ICAI 2023
12th International Conference on Applied Informatics
Eger, Hungary, 2023

A Pseudoanonymisation Tool for Hungarian

Péter Hatvani, László János Laki, Zijian Győző Yang

Hungarian Research Centre for Linguistics
hatvani9823@gmail.com,{yang.zijian.gyozo,laki.laszlo}@nytud.hu

Abstract

Nowadays, huge amounts of documents are created every day and the protection of personal data is an important task. Thus, anonymization is a current problem affecting many fields. The development of natural language processing research has made named entity recognition models more accurate. However, using a named entity recognition system to remove names from a text can make the text incoherent and corrupt the fluency of it. In our research, we introduce a Pseudoanonymization Tool for Hungarian that addresses this issue. Our tool implements a pipeline that integrates different named entity recognition, morphological parsing, and generation modules (see Figure 1). Instead of simply recognizing and removing named entities, our tool replaces found names with other names consistently throughout the document, resulting in a coherent and fluent text. We evaluate our tool using several testing corpora and show that the results are consistent across them.

Our Hungarian Pseudoanonymization Tool integrates different named entity recognition, morphological parsing, and generation modules. The tool uses HuSpaCy [4] and emMorph [6], a Hungarian morphological analyzer [1], to ensure consistent results across multiple testing corpora. The tool is designed for various use cases such as legal, medical documents, and other sensitive texts. However, the testing of the model has been limited to crawled comments, excerpts from programs of the Hungarian Kossuth Radio, and Hungarian literature, serving as proof of concept for the tool's ability to maintain text coherence and fluency in the anonymized output.

In our pipeline, the morphological analyzer achieved 90% accuracy in token classification tasks, while the NER module [7] achieved an F1 score of 80.75% on NerKor [5]. The tool also uses PurePos 2.0 [3], which achieved 96.72% accuracy in

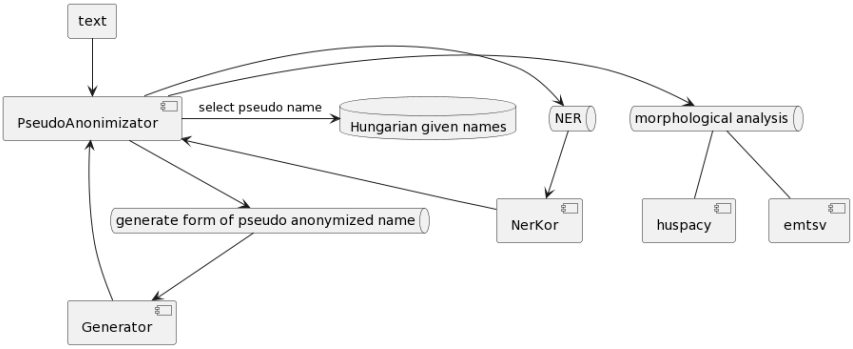

Figure 1. Structure diagram of the Pseudoanonymization Tool

part-of-speech recognition tasks.

In our first experiment, we evaluated the performance of the emMorph and the universal dependency (UD) morphology generators on named entities. We used the NerKor corpus to collect morphologically analyzed tokens and filter them to obtain unique names. We then applied the morphology generators on these names. The results of the evaluation are presented in Table 1. The column labeled "All" displays the performance of the models on all names, but it must be noted that these results may be biased because nearly 86% (emMorph: 85.62%; UD: 86.07%) of names are in the nominative case. To account for this bias, we also present the performance of the models on the non-nominative names in the "Filtered" column.

Table 1. Performance of morphology generators on named entities

	All	Filtered
emMorph	95,10%	84,51%
UD	92,85%	79,39%

In Table 2 and Table 3, you can see the manual evaluation of the Pseudoanonymization tool. In the spok (excerpt from the Hungarian Spoken Speech part of the Hungarian Gigaword Coprus) [2] corpus with the emMorph analyzer the distinction from the worst (p=0.0001) and the ideal (p=0.0029) are hairly distinct compared to the HuSpaCy analyzed it is not significantly distinct (p=0.5933). The same can

Table 2. Performance of pseudoanonymization with emMorph morphological analyzer

text	Found (real)	Found (false positive)	Missed	All
spok	7	8	2	13
comments	42	25	9	67
Ady letters	11	4	9	20

Table 3. Performance of pseudoanonymization with HuSpaCy morphological analyzer

text	Found (real)	Found (false positive)	Missed	All
spok	6	5	5	11
comments	58	19	3	77
Ady letters	6	5	9	15

be observed with the analysis of the two other corpora. However the similarity of the same pipeline on different texts is not as close as one might think emMorph "spok" and "comments" are not close (p=0.0001), only the Ady letter text and the spok corpus with the HuSpaCy pipeline was close (p=0.0454) but they still differed significantly.

References

[1] L. J. Laki, N. Ligeti-Nagy, N. Vadász, Z. G. Yang: *Neural Morphological Generators for Hungarian*, in: XIX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged: Szegedi Tudományegyetem, 2023, pp. 331–340.

[2] C. Oravecz, B. Sass, T. Váradi: *Mennyiségből minőséget. Nyelvtechnológiai kihívások és tanulságok az MNSz új változatának elkészítésében*, in: XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged: Szegedi Tudományegyetem, 2015, pp. 109–121.

[3] G. Orosz, A. Novák: *PurePos 2.0: a hybrid tool for morphological disambiguation*, in: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, Sept. 2013, pp. 539–545, URL: https://aclanthology.org/R13-1071.

[4] G. Orosz, Z. Szántó, P. Berkecz, G. Szabó, R. Farkas: *HuSpaCy: an industrial-strength Hungarian natural language processing toolkit*, Szeged, 2022.

[5] E. Simon, N. Vadász: *Introducing NYTK-NerKor, A Gold Standard Hungarian Named Entity Annotated Corpus*, in: Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings, ed. by K. Ekstein, F. Pártl, M. Konopik, vol. 12848, Lecture Notes in Computer Science, Springer, 2021, pp. 222–234, DOI: 10.1007/978-3-030-83527-9_19.

[6] V. Trón, G. Gyepesi, P. Halácsy, A. Kornai, L. Németh, D. Varga: *Hunmorph: open source word analysis*, in: Proceedings of the ACL 2005 Software Workshop, ed. by M. Jansche, Ann Arbor: ACL, 2005, pp. 77–85.

[7] Z. G. Yang, T. Váradi: *Training language models with low resources: RoBERTa, BART and ELECTRA experimental models for Hungarian*, in: Proceedings of 12th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2021), Online: IEEE, 2021, pp. 279–285.