MINISTERUL EDUCAŢIEI ȘI CERCETĂRII AL REPUBLICII MOLDOVA

Universitatea Tehnică a Moldovei
Facultatea Calculatoare, Informatică și Microelectronică
Departamentul Informatică și Ingineria Sistemelor

Admis la susținere
Șef departament:
Sudacevschi Viorica, conf. univ., dr.
_____
„____” _____ 2023

Instrumente de parafrazare a textului utilizând metodele PLN

Teză de master

Student: **Petrachi Alexandru, MAI-211M**

Conducător: **Moraru Vasile, prof. univ., dr.**

Chișinău, 2023

ADNOTARE

Cuvinte cheie: Procesarea limbajului natural (PLN), Parafrazarea, Reglaj fin (fine-tune), Similaritate semantică, Traducere, Pre-antrenare, Gramatică, Set de date (dataset), Învățare automată, Invatare profunda, Rețele neuronale, Sinonime, Traducere inversă, Conductă (Pipeline), Evaluarea umană, Comunitatea PLN.

Prin intermediul tezei de master ne-am propus să creăm și să evaluăme diverse metode de parafrazare în limba română. Proiectul a început prin ajustarea fină a unui model de similaritate semantică și a modelelor de traducere (EN-RO, RO-EN și FI-RO). A urmat reglarea fină a unui model gramatical și a unui model de parafrazare. A fost creat un set de date pentru pre-training și a fost actualizat constant, ajungând la peste patru milioane de înregistrări.

În ceea ce privește rezultatele metodelor de parafrazare, am testat mai multe abordări, inclusiv înlocuirea sinonimelor, traducerea inversa și diverse modele de parafrazare precum Flan-t5-small, Flan-t5-base, Flan-t5-large și pipeline. Am evaluat aceste metode folosind o abordare de evaluare umană, care a arătat că metoda pipeline a avut cea mai bună performanță.

În plus, această lucrare are o valoare semnificativă pentru comunitatea PLN, deoarece oferă acces la mai multe modele ajustate care pot fi utilizate pentru diverse aplicații, cum ar fi clasificarea textului, analiza sentimentelor, traducerea automată și multe altele. În plus, munca noastră evidențiază importanța creării de seturi de date mari și diverse pentru modelele de pre-instruire, precum și necesitatea evaluării umane pentru a evalua cu precizie performanța acestor modele.

Pe scurt, această lucrare a contribuit la avansarea procesării limbajului natural în limba română și poate servi drept resursă valoroasă pentru cercetătorii și practicienii din acest domeniu. Sperăm că munca noastră va inspira cercetări și dezvoltare în continuare în acest domeniu și, în cele din urmă, va conduce la aplicații de procesare a limbajului natural mai precise și mai eficiente pentru limba română.

ANNOTATION

Keywords: Natural Language Processing (NLP), Paraphrasing, Fine-tune, Semantic Similarity, Translation, Pre-training, Grammar, Dataset, Machine Learning, Deep Learning, Neural Networks, Synonyms, Translation reverse, Pipeline, Human evaluation, PLN Community.

Through the master's thesis, we set out to create and evaluate various methods of paraphrasing in Romanian. The project started by fine-tuning a semantic similarity model and translation models (EN-RO, RO-EN and FI-RO). Fine-tuning a grammar model and a paraphrasing model followed. A pre-training dataset was created and constantly updated, reaching over four million records.

Regarding the results of paraphrasing methods, we tested several approaches, including synonym replacement, reverse translation, and various paraphrasing models such as Flan-t5-small, Flan-t5-base, Flan-t5-large, and pipeline. We evaluated these methods using a human evaluation approach, which showed that the pipeline method performed best.

Furthermore, this work is of significant value to the PLN community as it provides access to several tuned models that can be used for various applications such as text classification, sentiment analysis, machine translation, and more. Furthermore, our work highlights the importance of creating large and diverse datasets for pre-training models, as well as the need for human evaluation to accurately assess the performance of these models.

In short, this work has contributed to the advancement of natural language processing in Romanian and can serve as a valuable resource for researchers and practitioners in this field. We hope that our work will inspire further research and development in this area and ultimately lead to more accurate and efficient natural language processing applications for the Romanian language.

CUPRINS

INTRODUCERE

În ultimii ani, procesarea limbajului natural (PLN) a devenit unul dintre cele mai dinamice domenii din informatică, cu o gamă largă de aplicații, de la chatbot și asistenți virtuali până la traducerea automată și analiza sentimentelor. Odată cu explozia datelor digitale, PLN a câștigat din ce în ce mai multă atenție, iar cercetătorii s-au străduit să dezvolte noi modele și tehnici care pot îmbunătăți acuratețea și eficiența diferitelor sarcini PLN.

Această teză de master se concentrează pe dezvoltarea și evaluarea metodelor de parafrazare a textului în limba română, o sarcină importantă și provocatoare în PLN. Parafrazarea este procesul de reformulare a unei propoziții sau a unei fraze, păstrând în același timp sensul acesteia, dar cu o formulare, structură și stil diferit. Parafrazarea are multe aplicații practice, de la simplificarea și rezumarea textului până la creșterea datelor și transferul de stil. Cu toate acestea, parafrazarea este, de asemenea, o sarcină complexă, deoarece necesită nu numai cunoștințe lingvistice, ci și creativitate și conștientizarea contextului.

Valoarea acestei lucrări constă în mai multe aspecte. În primul rând, această lucrare contribuie la îmbunătățirea modelelor de ultimă generație pentru parafrazarea în limba română, care este un limbaj cu resurse insuficiente în PLN. În al doilea rând, această lucrare propune o abordare nouă a parafrazării care combină mai multe tehnici, cum ar fi înlocuirea sinonimelor, traducerea inversă și modelele de parafrazare neuronală, pentru a obține rezultate mai bune. În al treilea rând, această lucrare prezintă o evaluare amănunțită a evaluarii umane, care oferă informații valoroase asupra punctelor forte și slabe ale metodelor.

Restul acestei teze este organizat după cum urmează. Capitolul 1 oferă o privire de ansamblu asupra lucrărilor conexe în parafrazare și PLN pentru limba română. Capitolul 2 descrie setul de date și metodologia utilizată pentru reglarea fină și evaluarea modelelor de parafrazare. Capitolul 3 prezintă dezvoltarea unei aplicații de parafrazare folosind metodele propuse, rezultatele experimentelor și evaluarea metodelor.

Bibliografia

1. DE BEAUGRANDE, R. and DRESSLER. W. V. *Introduction to Text Linguistics.* Longman. New York, NY, 1981.
2. HIRST, G. *Paraphrasing paraphrased*. Discurs invitat la ACL International Workshop on Parafrasing. Sapporo, 2003.
3. MEL'CUK, I. *Semantics: From Meaning to Text*. John Benjamins Publishing Co. Philadelphia, PA, 2012.
4. CLARK, E. V. *Conventionality and contrasts: Pragmatic principles with lexical consequences*. În Andrienne Lehrer și Eva Feder Kittay, New Essays in Semantic Lexical Organization. Lawrence Erlbaum Associates, Hillsdale, NJ, 1992.
5. HARRIS, Z. *Co-occurrence and transformation in linguistic structure*. In Henry Hiz. Reidel Publishing Co. Dordrecht, 1981. paginile 143–210.
6. HONECK, R. P. *A study of paraphrases*. Journal of Verbal Learning and Verbal Behavior, 1971. Disponibil: doi.org/10.1016/S0022-5371(71)80035-X.
7. VAN ROSSUM, G., DRAKE JR, F. L. *The python language reference*. Python Software Foundation: Wilmington, DE, USA, 2014.
8. JUN, L., LING, L. 2010. *Comparative research on Python speed optimization strategies.* În 2010, Conferința Internațională privind calculul inteligent și sistemele integrate, 2010. paginile 57-59. Disponibil: 10.1109/ICISS.2010.5655011.
9. WILBERS, I. M., LANGTANGEN, H. P., ØDEGÅRD, Å. *Using cython to speed up numerical python programs*. Proceedings of MekIT, 2009. paginile 495-512. Disponibil: https://www.researchgate.net/publication/51994561_Using_Cython_to_Speed_Up_Numerical_Python_Programs.
10. LOPER, E., BIRD, S. *Nltk: The natural language toolkit*. 2002. Disponibil: doi.org/10.3115/1118108.1118117.
11. VASILIEV, Y. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press, 2020. ISBN: 9785446115068.
12. MCMILLAN-MAJOR, A., OSEI, S., RODRIGUEZ, J. D., AMMANAMANCHI, P. S., GEHRMANN, S., JERNITE, Y. *Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards*. 2021. Disponibil: doi.org/10.48550/arXiv.2108.07374.
13. KHYANI, D., SIDDHARTHA, B. S., NIVEDITHA, N. M., DIVYA, B. M. *An interpretation of lemmatization and stemming in natural language processing*. Jurnalul

Universității din Shanghai pentru Știință și Tehnologie, 2021. paginile 350-357. Disponibil: https://www.researchgate.net/publication/348306833_An_Interpretation_of_Lemmatization_and_Stemming_in_Natural_Language_Processing.

14. MORATO, J., MARZAL, M. A., LLORÉNS, J., MOREIRO, J. *Wordnet applications*. În Proceedings of GWC, 2004. paginile 20-23. Disponibil: https://www.researchgate.net/publication/220046020_WordNet_Applications.
15. DUMITRESCU, S. D., AVRAM, A. M., MOROGAN, L., TOMA, S. A. *Rowordnet–a python api for the romanian wordnet.* În a 10-a Conferință Internațională pentru Electronică, Calculatoare și Inteligență Artificială (ECIA), 2018. paginile 1-6. Disponibil: doi.org/10.1109/ECAI.2018.8679089.
16. COLIN Raffel, NOAM Shazeer, ADAM Roberts, KATHERINE Lee, SHARAN Narang, MICHAEL Matena, YANQI Zhou, WEI Li, PETER J. Liu. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* Journal of Machine Learning Research 21, 2019. Disponibil: doi.org/10.48550/arXiv.1910.10683.
17. HYUNG Won Chung, LE Hou, SHAYNE Longpre, BARRET Zoph, YI Tay, WILLIAM Fedus, YUNXUAN Li, XUIZHI Wang, MOSTAFA Dehghani, SIDDHATHA Brahma, ALBERT Webson, SHIXIANG Shane Gu, ZHUYUN Dai, MIRAC Suzgun, XINYUN Chen, AAKANKSHA Chowdhery, ALEX Castro-Ros, MARIE Pellat, KEVIN Robinson, DASHA Valter, SHARAN Narang, GAURAV Mishra, ADAMS Yu, VINCENT Zhao, YANPING Huang, ANDREW Dai, HONGKUN Yu, SLAV Petrov, ED H. Chi, JEFF Dean, JACOB Devlin, ADAM Roberts, DENNY Zhou, QUOC V. Le, JASON Wei. *Scaling Instruction-Finetuned Language Models.* Journal of Machine Learning Research 24, 2022. Disponibil: doi.org/10.48550/arXiv.2210.11416.
18. JACOB Devlin, MING-WEI Chang, KENTON Lee, KRISTINA Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. Disponibil: doi.org/10.48550/arXiv.1810.04805.
19. ASHISH Vaswani, NOAM Shazeer, NIKI Parmar, JAKOB Uszkoreit, LLION Jones, AIDAN N. Gomez, LUKASZ Kaiser, ILIA Polosukhin. *Attention is All you Need*. 2017. Disponibil: doi.org/10.48550/arXiv.1706.03762.
20. DANIEL Cer, MONA Diab, ENEKO Agirre, IIGO LopezGazpio, LUCIA Specia. *SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation*. 2017. Disponibil: doi.org/10.18653/v1/S17-2001.
21. YINHAN Liu, MYLE Ott, NAMAN Goyal, JINGFEI Du, MANDAR Joshi, DANQI Chen, OMER Levy, MIKE Lewis, LUKE Zettlemoyer, VESELIN Stoyanov. *RoBERTa:*

A Robustly Optimized BERT Pretraining Approach. 2019. Disponibil: doi.org/10.48550/arXiv.1907.11692.

22. YINFEI Yang, STEVE Yuan, DANIEL Cer, SHENG-YI Kong, NOAH Constant, PETR Pilar, HEMING Ge, YUN-HSUAN Sung, BRIAN Strope, RAY Kurzweil. *Learning Semantic Textual Similarity from Conversations.* 2018. Disponibil: doi.org/10.18653/v1/W18-3022.
23. CHANDLER May, ALEX Wang, SHIKHA Bordia, SAMUEL R. Bowman, RACHEL Rudinger. *On Measuring Social Biases in Sentence Encoders*. 2019. Disponibil: doi.org/10.48550/arXiv.1903.10561.
24. TIANYI Zhang, VARSHA Kishore, FELIX Wu, KILIAN Q. Weinberger, YOAV Artzi. *BERTScore: Evaluating Text Generation with BERT*. 2019. Disponibil: doi.org/10.48550/arXiv.1904.09675.
25. YIFAN Qiao, Chenyan XIONG, ZHENG-HAO Liu, ZHIYUAN Liu. Understanding the Behaviors of BERT in Ranking. 2019. Disponibil: doi.org/10.48550/arXiv.1904.07531.
26. RYAN Kiros, YUKUN Zhu, RUSLAN R. Salakhutdinov, RICHARD Zemel, Raquel Urtasun, ANTONIO Torralba, SANJA Fidler. *Skip-Thought Vectors*. 2015. Disponibil: doi.org/10.48550/arXiv.1506.06726.
27. ALEXIS Conneau, DOUWE Kiela, HOLGER Schwenk, LOIC Barrault, ANTOINE Bordes. *Supervised Learning of Universal Sentence Representations from Natural Language Inference Data*. 2017. Disponibil: doi.org/10.18653/v1/D17-1070.
28. SAMUEL R. Bowman, GABOR Angeli, CHRISTOPHER Potts, CHRISTOPHER D. Manning. *A large annotated corpus for learning natural language inference*. 2015. Disponibil: doi.org/10.18653/v1/D15-1075.
29. ADINA Williams, NIKITA Nangia, SAMUEL Bowman. *A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference*. 2018. Disponibil: doi.org/10.18653/v1/N18-1101.
30. DUMITRESCY, S. D., REBEJA, P., LORINCZ, B., GAMAN, M., AVRAM A., ILIE, M., PRITEANU, A., STAN, A., ROSIA, L., IACOPESCU, C., și alții. *Benchmark and leaderboard for romanian language tasks. Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2021. Disponibil: https://github.com/dumitrescustefan/RO-STS.
31. Model de similaritate semantică. BlackKakapo/stsb-xlm-r-multilingual-ro. 2022. Disponibil: https://huggingface.co/BlackKakapo/stsb-xlm-r-multilingual-ro.
32. JOHN K. Pate, MARK Johnson. *Grammar induction from (lots of) words alone*. 2016. Disponibil: aclanthology.org/C16-1003.

33. QIZHE Xie, ZIHANG Dai, EDUARD Hovy, MINH-THANG Luong, QUOC V. Le. *Unsupervised Data Augmentation for Consistency Training*. 2020. Disponibil: doi.org/10.48550/arXiv.1904.12848.
34. QIU Ran, YANKAI Lin, PENG Li, JIE Zhou, ZHIYUAN Liu. *NumNet: Machine Reading Comprehension with Numerical Reasoning*. 2019. Disponibil: doi.org/10.18653/v1/D19-1251.
35. MARCIN Junczys-Dowmunt, ROMAN Grundkiewicz, TOMASZ Dwojak, HIEU Hoang, KENNETH Heafield, TOM Neckermann, FRANK Seide, ULRICH Germann, ALHAM Fikri Aji, NIKOLAY Bogoychev, ANDRE F. T. Martins, ALEXANDRA Birch. *Marian: Fast Neural Machine Translation in C++*. 2018. Disponibil: doi.org/10.48550/arXiv.1804.00344.
36. HOLGER Schwenk, GUILLAUME Wenzek, SERGEY Edunov, EDUARD Grave, ARMAND Joulin, ANGELA Fan. *CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web*. 2021. Disponibil: doi.org/10.18653/v1/2021.acl-long.507.
37. ALEC Radford, KARTHIK Narasimhan, TIM Salimans, ILYA Sutskever. *Improving Language Understanding by Generative Pre-Training*. 2018.
38. PRAJIT Ramachandran, PETER Liu, QUOC Le. *Unsupervised Pretraining for Sequence-to-Sequence Learning*. 2016. Disponibil: doi.org/10.48550/arXiv.1611.02683.
39. KISHORE Papineni, SALIM Roukos, TODD Ward, WEI-JING Zhu. *BLEU: a Method for Automatic Evaluation of Machine Translation.* 2002. Disponibil: https://aclanthology.org/P02-1040.pdf
40. CHIN-YEW Lin. *ROUGE: A Package for Automatic Evaluation of Summaries.* 2004 . Disponibil: https://aclanthology.org/W04-1013.pdf