemă, însă, a rămas constantă în NLP, deoarece face parte dintr-o problemă mai largă din ML: reprezentarea informaţiilor lingvistice în mod digital (Bowman et al. 2017). Complexitatea acestei probleme se dezvăluie atunci când ne uităm la sarcini NLP pentru limbi cu o variabilitate ridicată a datelor, cu spaţiu împrăştiat (*sparse*) al datelor de intrare sau al datelor de ieşire, deoarece datele sunt abundente, dar nu în modul în care am vrea noi să fie (Jain 2016).

Modelarea variaţiei (alternanţei) la nivelul unităţii în NLP este o problemă importantă. În limbile cu o morfologie inflexionară bogată, NLP a văzut sarcini precum *stemming* sau *lematizare* drept încercări de a reduce algoritmic cuvântul inflectat (de exemplu, *înveţi*) la forma sa canonică (de exemplu, *învăţa*) pentru a îmbunătăţi scorul TF-IDF al documentelor în sarcini de Regăsire a Informaţiei (Kamps et al. 2004), sau pentru a micşora dimensiunea vocabularului şi, astfel, reduce dimensiunea spaţiului datelor de intrare în modelele bag-of-words utilizate în Clasificarea Textului. În lingvistica diacronică şi similaritatea limbajului, schimbarea aceluiaşi cuvânt de-a lungul timpului sau de la o limbă la alta sunt fenomene lingvistice importante care trebuie modelate cu exactitate pentru sarcinile din Traducere Automată (Kondrak et al. 2003, Ciobanu and Dinu 2014). În stilistică şi lingvistică criminalistică sau socio-lingvistică, a fi capabil să modeleze variaţia în stilul unui autor (adică idiolect) sau al unui grup dintr-o populaţie (adică varietate de limbă sau dialect) este esenţial pentru cercetătorii ce se preocupă cu sarcinile de identificare sau profilare a autorului (Coulthard (2004), Wright (2014), Zampieri et al. (2018)). Chiar mai mult, în sarcinile de înţelegere a limbajului natural (*Natural Language Understanding* - NLU) sau inferenţă în limbajul natural (*Natural Language Inference* - NLI), cercetătorii încearcă să modeleze alternanţa de sens la nivelul propoziţiei sau al discursului şi să-l separe de stilul în care este comunicat (Wang et al. 2019).

Tematica acestei teze se referă la alternanţa lingvistică. Ne uităm la modalităţi

de calcul pentru a modela alternanţele în contextul identităţii lingvistice şi prezentăm rezultatele în domeniile Morfologiei Computaţionale, Traducerii Automate, Extragerii Informaţiei, Profilării şi Atribuirii Autorului, Inferenței în Limbajului Natural şi Clasificării Textului.

2. Contribuţii

În această teză, prezentăm şi extindem rezultatele noastre publicate. Lista completă de 21 de articole se găseşte în secţiunea 4. În ceea ce urmează, facem rezumatul principalelor contribuţii.

În Morfologia Computaţională, am propus o metodă de *învăţare slab-supervizată* pentru a prezice inflexiunea substantivelor (Şulea 2016) şi a verbelor (Dinu et al. 2011, 2012*b*) în limba română, o limbă bogată din punct de vedere al morfologiei sale inflexionare. Am introdus un mod inedit de modelare a tiparelor de inflexiune folosind seturi fixe de expresii regulate cu care am etichetat datele noastre substantivale şi verbale şi am antrenat clasificatori liniari cu vector de suport (SVC) pe n-grame de caractere extrase din formele lipsite de inflexiune (adică formele canonice, sau de dicţionar) pentru a prezice clasele de inflexiune ale cuvintelor. Am arătat că n-gramele de caractere, în special atunci când se evidenţiază sfârşitul cuvântului în formă canonică prin adăugarea unui marcaj de încheiere, sunt caracteristici suficiente pentru a prezice cu exactitate clasa verbală şi uşor mai puţin puternice în prezicerea clasei nominale pentru limba română. Explicăm mica scădere în performanţă a clasificatorilor nominali a fi cauzată de o inflexiune mai bogată a substantivelor româneşti parţial neregulate. Arătăm că, deşi putem prezice clasa de inflexiune din forma canonică, reversul nu e valid. Anume, nu putem genera exact formele inflectate dacă pornim de la regulile clasei de inflexiune şi formele de dicţionar (infinitivul pentru verbe şi forma de nominativ-acuzativ singular pentru substantive).

De asemenea, am definit învăţarea inflexiunii morfologice drept o sarcină de *etichetare a secvenţei* şi am considerat fiecare caracter din forma de dicţionar ca o *etichetă* sau un *tip de literă*: caracterul grafic *t* din verbul *a cânta* devine acum eticheta t_0 (Dinu, Şulea and Niculae 2013). Am folosit Conditional Random Fields (CRF) pentru a prezice eticheta fiecărui caracter din infinitivul verbelor româneşti şi am arătat că acest lucru duce la îmbunătăţiri ale problemei de învăţare a inflexiunii, dar nu soluţionează necesitatea etichetării manuale, o problemă şi cu metoda slab-supervizată.

Am explorat influenţa ambiguităţii lexicale şi fonologice asupra inflexiunii prezentând sistemul nostru de clasificare a genurilor pentru substantivele româneşti (Dinu et al. 2012*d*,*a*) şi sistemul nostru de etichetare a secvenţelor pentru silabisirea în limba română (Dinu, Niculae and Şulea 2013). Am arătat apoi aplicabilitatea unui model de învăţare

profundă (*deep learning*) a transformării unei secvenţe într-o altă secvenţă (*sequence to sequence learning* - Seq2Seq) la sarcina învăţării inflexiunii şi silabisirii ortografice (Şulea et al. 2018), îmbunătăţind performanţa modelelor noastre anterioare de învăţare a inflexiunii şi necesitând mai puţină supraveghere umană.

În cele din urmă în studiile noastre din morfologia computaţională, am propus o metodă nesupervizată pentru a genera inflexiunea completă a verbelor în română, spaniolă şi finlandeză (Şulea et al. 2019) şi a substantivelor în română, germană şi finlandeză (Şulea and Young 2019) folosind *modelarea neuronală a limbii*.

La nivel de frază, am propus o modalitate de a antrena reprezentări vectoriale bilingve pentru cuvinte (*bilingual word embeddings*), astfel încât reprezentările aceloraşi entităţi în limbi diferite să rămână similare (Şulea et al. 2016). Acest mecanism am arătat că duce la îmbunătăţirea traducerii numelor entităţilor şi la evitarea necesităţii construirii separat a unor sisteme de recunoaştere a entităţilor numite (*Named Entity Recognition* - NER) pentru diferite limbi. De asemenea, am analizat legăturile între texte scurte provenite de pe Twitter şi am arătat că, şi atunci când subiectul rămâne acelaşi, schimbarea limbii pe Twitter se întâmplă prea repede pentru ca modelele de indentificare a legăturii textuale (*Recognizing Textual Entailment* - RTE) să rămână robuste în timp (Şulea 2017). sulea De asemenea referitor la schimbările de limbă, am introdus o metodă fiabilă pentru clasificarea temporală a textelor literare româneşti (Ciobanu, Dinu, Şulea, Dinu and Niculae (2013), Ciobanu, Dinu, Dinu, Niculae and Şulea (2013)) şi am arătat că aceleaşi strategii nu sunt la fel de solide pentru textul francez provenind din domeniul juridic (Şulea, Zampieri, Malmasi, Vela, Dinu and van Genabith (2017), Şulea, Zampieri, Vela and van Genabith (2017)). În domeniul juridic, am prezentat şi clasificatorul nostru capabil să prezică hotărârea Curţii Supreme franceze într-un caz, având în vedere doar descrierea sa. În acest scop, prezentăm, de asemenea, tehnica noastră de reducere a spaţiului etichetelor, prin utilizarea clusterizării ierarhice şi alegerea unei tăieturi la nivel superior în dendogramă.

La nivel de autor, am introdus un model care să facă distincţia între textele literare de la un autor şi pastişele realizate de alţii (Dinu et al. 2012*c*). De asemenea, prezentăm modelele noastre pentru a prezice cu exactitate caracteristicile demografice ale utilizatorilor de Twitter doar pe baza textelor lor (Şulea and Dichiu 2015).

La nivel de limbă şi mai departe, introducem o metodă de a distinge între varietăţile limbii portugheze (Zampieri et al. 2016) şi o metodă de identificare a adreselor URL care indică site-uri web cu un conţinut periculos (Şulea et al. 2015), arătând că şi adresele URL pot fi, de asemenea, descrise foarte precis pe baza punctuaţiei şi se pot aplica în acest sens tehnici similare de clasificare precum în cazul textului scris în limbaj natural. În sfârşit, am folosit aceeaşi tehnică din capitolul 2 pentru a genera cu exactitate URL-uri noi, valide.

În ultimul capitol, propunem o modificare la Testul Turing, redenumindu-l *Testul*

Imortalităţii. În esenţă, această formă modificată a clasicului test de inteligenţa pentru un agent artificial (calculatorul) cere validatorului (omului) să treacă de la chestiunea originală, *maşină vs. om*, şi să se întrebe în schimb: *eu sau maşina?* Astfel, ne referim la ultimele rezultate din NLP care merg spre personalizarea modelelor neuronale de limbă (*neural language models*) la nivel de individ, dar mai ales această modificare este creată pentru a lua în considerare viziunea solipsistică asupra naturii umane, unde singurul lucru care poate fi validat este validatorul în sine.

3. Structura Tezei

În cele ce urmează, prezentăm structura acestei teze.

În capitolul 2, discutăm toate lucrările noastre din domeniul Morfologiei Computaţionale. În prima secţiune, explicăm conceptul de limbi bogate morfologic din perspectiva inflexiunii. Explicăm fenomenele lingvistice de *apofonie* şi *alomorfism*. Exemplificăm aceste concepte în limba română, germană şi spaniolă. De asemenea, discutăm ambiguitatea la nivel de *gen* şi între *diftong* şi *hiat* pentru limba română.

În secţiunea 2, prezentăm două strategii pentru învăţarea automată a inflexiunii: un model slab supervizat şi un model de etichetare a secvenţei. Pentru primul model, am prezentat regulile de inflexiune create pentru a eticheta substantive şi verbe româneşti şi am detaliat modul în care am antrenat clasificatori pe bază de vectori suport (SVC) pentru a învăţa asocierea dintre forma de dicţionar a unui cuvânt (forma canonică) şi regula lui de inflexiune folosind doar n-grame de caractere ale formei canonice. Arătăm că procesul invers, în sensul de a genera formele inflectate din regula inflexionară, nu este fezabil prin această metodă. Pentru a doua strategie, considerând învăţarea inflexiunii ca o sarcină de etichetare a secvenţei, detaliem modul în care am reprezentat fiecare caracter din forma de dicţionar a unui verb românesc ca o etichetă care reprezintă modelul său de a alterna şi cum am folosit CRF-uri pentru a prezice aceste etichete. În secţiunea 3 discutăm activitatea noastră în clasificarea pe gen şi silabisirea cuvintelor în limba română, dezvăluind influenţa indirectă pe care genul şi structura de silabe o au asupra inflexiunii.

Secţiunea 4 din capitolul 2 prezintă experimentele noastre ce consideră învăţarea inflexiunii drept o sarcină de învăţare a transformării unei secvenţe într-o altă secvenţă (*sequence to sequence learning* - Seq2Seq) şi folosim o arhitectură pe bază de reţele neuronale recurente (RNN). În cele din urmă, secţiunea 5 a aceluiaşi capitol prezintă eforturile noastre în generarea tuturor formelor inflectate, nesupervizat, doar din forma canonică. Arătăm că acest lucru poate fi obţinut pentru verbele româneşti, spaniole şi finlandeze şi substantivele româneşti folosind destul de uşor un model de limbă bazat pe RNN cu un mecanism de atenţie şi că acelaşi lucru este mai greu de obţinut pentru

substantivele germane şi finlandeze. De asemenea, discutăm rezultatele noastre din experimentarea cu modele de limbaj *tabula rasa* şi modele pre-antrenate.

Capitolul 3 prezintă lucrările noastre în modelarea alternanţei la nivel de frază şi propoziţie. Secţiunea 1 discută metoda noastră de antrenare a unor reprezentări vectoriale bilingve pentru cuvinte (*bilingual word embeddings*) pentru a traduce nume de entităţi (*named entities*), în timp ce secţiunea 2 discută arhitectura noastră de reţea neuronală pentru a determina dacă două tweet-uri (texte scurte) spun lucruri similare sau contradictorii. În secţiunea 2, analizăm, de asemenea, efectul dimensiunii şi distribuţiei datelor şi a diferitelor metrici de similaritate semantică asupra performanţei clasificatorului.

Capitolul 4 discută alternanţa la nivel de text. Secţiunea 1 se concentrează pe rezultatele noastre în clasificarea temporală a textului pentru limba română, în timp ce Secţiunea 2 se aruncă profund în activitatea noastră în clasificarea textului legal. În această a doua secţiune, prezentăm atât un model de clasificare temporală, cât şi un model capabil să prezică aria juridică, precum şi decizia finală din descrierea unui caz. De asemenea, discutăm mecanismul nostru pentru reducerea dimensionalităţii spaţiului etichetelor.

Capitolul 5 analizează alternanţele la nivel de autor şi prezintă lucrările noastre în Atribuirea şi Profilarea Autorului. Secţiunea 1, în mod special, prezintă experimentele noastre de detectare a pastişei, în timp ce Secţiunea 2 discută despre detectarea genului, vârstei şi personalităţii unui autor.

Capitolul 6 analizează alternanţa dintre şi dincolo de limbi. Prima secţiune prezintă clasificatorul nostru capabil să distingă între cele trei varietăţi ale limbii portugheze regăsite în textele jurnalistice. A doua secţiune analizează sistemul nostru pentru identificarea adreselor URL ce se îndreptă spre site-uri web dăunătoare utilizatorilor de Internet. În cele din urmă, a treia secţiune arată cum antrenarea unui model de limbaj neuronal (*neural language model*) pe un set de date care conţine doar adrese URL duce la generarea precisă de adrese URL noi.

În capitolul 7, tragem concluziile, unificăm şi distilăm rezultatele noastre şi argumentăm în favoarea înţelegerii alternanţei lingvistice ca o caracteristică universală a datelor generate de om, date care pot fi modelate statistic cu succes. Tot aici, anticipând lucrările viitoare, propunem o formă modificată a testului Turing, pe care îl denumim: *Testul Imortalităţii*.

4. Lista de Publicaţii

1. Octavia-Maria Şulea, Steve Young (2019), Unsupervised inflection generation using neural language modelling, in I. Rojas, G. Joya and A. Catala, eds, 'Advances in Computational Intelligence - 15th International Work-Conference on Artificial Neural Networks, IWANN 2019, Proceedings, Part I', Gran Canaria, Spain, June 12-14, 2019, Vol. 11506 of Lecture Notes in Computer Science, Springer, pp. 668–678.
2. Octavia-Maria Şulea, Steve Young, Liviu P. Dinu (2019), MorphoGen: Full Inflection Generation Using Recurrent Neural Networks. '20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)', La Rochelle, France, April 2019 (to appear)
3. Octavia-Maria Şulea, Liviu P. Dinu, Bogdan Dumitru (2018), Full Inflection Learning Using Deep Neural Networks. '19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2018)'. Hanoi, Vietnam, March 2018 (to appear)
4. Octavia-Maria Şulea (2017), Recognizing textual entailment in Twitter using word embeddings, in S. R. Bowman, Y. Goldberg, F. Hill, A. Lazaridou, O. Levy, R. Reichart and A. Søgaard, eds, 'Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP, RepEval@EMNLP 2017', Copenhagen, Denmark, September 8, 2017, Association for Computational Linguistics, pp. 31–35.
5. Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, Josef van Genabith (2017), Predicting the Law Area and Decisions of French Supreme Court Cases. In Mitkov, R. and Angelova, G., eds, 'Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017', Varna, Bulgaria, September 2 - 8, 2017, INCOMA Ltd., pp. 716-722
6. Octavia-Maria Şulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu, Josef van Genabith (2017), Exploring the use of text classification in the legal domain, in K. D. Ashley, K. Atkinson, L. K. Branting, E. Francesconi, M. Grabmair, M. Lauritsen, V. R. Walker and A. Z. Wyner, eds, 'Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal

Texts co-located with the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)', London, UK, June 16, 2017, Vol. 2143 of CEUR Workshop Proceedings, CEUR-WS.org.

7. Octavia-Maria Şulea (2016), Semi-supervised approach to Romanian noun declension, in R. J. Howlett, L. C. Jain, B. Gabrys, C. Toro and C. P. Lim, eds, 'Knowledge-Based and Intelligent Information and Engineering Systems: Proceedings of the 20th International Conference KES-2016', York, UK, 5-7 September 2016, Vol. 96 of Procedia Computer Science, Elsevier, pp. 664–671.
8. Marcos Zampieri, Shervin Malmasi, Octavia-Maria Şulea, Liviu P. Dinu (2016), A Computational Approach to the Study of Portuguese Newspapers Published in Macau, in Larry Birnbaum, Octavian Popescu and Carlo Strapparava, eds, 'Proceedings of the Workshop on Natural Language Processing Meets Journalism (NLPMJ16) co-located with the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)', New York, United States, July 2016, pp. 47-52
9. Octavia-Maria Şulea, Sergiu Nisioi, Liviu P. Dinu (2016), Using word embeddings to translate named entities, in N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk and S. Piperidis, eds, 'Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016', Portorož, Slovenia, May 23-28, 2016, European Language Resources Association (ELRA), pp. 3362-3366
10. Octavia-Maria Şulea, Liviu P. Dinu, Alexandra Peşte (2015), Using NLP specific tools for non-nlp specific tasks. A web security application, in S. Arik, T. Huang, W. K. Lai and Q. Liu, eds, 'Neural Information Processing - 22nd International Conference, ICONIP 2015, Proceedings, Part IV', Istanbul, Turkey, November 9-12, 2015, Vol. 9492 of Lecture Notes in Computer Science, Springer, pp. 631–638.
11. Octavia-Maria Şulea, Daniel Dichiu (2015), Automatic Profiling of Twitter Users Based on Their Tweets: Notebook for PAN at CLEF 2015, in Cappellato, L., Ferro, N., Jones, G. J. F. and SanJuan, E., eds, 'Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum', Toulouse, France, September 8-11, 2015, Vol. 1391 of CEUR Workshop Proceedings, CEUR-WS.org.
12. Liviu P. Dinu, Vlad Niculae, Octavia-Maria Şulea (2013), Romanian syllabication using machine learning, in I. Habernal and V. Matousek, eds, 'Text, Speech, and Dialogue - 16th International Conference (TSD 2013), Pilsen, Czech Republic, September 1-5, 2013. Proceedings', Vol. 8082 of Lecture Notes in Computer Science, Springer, pp. 450–456.
13. Liviu P. Dinu, Octavia-Maria Şulea, Vlad Niculae (2013), Sequence tagging for

verb conjugation in Romanian, in Angelova, G., Bontcheva, K. and Mitkov, R., eds, 'Recent Advances in Natural Language Processing (RANLP 2013)', Hissar, Bulgaria, 9-11 September, 2013, INCOMA Ltd, pp. 215-220

14. Alina Maria Ciobanu, Liviu P. Dinu, Octavia-Maria Şulea, Anca Dinu, Vlad Niculae (2013), Temporal Text Classification for Romanian Novels set in the Past. in Angelova, G., Bontcheva, K. and Mitkov, R., eds, 'Recent Advances in Natural Language Processing, RANLP 2013', Hissar, Bulgaria, 9-11 September, 2013, INCOMA Ltd, pp. 136-140

15. Alina Maria Ciobanu, Anca Dinu, Liviu P. Dinu, Vlad Niculae, Octavia-Maria Şulea (2013), Temporal classification for historical Romanian texts, in P. Lendvai and K. Zervanou, eds, 'Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH@ACL 2013', Sofia, Bulgaria, August 8, 2013, The Association for Computer Linguistics, pp. 102–106.

16. Liviu P. Dinu, Vlad Niculae, Octavia-Maria Şulea (2012a), Dealing with the grey sheep of the Romanian gender system, the neuter, in M. Kay and C. Boitet, eds, 'COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Demonstration Papers', Mumbai, India, 8-15 December 2012, Indian Institute of Technology Bombay, pp. 119–124.

17. Iulia Dănăilă, Liviu P. Dinu, Vlad Niculae, Octavia-Maria Şulea (2012), String Distances for Near-duplicate Detection. Polibits, Research journal on Computer science and computer engineering with applications. June 2012, pp. 21-25

18. Liviu P. Dinu, Vlad Niculae, Octavia-Maria Şulea (2012b), The Romanian neuter examined through A two-gender n-gram classification system, in N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk and S. Piperidis, eds, 'Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012', Istanbul, Turkey, May 23-25, 2012, European Language Resources Association (ELRA), pp. 907–910.

19. Liviu P. Dinu, Vlad Niculae, Octavia-Maria Şulea (2012c). Pastiche detection based on stopword rankings: Exposing impersonators of a Romanian writer, in Eileen Fitzpatrick, Joan Bachenko and Tommaso Fornaciari, eds, 'Proceedings of the Workshop on Computational Approaches to Deception Detection, EACL 2012,' Avignon, France, April 2012, The Association for Computational Linguistics, pp. 72–77.

20. Liviu P. Dinu, Vlad Niculae, Octavia-Maria Şulea (2012d), Learning how to conjugate the Romanian verb. Rules for regular and partially irregular verbs, in W. Daelemans, M. Lapata and L. Marquez, eds, '13th Conference of the European

Chapter of the Association for Computational Linguistics (EACL 2012)', Avignon, France, April 23-27, 2012', The Association for Computer Linguistics, pp. 524–528.

21. Liviu P. Dinu, Emil Ionescu, Vlad Niculae, Octavia-Maria Şulea (2011), Can alternations be learned? A machine learning approach to Romanian verb conjugation, in G. Angelova, K. Bontcheva, R. Mitkov and N. Nicolov, eds, 'Recent Advances in Natural Language Processing, RANLP 2011', Hissar, Bulgaria, 12-14 September, 2011, INCOMA Ltd, pp. 539–544.

REFERINŢE

Angelova, G., Bontcheva, K. and Mitkov, R., eds (2013), *Recent Advances in Natural Language Processing, RANLP 2013, 9-11 September, 2013, Hissar, Bulgaria*, RANLP 2013 Organising Committee / ACL.

Borowiec, S. and Lien, T. (2016), 'Alphago beats human go champ in milestone for artificial intelligence'. Retrieved from http://www.latimes.com.

Bowman, S. R., Goldberg, Y., Hill, F., Lazaridou, A., Levy, O., Reichart, R. and Søgaard, A., eds (2017), *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP, RepEval@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, Association for Computational Linguistics.

URL: *http://aclanthology.info/volumes/proceedings-of-the-2nd-workshop-on-evaluating-vector-space*

Cappellato, L., Ferro, N., Jones, G. J. F. and SanJuan, E., eds (2015), *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*, Vol. 1391 of *CEUR Workshop Proceedings*, CEUR-WS.org.

URL: *http://ceur-ws.org/Vol-1391*

Ciobanu, A. M., Dinu, A., Dinu, L. P., Niculae, V. and Şulea, O. (2013), Temporal classification for historical romanian texts, *in* P. Lendvai and K. Zervanou, eds, 'Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH@ACL 2013, August 8, 2013, Sofia, Bulgaria', The Association for Computer Linguistics, pp. 102–106.

URL: *http://aclweb.org/anthology/W/W13/W13-2714.pdf*

Ciobanu, A. M. and Dinu, L. P. (2014), Automatic detection of cognates using orthographic alignment, *in* 'Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers', The Association for Computer Linguistics, pp. 99–105.

URL: *https://www.aclweb.org/anthology/P14-2017/*

Ciobanu, A. M., Dinu, L. P., Şulea, O., Dinu, A. and Niculae, V. (2013), Temporal text classification for romanian novels set in the past, *in* Angelova et al. (2013), pp. 136–140.

URL: *http://aclweb.org/anthology/R/R13/R13-1018.pdf*

Coulthard, M. (2004), 'Author identification, idiolect and linguistic uniqueness', *Applied Linguistics* **25**.

Davies, A. (2016), 'Uber's self-driving truck makes its first delivery: 50,000 beers', *Wired* . Retrieved from http://www.wired.com.

Deangelis, S. F. (2014), 'The growing importance of natural language processing', *Wired* . Retrieved from http://www.wired.com.

Deutsch, D. (2012), 'How close are we to creating artificial intelligence?', *Aeon* .

Dinu, L. P., Ionescu, E., Niculae, V. and Şulea, O. (2011), Can alternations be learned? A machine learning approach to romanian verb conjugation, *in* G. Angelova, K. Bontcheva, R. Mitkov and N. Nicolov, eds, 'Recent Advances in Natural Language Processing, RANLP 2011, 12-14 September, 2011, Hissar, Bulgaria', RANLP 2011 Organising Committee, pp. 539–544.

URL: *http://www.aclweb.org/anthology/R11-1075*

Dinu, L. P., Niculae, V. and Şulea, O. (2012*a*), Dealing with the grey sheep of the romanian gender system, the neuter, *in* M. Kay and C. Boitet, eds, 'COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Demonstration Papers, 8-15 December 2012, Mumbai, India', Indian Institute of Technology Bombay, pp. 119–124.

URL: *http://aclweb.org/anthology/C/C12/C12-3015.pdf*

Dinu, L. P., Niculae, V. and Şulea, O. (2012*b*), Learning how to conjugate the romanian verb. rules for regular and partially irregular verbs, *in* W. Daelemans, M. Lapata and L. Màrquez, eds, 'EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012', The Association for Computer Linguistics, pp. 524–528.

URL: *http://aclweb.org/anthology/E/E12/E12-1053.pdf*

Dinu, L. P., Niculae, V. and Şulea, O. (2012*c*), Pastiche detection based on stopword rankings: Exposing impersonators of a romanian writer, *in* 'Proceedings of the Workshop on Computational Approaches to Deception Detection', EACL 2012, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 72–77.

URL: *http://dl.acm.org/citation.cfm?id=2388616.2388627*

Dinu, L. P., Niculae, V. and Şulea, O. (2012*d*), The romanian neuter examined through A two-gender n-gram classification system, *in* N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk and S. Piperidis, eds, 'Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012', European Language Resources Association (ELRA), pp. 907–910.

URL: *http://www.lrec-conf.org/proceedings/lrec2012/summaries/651.html*

Dinu, L. P., Niculae, V. and Şulea, O. (2013), Romanian syllabication using machine learning, *in* I. Habernal and V. Matousek, eds, 'Text, Speech, and Dialogue - 16th International Conference, TSD 2013, Pilsen, Czech Republic, September 1-5, 2013. Proceedings', Vol. 8082 of *Lecture Notes in Computer Science*, Springer, pp. 450–456.

URL: *https://doi.org/10.1007/978-3-642-40585-3_57*

Dinu, L. P., Şulea, O. and Niculae, V. (2013), Sequence tagging for verb conjugation in

romanian, *in* Angelova et al. (2013), pp. 215–220.
URL: *http://aclweb.org/anthology/R/R13/R13-1028.pdf*
Graves, A., Wayne, G. and Danihelka, I. (2014), 'Neural turing machines', *CoRR* **abs/1410.5401**.
URL: *http://arxiv.org/abs/1410.5401*
Gross, R., Gu, Y., Li, W. and Gauci, M. (2017), Generalizing gans: A turing perspective, *in* 'Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA', pp. 6319–6329.
Jain, A. (2016), 'The 5 v's of big data', *IBM Watson Health Perspectives* . Retrieved from https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/.
Kamps, J., Monz, C., de Rijke, M. and Sigurbjörnsson, B. (2004), Language-dependent and language-independent approaches to cross-lingual text retrieval, *in* C. Peters, J. Gonzalo, M. Braschler and M. Kluck, eds, 'Comparative Evaluation of Multilingual Information Access Systems', Vol. 3237, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 152–165.
Kondrak, G., Marcu, D. and Knight, K. (2003), Cognates can improve statistical translation models, *in* 'Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–short Papers - Volume 2', NAACL-Short '03, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 46–48.
URL: *https://doi.org/10.3115/1073483.1073499*
Merchant, B. (2015), 'The poem that passed the turing test', *Motherboard* . Retrieved from http://www.vice.com.
Metz, C. (2015), 'AI's next frontier: Machines that understand language', *Wired* . Retrieved from http://www.wired.com.
Mitkov, R. and Angelova, G., eds (2017), *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, INCOMA Ltd.
URL: *https://aclanthology.info/volumes/proceedings-of-the-international-conference-recent-advances*
Nieva, R. (2018), 'Alphabet chairman says google duplex passes turing test in one specific way'. Retrieved from http://www.cnet.com.
Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (n.d.), 'Language models are unsupervised multitask learners'.
Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., Larochelle, H. and Zemel, R. S. (2018), Meta-learning for semi-supervised few-shot classification, *in* '6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings', OpenReview.net.

URL: *https://openreview.net/forum?id=HJcSzz-CZ*

Schmelzer, R. (2018), 'Google duplex and the problem with the turing test'. Retrieved from https://ctovision.com.

Şulea, O. (2016), Semi-supervised approach to romanian noun declension, *in* R. J. Howlett, L. C. Jain, B. Gabrys, C. Toro and C. P. Lim, eds, 'Knowledge-Based and Intelligent Information and Engineering Systems: Proceedings of the 20th International Conference KES-2016, York, UK, 5-7 September 2016.', Vol. 96 of *Procedia Computer Science*, Elsevier, pp. 664–671.

URL: *https://doi.org/10.1016/j.procs.2016.08.248*

Şulea, O. (2017), Recognizing textual entailment in twitter using word embeddings, *in* Bowman et al. (2017), pp. 31–35.

URL: *https://aclanthology.info/papers/W17-5306/w17-5306*

Şulea, O. and Dichiu, D. (2015), Automatic profiling of twitter users based on their tweets: Notebook for PAN at CLEF 2015, *in* Cappellato et al. (2015).

URL: *http://ceur-ws.org/Vol-1391/48-CR.pdf*

Şulea, O., Dinu, L. P. and Dumitru, B. (2018), 'Full inflection learning using deep neural networks'. 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2018).

Şulea, O., Dinu, L. P. and Peşte, A. (2015), Using NLP specific tools for non-nlp specific tasks. A web security application, *in* S. Arik, T. Huang, W. K. Lai and Q. Liu, eds, 'Neural Information Processing - 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings, Part IV', Vol. 9492 of *Lecture Notes in Computer Science*, Springer, pp. 631–638.

Şulea, O., Nisioi, S. and Dinu, L. P. (2016), Using word embeddings to translate named entities, *in* N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk and S. Piperidis, eds, 'Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portoroz, Slovenia, May 23-28, 2016.', European Language Resources Association (ELRA).

URL: *http://www.lrec-conf.org/proceedings/lrec2016/summaries/1167.html*

Şulea, O. and Young, S. (2019), Unsupervised inflection generation using neural language modelling, *in* I. Rojas, G. Joya and A. Català, eds, 'Advances in Computational Intelligence - 15th International Work-Conference on Artificial Neural Networks, IWANN 2019, Gran Canaria, Spain, June 12-14, 2019, Proceedings, Part I', Vol. 11506 of *Lecture Notes in Computer Science*, Springer, pp. 668–678.

URL: *https://doi.org/10.1007/978-3-030-20521-8_55*

Şulea, O., Young, S. and Dinu, L. P. (2019), 'Morphogen: Full inflection generation using recurrent neural networks'. 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019).

Şulea, O., Zampieri, M., Malmasi, S., Vela, M., Dinu, L. P. and van Genabith, J. (2017), Exploring the use of text classification in the legal domain, *in* K. D. Ashley, K. Atkinson, L. K. Branting, E. Francesconi, M. Grabmair, M. Lauritsen, V. R. Walker and A. Z. Wyner, eds, 'Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Texts co-located with the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017), London, UK, June 16, 2017.', Vol. 2143 of *CEUR Workshop Proceedings*, CEUR-WS.org.
URL: *http://ceur-ws.org/Vol-2143/paper5.pdf*

Şulea, O., Zampieri, M., Vela, M. and van Genabith, J. (2017), Predicting the law area and decisions of french supreme court cases, *in* Mitkov and Angelova (2017), pp. 716–722.
URL: *https://doi.org/10.26615/978-954-452-049-6_092*

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C. and Liu, C. (2018), A survey on deep transfer learning, *in* V. Kurková, Y. Manolopoulos, B. Hammer, L. S. Iliadis and I. Maglogiannis, eds, 'Artificial Neural Networks and Machine Learning - ICANN 2018 - 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III', Vol. 11141 of *Lecture Notes in Computer Science*, Springer, pp. 270–279.
URL: *https://doi.org/10.1007/978-3-030-01424-7_27*

Turing, A. M. (1950), 'Computing machinery and intelligence', *Mind* **59**(236), 433–460.

Wang, K., Hua, H. and Wan, X. (2019), 'Controllable unsupervised text attribute transfer via editing entangled latent representation', *CoRR* **abs/1905.12926**.
URL: *http://arxiv.org/abs/1905.12926*

Weiner, K. (2018), 'Can AI create true art?', *Scientific American* . Retrieved from http://www.blogs.scientificamerican.com.

Weiss, K. R., Khoshgoftaar, T. M. and Wang, D. (2016), 'A survey of transfer learning', *J. Big Data* **3**, 9.
URL: *https://doi.org/10.1186/s40537-016-0043-6*

Wright, D. (2014), Stylistics versus Statistics: A corpus linguistic approach to combining techniques in forensic authorship analysis using Enron emails, PhD thesis, The University of Leeds.

Zampieri, M., Malmasi, S., Nakov, P., Ali, A., Shon, S., Glass, J. R., Scherrer, Y., Samardzic, T., Ljubesic, N., Tiedemann, J., van der Lee, C., Grondelaers, S., Oostdijk, N., Speelman, D., van den Bosch, A., Kumar, R., Lahiri, B. and Jain, M. (2018), Language identification and morphosyntactic tagging: The second vardial evaluation campaign, *in* M. Zampieri, P. Nakov, N. Ljubesic, J. Tiedemann, S. Malmasi and A. Ali, eds, 'Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial@COLING 2018, Santa Fe, New Mexico, USA, August 20, 2018',

Association for Computational Linguistics, pp. 1–17.
URL: *https://aclanthology.info/papers/W18-3901/w18-3901*
Zampieri, M., Malmasi, S., Şulea, O.-M. and Dinu, L. P. (2016), A computational approach to the study of portuguese newspapers published in macau, *in* 'Proceedings of the Workshop on Natural Language Processing Meets Journalism (NLPMJ)'.
Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F. and Choi, Y. (2019), 'Defending against neural fake news', *CoRR* **abs/1905.12616**.
URL: *http://arxiv.org/abs/1905.12616*