Transferability of General Polish NER to Electronic Health Records

Krištof Anetta and Mahmut Arslan

Natural Language Processing Centre,
Faculty of Informatics, Masaryk University
Botanická 68a, Brno, Czech Republic
xanetta@fi.muni.cz, xarslan@fi.muni.cz

Abstract. This paper investigates the transferability of general Polish named entity recognition tools to the analysis of Polish health records. The tools, namely PolDeepNer2, spaCy's *pl_core_news_lg* pipeline and Spark NLP's *entity_recognizer_md* pipeline for Polish, were run on the *pl_ehr_cardio* corpus and their results were analyzed, paying special attention to their performance when processing these highly specific texts and to the applicability of the results in the healthcare domain. Even though the precision of PolDeepNer2 proved to be superior to both spaCy and Spark NLP, the paper concludes that without additional training, general named entity recognition tools for Polish have very limited use in the medical analysis of electronic health records. However, they could be helpful in partial tasks ranging from de-identification to entity disambiguation and discovery of mistyped entities or candidate entities that are not present in medical dictionaries.

Keywords: EHR · Electronic health records · Healthcare texts · NER · Named entity recognition · NLP · Natural language processing · Slavic languages · Polish · PolDeepNer2 · spaCy · Spark NLP

1 Introduction

In the past decade, NLP for healthcare, especially entity recognition, has been growing rapidly in the English-speaking world. However, low-resourced languages like Polish have been progressing much more slowly due to the combined effects of a lack of resources at every level of processing. The key disadvantage is the absence of a Polish UMLS translation - while English UMLS boasts more than 9 million terms [1], facilitating knowledge extraction, Polish only has around 50,000 terms in the MeSH subset, which is both too sparse and too general to be of use in health records. Until better Polish healthcare dictionaries are developed, researchers have the option to train deep learning entity recognition systems to find strings which are likely to be medical entities based on their features. As there are currently no benchmark tools for discovering Polish medical entities (notable work has been done by [2], but without generalizable search for new entities), this paper surveys the borderland between general entity recognition and healthcare entity recognition, trying to find out to what extent the

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021, pp. 151–159, 2021. © Tribun EU 2021

Table 1: Mapping of entity categories

	PolDeepNer2	spaCy	Spark NLP
PER	nam_liv	persName	PER
ORG	nam_org	orgName	ORG
LOC	nam_loc nam_fac	placeName geogName	LOC
MISC	nam_eve nam_pro nam_adj nam_num nam_oth	date time	MISC

existing general Polish entity recognition systems can be ported to the healthcare domain.

When looking for named entities in Polish text, there are several options to consider [3], ranging from deep learning to dictionary-based approaches. In this paper, three recently updated options were chosen for comparison - PolDeepNer2 [4] with the KPWr n82 NER model [5] was chosen as the state-of-the-art, custom-made deep learning approach (categories were simplified for the statistics), spaCy's [6] *pl_core_news_lg* pipeline was chosen based on its effortless availability to any spaCy user, and Spark NLP's [7] *entity_recognizer_md* pipeline for Polish was chosen because of Spark NLP's noticeable presence in healthcare text processing - there are already clinical NLP models for English, German, and Spanish, which hints at potential future extensibility of Spark NLP's general Polish entity recognition into clinical entity recognition.

The analyzed corpus, *pl_ehr_cardio* [8], consists of more than 50,000 health records related to cardiology collected over 18 years at the Medical University of Silesia in Katowice, Poland. The corpus contains more than 23 million words.

2 NER Results in *pl_ehr_cardio*

In order to compare the results, a mapping between categories used by individual tools had to be decided. PER, ORG, LOC and MISC were chosen as the unifying categories with the mapping shown in Table 1. Table 2 compares the total counts and ratios of entities found in the corpus. Tables 3, 4, and 5 show entity statistics for the entire corpus processed by PolDeepNer2, spaCy, and Spark NLP, respectively.

Table 2: Total counts of entities in the *pl_ehr_cardio* corpus. Total word count of the corpus is 23,831,785.

	PolDeepNer2		spaCy		Spark NLP	
all	**725,198**		**965,225**		**3,428,457**	
PER	170,969	23.6%	350,749	36.3%	381,543	11.1%
ORG	119,321	16.5%	248,115	25.7%	502,457	14.7%
LOC	21,026	2.9%	78,888	8.2%	1,350,885	39.4%
MISC	413,882	57.1%	287,473	29.8%	1,193,572	34.8%

3 Performance Analysis

3.1 Analyzed Sample Characteristics

The sample chosen for manual analysis consisted of a pseudo-random selection of 17 patient records totaling 9382 words, evenly distributed across the 18-year timespan of the *pl_ehr_cardio* corpus. Table 6 summarizes the precision achieved by individual tools, in total and per category. The MISC category is not evaluated because it has a different meaning for each tool and its boundaries are fuzzy - furthermore, the status of a named entity is especially difficult to establish in medical terminology.

3.2 PolDeepNer2

PolDeepNer2 identified 193 named entities in the analyzed sample. It was the smallest number of entities of all the tools, but they were identified with significantly greater precision.

Names of people Within the sample chosen for analysis, 100% (54/54) of what PolDeepNer2 identified as names of people was correct, even though in most

Table 3: PolDeepNer2 statistics for entities. The ◁ symbol separates values for the minimum, average and maximum number of entities per the specified text block.

per	any entity	PERson	ORGanization	LOCation	MISCellaneous
sentence	0 ◁ 2.1 ◁ 32	0 ◁ 1.3 ◁ 11	0 ◁ 1.3 ◁ 12	0 ◁ 1.2 ◁ 13	0 ◁ 2.4 ◁ 31
paragraph	0 ◁ 4.0 ◁ 92	0 ◁ 2.0 ◁ 33	0 ◁ 1.6 ◁ 36	0 ◁ 1.4 ◁ 13	0 ◁ 4.1 ◁ 67
epicrisis physicalexam	0 ◁ 2.7 ◁ 38	0 ◁ 1.2 ◁ 8	0 ◁ 1.4 ◁ 10	0 ◁ 1.2 ◁ 6	0 ◁ 2.3 ◁ 24
epicrisis recommendation	0 ◁ 3.4 ◁ 25	0 ◁ 1.1 ◁ 8	0 ◁ 1.7 ◁ 11	0 ◁ 1.6 ◁ 12	0 ◁ 3.0 ◁ 21
interview onset	0 ◁ 5.8 ◁ 92	0 ◁ 1.4 ◁ 11	0 ◁ 1.8 ◁ 36	0 ◁ 1.4 ◁ 13	0 ◁ 5.2 ◁ 67
interview physicalexam	0 ◁ 2.3 ◁ 76	0 ◁ 2.1 ◁ 33	0 ◁ 1.0 ◁ 9	0 ◁ 1.3 ◁ 13	0 ◁ 1.4 ◁ 44
document	0 ◁ 9.5 ◁101	0 ◁ 2.4 ◁ 33	0 ◁ 2.5 ◁ 38	0 ◁ 1.6 ◁ 14	0 ◁ 6.9 ◁ 74

Table 4: Spacy statistics for entities. The ◁ symbol separates values for the minimum, average and maximum number of entities per the specified text block.

per	any entity	PERson	ORGanization	LOCation	MISCellaneous
sentence	0 ◁ 2.0 ◁ 70	0 ◁ 1.4 ◁ 16	0 ◁ 1.1 ◁ 11	0 ◁ 1.0 ◁ 5	0 ◁ 2.4 ◁ 58
paragraph	0 ◁ 5.0 ◁ 110	0 ◁ 2.7 ◁ 57	0 ◁ 1.8 ◁ 42	0 ◁ 1.7 ◁ 24	0 ◁ 4.0 ◁ 72
epicrisis phys.exam	0 ◁ 3.3 ◁ 50	0 ◁ 1.4 ◁ 14	0 ◁ 1.6 ◁ 12	0 ◁ 1.1 ◁ 6	0 ◁ 2.6 ◁ 37
epicrisis recomm.	0 ◁ 2.8 ◁ 25	0 ◁ 2.2 ◁ 16	0 ◁ 1.3 ◁ 9	0 ◁ 1.1 ◁ 4	0 ◁ 3.0 ◁ 21
interview onset	0 ◁ 6.5 ◁ 110	0 ◁ 2.1 ◁ 24	0 ◁ 2.2 ◁ 42	0 ◁ 1.4 ◁ 10	0 ◁ 4.6 ◁ 62
interview phys.exam	0 ◁ 4.9 ◁ 110	0 ◁ 3.1 ◁ 57	0 ◁ 1.6 ◁ 17	0 ◁ 2.0 ◁ 24	0 ◁ 6.3 ◁ 72
document	0 ◁ 12.1 ◁ 136	0 ◁ 4.7 ◁ 65	0 ◁ 3.6 ◁ 45	0 ◁ 2.2 ◁ 24	0 ◁ 5.7 ◁ 89

cases these names were parts of the names of medical examinations, conditions, and methods named after their discoverers or inventors ("objaw **Chełmońskiego / Blumberga / Goldflamma / Babińskiego**", "choroba **Buergera**", "metodą **Holtera**"). This unintended capability proves especially useful in cardiology where discoverer-based medical concept names are common. With some additional rule-based evaluation on top of PolDeepNer2's person name recognition, it could be a useful addition to a Polish healthcare text processing system.

Names of organizations Medical organization names were more difficult for PolDeepNer2, but it still fared quite well - in the analyzed sample, 81.8% (36/44) of strings identified as organization names were in fact names of organizations or individual departments and offices of those organizations ("**Poradni Kardiologicznej i Diabetologicznej**", "**Szpitala w Tychach**", "**Szpitala w Świętochłowicach**", "**Oddziału Intensywnej Terapii z Nadzorem Kardiologicznym**"). Almost all of the errors occurred in the most difficult kind of organization names - capitalized abbreviations. Apart from surprising success with some instances ("**OAITK zNK**", "**OITK**", "**POChP**", "**MIC**", "**POZ**"), there were some non-organization abbreviations that slipped in ("**LAD**", "**UKG**",

Table 5: Spark NLP statistics for entities. The ◁ symbol separates values for the minimum, average and maximum number of entities per the specified text block.

per	any entity	PERson	ORGanization	LOCation	MISCellaneous
sentence	0 ◁ 1.9 ◁ 82	0 ◁ 1.2 ◁ 10	0 ◁ 1.2 ◁ 15	0 ◁ 1.3 ◁ 49	0 ◁ 1.8 ◁ 36
paragraph	0 ◁ 19.0 ◁ 536	0 ◁ 2.7 ◁ 38	0 ◁ 6.1 ◁ 144	0 ◁ 8.1 ◁ 214	0 ◁ 7.3 ◁ 178
epicrisis phys.exam	0 ◁ 2.7 ◁ 536	0 ◁ 1.4 ◁ 11	0 ◁ 1.4 ◁ 19	0 ◁ 10.8 ◁ 214	0 ◁ 2.3 ◁ 24
epicrisis recomm.	0 ◁ 5.3 ◁ 52	0 ◁ 1.7 ◁ 10	0 ◁ 2.1 ◁ 17	0 ◁ 2.3 ◁ 16	0 ◁ 3.4 ◁ 32
interview onset	0 ◁ 11.7 ◁ 272	0 ◁ 2.2 ◁ 31	0 ◁ 2.2 ◁ 29	0 ◁ 5.2 ◁ 117	0 ◁ 6.0 ◁ 104
interview phys.exam	0 ◁ 2.3 ◁ 76	0 ◁ 3.3 ◁ 38	0 ◁ 8.2 ◁ 144	0 ◁ 1.3 ◁ 13	0 ◁ 10.3 ◁ 178
document	0 ◁ 40.5 ◁ 567	0 ◁ 5.0 ◁ 38	0 ◁ 8.8 ◁ 145	0 ◁ 15.9 ◁ 218	0 ◁ 15.6 ◁ 188

"**POWLOK**", "**EKG**", "**LOC**"), likely due to the notorious syntactical insufficiency of health records that confused the contextual classifier.

Even though this might raise a suspicion that PolDeepNer2 chose these abbreviations superficially, based on the capitalization of all of their letters, it proves to be unfounded upon closer analysis - there were more than 300 capitalized abbreviations in the sample and only 12 of those were recognized as organization names, demonstrating the high specificity of PolDeepNer2's criteria.

In addition to the above, PolDeepNer2 was able to identify incomplete references to organizations ("**Kliniki**") and recognize an entity in spite of an error in a crucial noun ("**Kliniii Chirurgii Ogólnej i Naczyń**").

Names of locations In the analyzed sample, PolDeepNer2 only identified 1 occurrence of a location, which is not enough to evaluate its performance. This occurrence was labeled incorrectly, as it was a general reference to organizations ("w **Poradnich**") the syntactical use of which resembled a geographical name.

Miscellaneous names Miscellaneous is perhaps the most interesting category, since it has the potential to discover names that are actually relevant for medicine. PolDeepNer2 found 94 miscellaneous names, further divided into 16 product names, 9 event names, and 69 "other" names. Of the product names, 68.8% (11/16) can be considered correct, including 9 medicine names (e. g. "**Biosotal**","**Mixtrad**", "**Encorton**", "**Theovent**", "**Pentohexal 600**") and 2 device names ("w **Holterze**", "**EKG**"). Of the event names, 44.4% (4/9) were correct, identifying 2 heart attacks ("**NSTEMI**", "**Przebyty udar**") and 2 medical procedures (e. g. "**POBA**"). Errors in the product and event categories resulted from incorrectly labeling capitalized abbreviations with insufficient syntactical context, namely 100% (10/10) of errors were strings either entirely composed of capital letters and numbers or including a capitalized non-word substring (e. g. "**PTCA LAD**", "**Stan po POBA**", "**R57**").

The "other" category is more difficult to evaluate because almost anything in health records can be considered an entity, even though rarely a proper name. Of the 69 strings labeled in this way, there were 16 additional medicine names

Table 6: Performance comparison for commensurable categories. Precision was manually evaluated on a subset of records.

	PolDeepNer2		spaCy		Spark NLP	
all	**90.9%**	**90/99**	**40.3%**	**104/258**	**7.6%**	**59/780**
PER	100%	54/54	41.1%	53/129	34.4%	45/131
ORG	81.8%	36/44	50.5%	51/101	6.1%	11/179
LOC	0%	0/1	0%	0/28	0.6%	3/470

(in addition to the ones identified as product names) and a varied collection of medical states, procedure names, and institution name abbreviations. 79.7% (55/69) of the "other" names were strings that were either capitalized or exhibited a different sign of being an abbreviation, such as including a number (e. g. "**CCS II**", "**WZWB**", "**DDD**", "**Ao-OM2**", "**TILT**"). On the one hand, these matches seem to be highly relevant for medicine, but on the other hand, since the system has no idea what it has found, significant further processing or approach hybridization would be required to turn these discoveries into knowledge in big data.

3.3 spaCy

spaCy's *pl_core_news_lg* pipeline identified 403 entities in the analyzed sample.

Names of people spaCy identified 129 strings as names of people, but only 41.1% (53/129) were actual names, and they were exclusively the names of medical concepts named after their inventors, the very same ones that were described in the PolDeepNer2 section. 17.1% (22/129) were incorrectly labeled medicine names that probably confused the system by their capitalized first letter. Most of the remaining errors were standard words, often describing a body part or a characteristic looked for in the examination (e. g. "**TKANKA**", "**ODGLOS**", "**Wątroba**", "**Tony**"). Interestingly, there were cases where the first letter was not even capitalized ("**ablacją**", "**tężcowa**").

Names of organizations 101 strings were labeled as organization names, however, only 50.5% (51/101) were truly referring to organizations and their individual departments and offices. Similar to the names of people, the errors included 10 medicine names and a mix of regular words relating to medical examinations ("**Uczulenia**", "**stentem**", "**TARCZYCA**", "**Cholesterolu**")

Names of locations PolDeepNer2 already indicated that health records are not rich in location names and this was the case for spaCy as well. It identified 28 strings as names of locations, of which 0% (0/28) were correct in the proper, narrow sense of what a location is. There were, however, 15 instances of locations on the body ("**GRANICE DOLNE PLUC**", "**Spojówki**", "**Tarczyca**"), resulting from a syntactical similarity which could prove useful in the analysis of body references in health records.

Miscellaneous names spaCy's miscellaneous category only includes dates and times mentioned in the text, and is therefore quite different from the same category in the other tools. The performance of spaCy in this particular task was decent and potentially useful for temporal marking of health records. In the analyzed sample, 145 strings were identified as date or time, of which 97.2% (141/145) were correct. Errors included mistakenly labeled use of numbers, e. g. drug dosage or measurements ("**1-0-0**", "BMI **21.08**").

3.4 Spark NLP

Spark NLP's *entity_recognizer_md* pipeline for Polish proved to be overwhelmingly optimistic in its guesses. It found 1193 entities in the analyzed sample, 6 times as many as PolDeepNer2 and 3 times as many as spaCy, which was already too optimistic to start with.

Names of people Interestingly, despite being extremely liberal with other labels, Spark NLP identified 131 strings as names of people, a result very close to spaCy's 129. 34.4% (45/131) of these strings correctly captured a personal name, but often included other words that did not belong with the name ("**Objaw Goldflama**", "**Objawy Chełmońskiego**"), likely because of the capitalization of the neighboring words. Only 19.8% (26/131) were clean personal names.

Names of organizations Spark NLP's performance on organizations was outright abysmal. Only 6.1% (11/179) of the found strings were truly referring to organizations. The most obvious error pattern was related to capitalization - in 69.8% (125/179) of strings identified as organizations, more than half of all characters were either capital letters or numbers, thus resembling abbreviations and company/institution names, even if they were regular Polish words (e. g. "**WYPOWIADA**", "**SKORA**", "**CZASZKA**").

Names of locations Compared with PolDeepNer2's 1 and spaCy's 28, Spark NLP's 470 results for location names sounds too good to be true, and it is. Only 0.6% (3/470) of the strings identified as location names were geographical locations. Interestingly, 29.6% (139/470) of the strings represented locations on or within the body ("**Gałki**", "**Śledziona**", "**Brzuch**"), which, if the precision improved, could be useful for health record analysis. While body location errors can be explained by syntactical similarity, another notable error pattern is more difficult to explain: 6.8% (32/470) of the identified strings were medicine names ("**Acard**", "**Milurit**", "**Tertensif SR**") which often stand alone in the text, outside of sentence structure, and therefore there seems to be no reason to consider them location names apart from the capital letter at their beginning.

Miscellaneous names In short, the noise in this category renders the results unusable. The 413 identified strings were chosen for indecipherable reasons and they ranged from meaningless fragments (e. g. "**V**", "**Po**", "**(EF**", "**-0-10j).**") to regular words to abbreviations and codes. Capitalization and code-like nature seemed to matter, as 44.1% (182/413) of the strings were more than half capitals or numbers.

An interesting error in the miscellaneous category was the labeling of very long strings. 19.6% (81/413) of the strings identified as miscellaneous names were longer than 20 characters, 6.5% (27/413) were longer than 30 characters, and 1.9% (8/413) were longer than 40 characters. None of these longer strings was a proper name.

4 Conclusion

The tested named entity recognition tools were facing a highly improbable task and they met, and in the case of PolDeepNer2, exceeded the expectations set at the start. That said, if we were to ask the question whether existing general named entity recognition for Polish can render useful results for electronic health records, the answer is a clear no - even in the tasks they are relatively good at (PolDeepNer2's performance in names of people and organizations), recall is threatened by the syntactical poverty of health record text, and once the tools attempt to identify other types of entities, they no longer label them correctly, thus providing no information on how to handle them. In addition to all this, the basic entity categories that the models are looking for do not overlap well with what is relevant for medical science. Names of body parts, symptoms, and diagnoses do not fit anywhere, the often abbreviated names of procedures, even though sometimes identified as events, end up scattered amongst categories, and some, but not enough names of medicines are identified by PolDeepNer2 as products. Even with radically improved performance, the existing tools would not be looking for the relevant data in the first place.

Of course, this is an unfair question to ask, as these tools were never intended for such texts - their failure is expected and understandable. A more productive question is whether the existing tools could be useful with some additional training or as a part of a more complex processing pipeline, and here the results suggest a much more positive outlook - especially PolDeepNer2, apart from providing the obvious and highly demanded service of de-identification by finding personal names with great precision, might be able to enhance dictionary-based lookup techniques for medical entities by providing candidate entities that are either unknown to the lookup system or distorted by errors, or it could help disambiguate the meaning of previously identified entities by labeling them with their role. Additional training on medicine names could easily improve the recognition of product names, which could go beyond the available databases of medical products and identify alternative product names or even the medicinal use of products that are originally non-medical.

Research on Polish electronic health records is still in its infancy, but the rapid global development of transformer architectures together with Polish-specific research initiatives are quickly progressing towards their first successes in mining structured data from the cryptic, time-pressured writing produced in hospitals and doctors' offices.

Acknowledgements. This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101. The Polish data was provided by the Medical University of Silesia in Katowice, Poland, with the help of Prof. Krystian Wita, MD PhD, Prof. Wojciech Wojakowski, MD PhD, Prof. Wacław Kuczmik, MD PhD, Tomasz Jadczyk, MD PhD, and Marcin Wita, MD PhD.

References

1. Névéol, A., Dalianis, H., Velupillai, S., Savova, G., Zweigenbaum, P.: Clinical Natural Language Processing in Languages Other Than English: Opportunities and Challenges. *Journal of Biomedical Semantics*, vol. 9, issue 1, pp. 1-13, 2018. https://doi.org/10.1186/s13326-018-0179-8
2. Dobrakowski, A.G., Mykowiecka, A., Marciniak, M., Jaworski, W., Biecek, P.: Interpretable Segmentation of Medical Free-text Records Based on Word Embeddings. *Journal of Intelligent Information Systems*, 2021. https://doi.org/10.1007/s10844-021-00659-4
3. Marcińczuk, M., Wawer, A.: Named Entity Recognition for Polish. *Poznan Studies in Contemporary Linguistics* vol. 55, issue 2, pp. 239-269, 2019. https://doi.org/10.1515/psicl-2019-0010
4. Marcińczuk, M., Radom, J.: A Single-run Recognition of Nested Named Entities with Transformers. *Procedia Computer Science*, vol. 192, pp. 291-297, 2021. https://doi.org/10.1016/j.procs.2021.08.030
5. Marcińczuk, M.: KPWr n82 NER model (on Polish RoBERTa base). CLARIN-PL digital repository, 2020. `http://hdl.handle.net/11321/743`
6. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python. 2020. https://doi.org/10.5281/zenodo.1212303
7. Kocaman, V., Talby, D.: Spark NLP: Natural Language Understanding at Scale. *Software Impacts*, vol. 8, 100058, 2021. https://doi.org/10.1016/j.simpa.2021.100058
8. Anetta, K.: Data Mining from Free-Text Health Records: State of the Art, New Polish Corpus. In: Horák, A. (ed.) *Proceedings of the Fourteenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2020*, pp. 13-22. Tribun EU, Brno (2020). ISBN 978-80-263-1600-8