University of Bucharest
Faculty of Mathematics and Computer Science
Doctoral School of Computer Science

Doctoral Thesis

Computational Aspects in Natural Language Processing

Aspecte Computaţionale în Procesarea Limbajului Natural

Summary

Author: Octavia Maria Şulea
Scientific Advisor: Prof. dr. Liviu P. Dinu

2019

TABLE OF CONTENTS

1. Introduction

Artificial Intelligence (AI) as a research field within Computer Science is entering its 63rd year at the time of completing this thesis, with the famous Turing Test (Turing 1950) nearing its 70th year and with that its argued obsolescence (Schmelzer 2018). As media becomes more and more aware of machine generated output passing as or surpassing humans in more and more contexts (Merchant (2015), Davies (2016), Borowiec and Lien (2016), Nieva (2018), Weiner (2018)), computer generated content is becoming harder to detect even with the help of machines, as evidenced by top AI conferences like the Thirty-sixth International Conference on Machine Learning (ICML 2019) holding workshops for their detection [1], important workshops like PAN creating shared tasks to distinguish between machine and human generated content [2], and researchers becoming increasingly concerned about releasing their trained neural language models as they could be used maliciously, for propaganda (Radford et al. n.d., Zellers et al. 2019).

Initially a subfield of AI, Machine Learning (ML) slowly detached from the original goal of imitating human intelligence and turned to solving specific tasks with data driven techniques. It soon became apparent to researchers within ML, informed by various theories of mind, that learning from past data a generalizing hypothesis and then making predictions on future data based on this hypothesis – the main purpose of a supervised ML model – wasn't the only ingredient to human intelligence so the ML field matured into learning as a way of solving specific tasks, rather than seeking an Artificial General Intelligence (AGI) (Deutsch 2012). In the last few years, however, ML research has come full circle with papers such as Graves et al. (2014) incorporating the Turing Machine into Recurrent Neural Networks (RNNs) and Gross et al. (2017) comparing Generative Adversarial Networks (GANs) to the Turing Test itself, and the community at large shifting from supervised learning over feature engineered models for specific tasks to semi-supervised and fully unsupervised Deep Neural Network (DNN) models capable of transferring learned skills from one task to the other (Weiss et al. 2016, Tan et al. 2018) or generalizing from a few examples (Ren et al. 2018), thus drawing nearer to the seeming ultimate goal of AGI.

One research area within Computer Science that intersects with Machine Learning but is considered as old as AI, due to the Turing Test's requirement that the human-machine interaction be done via text, is Natural Language Processing (NLP). The goal within NLP, confused in recent years with the general AI goal of imitating human intelligence (Metz 2015), is to make the communication between computers and typical human users, as opposed to expert users (i.e. programmers), more *natural* (Deangelis

[1] Synthetic Realities: Deep Learning for Detecting AudioVisual Fakes: https://sites.google.com/view/audiovisualfakes-icml2019/

[2] PAN @ CLEF 2019's bots and gender profiling task: `https://pan.webis.de/clef19/pan19-web/author-profiling.html`

2014). Specifically, this means that between the computer, operated through *formal languages* (i.e. programming languages) and the human individual, cooperating with other individuals through *natural language*, NLP strives to find a happy medium. When we assume this to be the over-arching goal of NLP and observe it in the context of the over-abundance of textual data on the Internet, it becomes easier to understand the definition and solution of many tasks and subfields within NLP over the decades: Machine Translation, Information Extraction, Information Retrieval, Question Answering, Speech to Text, Text Classification, etc. One issue has remained constant within NLP as it is part of a larger issue within statistical ML: representing linguistic information digitally (Bowman et al. 2017). The complexity of this issue reveals itself when we look at tasks or languages with high variability in the data, with sparsity exhibited in the feature (i.e. input) or class (i.e. output) space, as data is abundant but not in the ways we would hope it to be (Jain 2016).

Modeling unit level variation (alternation) in NLP has been a great issue. In languages with rich inflectional morphology, NLP has seen tasks like *stemming* or *lemmatisation* attempt to algorithmically reduce the inflected word (e.g. learns) to its uninflected form (e.g. learn) to improve *term-frequency inverse document frequency* (TF-IDF) methods of representing text in Information Retrieval tasks (Kamps et al. 2004), or to shrink the size of the vocabulary and thusly reduce the sparsity of the feature space in *bag-of-words* models, used in Text Classification. In diachronic linguistics and language similarity, the change of the same word through time or from one language to another are important linguistic phenomena that need to be modeled accurately for downstream tasks within NLP areas like Machine Translation (Kondrak et al. 2003, Ciobanu and Dinu 2014). In stylistics and forensic or social linguistics, being able to model variation within the style of one author (i.e. idiolect) or one population group (i.e. language variety or dialect) is essential in Author Profiling or Deception Detection tasks (Coulthard (2004), Wright (2014), Zampieri et al. (2018)). Even more, in Natural Language Understanding (NLU) or Inference (NLI) tasks, researchers seek to model alternation of meaning at the sentence or discourse level and disentangle it from the style in which it is delivered (Wang et al. 2019).

The over-arching topic of this thesis is that of linguistic alternation. We look at computational ways to model alternations in the context of linguistic identity and present our works in Computational Morphology, Machine Translation, Information Extraction, Author Profiling and Attribution, Natural Language Inference, and Text Classification.

2. Contributions

In this thesis, we present and expand upon our published work. The complete list of 21 publications can be found in the List of Publications. In what follows, we summarize our main contributions.

In Computational Morphology, we proposed a *weakly supervised learning* method for predicting the inflection of nouns (Şulea 2016) and verbs (Dinu et al. 2011, 2012*b*) in Romanian, a morphologically rich language. We introduced a novel way of modeling inflection patterns using fixed sets of regular expressions with which we labeled our noun and verb datasets and trained linear Support Vector Classifiers (SVC) on character level n-grams extracted from the uninflected forms to predict their inflection classes. We showed that the character n-grams, especially when emphasizing the ending of the uninflected word by appending an artificial ending marker, are sufficient features to accurately predict the verbal class and slightly less powerful in predicting the nominal class for Romanian. We posit that the issue lies in a richer inflection for partially-irregular Romanian nouns. We show that, while we can predict the inflection class from the uninflected form, the reverse does not hold. Namely, we cannot generate the inflected forms accurately if we start off from the inflection class rules and the infinitive.

We also defined morphological inflection as a *sequence labeling task* and considered each character within the uninflected form as a *label* or *type of letter*: the graphical character *t* in the verb *a cânta* (to sing) now becomes the label t_0 (Dinu, Şulea and Niculae 2013). We used Conditional Random Fields (CRF) to predict the label for each character in the infinitive of Romanian verbs and showed that this leads to improvements in the over-arching inflection learning problem, but does not alleviate the need for hand-crafted labeling systems, an issue also with the weakly supervised method.

We explored the influence of lexical and phonological ambiguity on inflection by presenting our gender classification system for Romanian nouns (Dinu et al. 2012*d*,*a*) and our sequence labeling system for Romanian syllabication (Dinu, Niculae and Şulea 2013). We then show the applicability of a sequence to sequence (Seq2Seq) deep learning model to the task of morphological inflection learning and syllabication (Şulea et al. 2018), improving upon the performance of our previous inflection learning models and requiring less human supervision.

Finally, we proposed an unsupervised method to generate the full inflection of verbs in Romanian, Spanish and Finnish (Şulea et al. 2019) and nouns in Romanian, German and Finnish (Şulea and Young 2019) using *neural language modeling*.

At the phrasal level, we proposed a way to train bilingual word embeddings such that the representation of the same Named Entities in different languages stays similar (Şulea et al. 2016), this leading to improvements in Name Entity Translation and circumventing the need to build Named Entity Recognition systems for multiple languages

separately. We also looked at entailment of short text coming from Twitter and showed that, even though the topic stays the same, language shift on Twitter occurs too quickly for entailment models to remain robust over time (Şulea 2017).

Also related to language change, we introduced a reliable method for temporal text classification of Romanian literary texts (Ciobanu, Dinu, Şulea, Dinu and Niculae 2013, Ciobanu, Dinu, Dinu, Niculae and Şulea 2013) and showed that the same strategies are not as robust for French text coming from the legal domain (Şulea, Zampieri, Malmasi, Vela, Dinu and van Genabith 2017, Şulea, Zampieri, Vela and van Genabith 2017). Within the legal domain, we also present our classifier capable of predicting the French Supreme Court's ruling in a case, given its description. For this purpose, we also present our technique for reducing the label space by using hierarchical clustering and choosing a higher level cut-off in the dendogram.

At the author level, we introduced a model to distinguish between literary texts from one author and pastiches (spoofs) made by others (Dinu et al. 2012*c*). We also present our models to accurately predict demographic characteristics of microblog users based on their text output (Şulea and Dichiu 2015).

At the language level and beyond, we introduce a method to distinguish between different varieties of Portuguese (Zampieri et al. 2016) and we also introduce a method to identify URLs pointing to websites that contain malicious content (Şulea et al. 2015), showing that URLs can also very accurately be described based on punctuation and similar text classification techniques can be applied. We also used the same technique as in Chapter 2 for neural language modeling to accurately generate novel, valid URLs.

Finally, we propose a modification to the Turing Test which we re-dub the *Immortality Test*. In essence, this modified form of the classic test of intelligence for an artificial agent (i.e. computer) asks the validator (i.e. human) to switch from the original question of *man vs. machine* and ask instead: *myself or the machine?* This way, we update the test to account for the latest results in NLP of over-personalization and of training neural language models on individual user data, but, most importantly, this modified test for AI is designed to take into account the solipsistic view of human nature where the only thing conscious validators can validate is themselves.

3. Thesis Outline

In the following, we present the outline of this thesis.

In Chapter 2, we discuss all our published work in the area of Computational Morphology. In the first section, we explain the concept of morphologically rich languages from the perspective of inflection. We explain the linguistic phenomena of *apophony* and *stem allomorphy*. We exemplify these concepts on Romanian, German, and Span-

ish. We also discuss *gender class* and *diphthong-hiatus* ambiguity for Romanian.

In Section 2, we present two strategies for inflection learning: a weakly-supervised model and a sequence labelling model. For the first, we present the inflection rules created in order to label Romanian nouns and verbs and detail how we trained SVMs to learn the association between the dictionary form of a word and its inflection rule using only character n-grams of the uninflected form. We show that the reverse process, meaning to generate the inflected forms from the inflection rule, is not feasible. For the second strategy, considering inflection learning as a sequence labeling task, we detail how we modeled each character within the dictionary form of a Romanian verb as a label representing its alternation pattern and how we employed CRFs to predict these labels. Section 3 discusses our work in noun gender classification and syllabication for Romanian, revealing the indirect influence gender and syllable structure has on inflection.

Section 4 of Chapter 2 presents our work in modeling inflection as a sequence to sequence task and using a Recurrent neural Network (RNN) architecture. Finally, section 5 presents our efforts in generating all the inflected forms using no supervision from the uninflected form alone. We show that this can be achieved for Romanian, Spanish and Finnish verbs and Romanian nouns quite easily using an RNN-based language model with an attention mechanism, and that the same is harder to achieve for German and Finnish nouns. We also discuss our results from experimenting with tabula rasa and pretrained language models.

Chapter 3 presents our works in modeling alternation at the phrase and sentence level. Section 1 discusses our method of training bilingual word embeddings to translate Named Entities, while section 2 discusses our neural network to determine whether two tweets (short texts) are saying semantically similar or contradicting things. In Section 2, we also look at the effect of data size and data distribution on the classifier's performance.

Chapter 4 discusses text level alternation. Section 1 focuses on our results in temporal text classification for Romanian, while Section 2 dives deep into our work in legal text classification. In this section, we present both a temporal classification model and a model capable of predicting the legal area as well as the final decision from the description of a case. We also discuss our mechanism for dimensionality reduction of the label space.

Chapter 5 looks at author level alternations and presents our work in Authorship Attribution and Author Profiling. Section 1, specifically, presents our pastiche detection experiments, while Section 2 discusses our gender, age and personality detection.

Chapter 6 looks at alternation between and beyond languages. The first section presents our classifier capable of distinguishing between the three varieties of Portuguese in journalist texts. The second section looks at our system for identifying URLs

pointing to malicious websites. Finally, the third section shows how training a neural language model on a dataset containing URLs leads to accurate URL generation.

Finally, in Chapter 7 we draw our conclusions, unify and distill our results, and argue for the understanding of linguistic alternation as a universal characteristic of human generated data which can be successfully modeled. Here, we also introduce our modification to the Turing Test, which we dub the *Immortality Test*.

4. List Of Publications

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.

REFERENCES

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-representations-for-nlp*

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-in-natural-language-processing-ranlp-2017*

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 cam-

paign, *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*