

Transferability of General Polish NER to Electronic Health Records

A report on *pl_ehr_cardio*

Krištof Anetta, Mahmut Arslan
xanetta@fi.muni.cz, xarslan@fi.muni.cz

Natural Language Processing Centre
Faculty of Informatics, Masaryk University

December 11, 2021

TL;DR

Available Polish NER is almost useless for mining health records.

TL;DR

Available Polish NER is almost useless for mining health records.

- However, the exact proportions of its failures contain some food for thought for anyone interested in healthcare NLP or Polish language research.

Our data

- *pl_ehr_cardio*: Polish corpus of health records
 - 50,000+ hospitalizations
 - 18 years
 - cardiology only
 - ICD-10 diagnosis code for each hospitalization
- Unstructured text data entered by doctors consisting of 4 sections:
 1. Interview (onset): Arrival of the patient, reported symptoms
 2. Interview (physical examination): Doctor's findings after preliminary physical examination
 3. Discharge (physical examination): Summary of the hospitalization, procedures, diagnoses
 4. Discharge (recommendation): Recommended medications and behavior changes

Core motivation

Discover latent knowledge hidden in unstructured text.

Data sample

Lp.	Wywiad - Początek choroby - Treść	Wywiad - Badanie przedmiotowe - Treść	Epikryza - Badanie fizykalne - Treść	Epikryza - Zalecenia lekarskie - Treść	Rozp. końc. - choroba zasadnicza - Kod choroby
1	<p>Pacjent lat 64 po zawale serca ściany przedniej (2005), po wielokrotnych angioplastykach naczyń wieńcowych, został przyjęty do Oddziału z powodu nawrotu dolegliwości stenokardialnych. W ostatnim czasie epizody bólu w klp o charakterze piekącym, ostatnio epizod trwał 4 dni. W wywiadzie: Stan po zawale serca ściany przedniej 2005r. Stan po PCI LAD + stent (2005). Stan po PCI Cx + stent (2005). Stan po PCI-RCA+stent (11.2011) Zaburzenia lipidowe. Uczulenia: nie podaje. Wywiad rodzinny: brat - IM Użytki: nie pali od 15 lat - wcześniej do paczki/3 dni kawa - 1/dz, alkohol - okazjonalnie Szczepienie WZW: NIE</p>	<p>Pacjent przytomny, ułożenie dowolne, kontakt logiczny zachowany. Skóra czysta, bez wykwitów, tkanka podskórna prawidłowo rozwinięta, węzły chłonne niewyczuwalne. Głowa niebolesna opukowo, gałki oczne osadzone prawidłowo, symetryczne, źrenice równe, okrągłe, prawidłowo reagują na światło i zbieżność. Nos drożny, śluzówki jamy ustnej wilgotne, różowe, gardło blade, migdałki podniebienne niepowiększone, bez nalotów. Język prawidłowo ruchomy, bez nalotów. Uzębienie kompletne. Szyja: ruchomość czynna i bierna zachowana, tarczyca niepowiększona. Klatka piersiowa: ruchomość oddechowa zachowana, wypuk jawny, szmer pecherzykowy prawidłowy, drżenie głosowe zachowane. Akcja serca miarowa ok. 60/min, tony serca czyste, prawidłowo akcentowane, bez szmerów patologicznych. Brzuch miękki, niebolesny, bez oporów patologicznych. Powłoki wysklepione w poziomie klatki piersiowej, wątroba niepowiększona, śledziona niepowiększona, objawy: Chęłmońskiego i Blumbrga ujemne. Objawy otrzewnowe ujemne.</p>	<p>Pacjent lat 64 po zawale serca ściany przedniej (2005), po wielokrotnych angioplastykach naczyń wieńcowych, został przyjęty do Oddziału z powodu nawrotu dolegliwości stenokardialnych. W echokardiografii stwierdzono hypokinezę przegrody międzykomorowej, koniuszka, ściany dolnej przy frakcji wyrzutowej lewej komory 50%. Wykonana koronarografia wykazała krytyczną zmianę w tętnicy diagonalnej, wykonano jednoczasowo angioplastykę balonową zmiany. Dodatkowo opisano występowanie mostka mięśniowego w przebiegu gałęzi międzykomorowej przedniej (LAD). Po modyfikacji farmakoterapii pacjent wypisany do domu w stanie dobrym z w/w zaleceniami.</p>	<p>Okresowa kontrola w Poradni Kardiologicznej, Dieta małosolna, ubogotłuszczowa, Kontrola wartości ciśnienia tętniczego. Regularne zażywanie leków: Bisohexal 5mg 1/2-0-0 tabl., Vivace 2,5mg 1/2-0-0 tabl., Polocard 150mg 1-0-0 tabl., Simvagen 20mg 0-0-1 tabl., Agen 5mg 0-0-1., Areplex 75mg 1-0-0 tabl. przez okres co najmniej 3 miesięcy pod kontrolą morfologii krwi z płytkami.</p>	<p>I20.0</p>

Corpus statistics: Unit counts

Documents (hospitalizations)	50,469
Paragraphs (sections)	198,737
Sentences	2,573,000
Words	23,831,785
Interview (onset) sections	49,873
Interview (physical examination) sections	49,833
Discharge (physical examination) sections	49,722
Discharge (recommendation) sections	49,312

Used NER tools: PolDeepNer2

CLARIN-PL / PolDeepNer2 Public

[Code](#) [Issues](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#)

master 1 branch 0 tags

[Go to file](#) [Code](#)

mezcuk

Comparison of NER times for different datasets

ca585d6 on Apr 9 21 commits

docs/media	Comparison of NER times for different datasets	8 months ago
notebooks	Notebook on how to install and use PolDeepNer2	9 months ago
README.md	Comparison of NER times for different datasets	8 months ago
poleval_ner_test.py	Installation description. Added evaluation scripts.	9 months ago
process_poleval.py	Installation description. Added evaluation scripts.	9 months ago

README.md

PolDeepNer2

PolDeepNer2 is an Improved version of [PolDeepNer](#). The tool is designed to recognize and categorize named entities utilizing neural networks and transformer-based language models.

The tool is provided with a list of pre-trained models for Polish and other languages.

It contains a pre-trained model trained on the [NIKJP corpus](#) which recognizes nested annotations of the following types:

Contributors

- Michał Marcińczuk marcinczuk@gmail.com
- Jarema Radom

About

An improved tool for named entity recognition for Polish based on deep learning.

[deep-learning](#) [named-entities](#) [ner](#) [roberta](#) [fine-grained-ner](#) [nikjp](#) [kper](#) [polver](#)

[Readme](#)

Releases

No releases published

Packages

No packages published

Languages

Python 69.0%

Jupyter Notebook 31.0%

Used NER tools: spaCy

spaCy

★ Out now: spaCy v3.2

USAGE

MODELS

MODELS

Overview

TRAINED PIPELINES

Catalan

Chinese

Danish

Dutch

English

French

German

Greek

Italian

Japanese

Lithuanian

Macedonian

Multi-language

Norwegian Bokmål

Polish

● pl_core_news_sm

● pl_core_news_md

● pl_core_news_lg

Portuguese

Romanian

Russian

Spanish

pl_core_news_lg

RELEASE DETAILS

Latest: 3.2.8

Polish pipeline optimized for CPU. Components: tok2vec, morphologizer, tagger, parser, sender, ner, attribute_ruler, lemmatizer.

LANGUAGE	PL Polish
TYPE	CORE Vocabulary, syntax, entities, vectors
GENRE	NEWS written text (news, media)
SIZE	LG 583 MB
COMPONENTS ?	tok2vec, morphologizer, parser, tagger, sender, attribute_ruler, lemmatizer, ner
PIPELINE ?	tok2vec, morphologizer, parser, tagger, attribute_ruler, lemmatizer, ner
VECTORS ?	500k keys, 500k unique vectors (300 dimensions)
SOURCES ?	UD Polish PDB v2.8 (Wróblewska, Alina; Zeman, Daniel; Mašek, Jan; Rosa, Rudolf) National Corpus of Polish (Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, Marek Łaziński, Piotr Pęzik, Adam Przepiórkowski) National Corpus of Polish (Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, Marek Łaziński, Piotr Pęzik, Adam Przepiórkowski) PolMori (Woliński, Marcin; Miłkowski, Marcin; Ogrodniczuk, Maciej; Przepiórkowski, Adam; Szalkiewicz, Lukasz) Explosion fastText Vectors (cbow, OSCAR Common Crawl + Wikipedia) (Explosion)
AUTHOR	Explosion
LICENSE	GNU GPL 3.0

Used NER tools: Spark NLP

[Home](#)[Docs](#)[Learn](#)[Models](#)[Demo](#)

John Snow Labs | Mar 22, 2021

Recognize Entities DL Pipeline for Polish - Medium

[open_source](#)[polish](#)[entity_recognizer_md](#)[pipeline](#)[pl](#)

Description

The entity_recognizer_md is a pretrained pipeline that we can use to process text with a simple pipeline that performs basic processing steps. It performs most of the common text processing tasks on your dataframe

[Live Demo](#)[Open in Colab](#)[Download](#)

Why these 3?

- PolDeepNer2 [4] is the state of the art
- PolDeepNer2's KPWr model [3] has some categories potentially useful for medicine
- spaCy [1] and Spark NLP [2] are widely used, so their biggest Polish models (*pl_core_news_lg* and *entity_recognizer_md* for Polish, respectively) are natural entry points for researchers who primarily deal with other languages
- Part of Spark NLP's business are models for healthcare (so far in English, German, and Spanish) and future extension into Polish is possible

What we did



Figure: Performance was manually evaluated in the BRAT annotation tool on a balanced 10,000-word subset

Performance comparison

Table: Entity counts (corpus word count is 23,831,785)

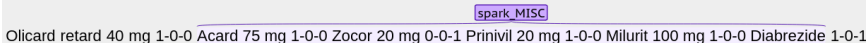
	PolDeepNer2		spaCy		Spark NLP	
all	725,198		965,225		3,428,457	
PER	170,969	23.6%	350,749	36.3%	381,543	11.1%
ORG	119,321	16.5%	248,115	25.7%	502,457	14.7%
LOC	21,026	2.9%	78,888	8.2%	1,350,885	39.4%
MISC	413,882	57.1%	287,473	29.8%	1,193,572	34.8%

Table: Precision comparison (manually evaluated, MISC cannot be compared)

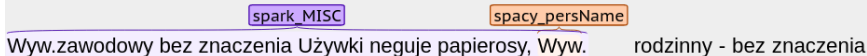
	PolDeepNer2		spaCy		Spark NLP	
all	90.9%	90/99	40.3%	104/258	7.6%	59/780
PER	100%	54/54	41.1%	53/129	34.4%	45/131
ORG	81.8%	36/44	50.5%	51/101	6.1%	11/179
LOC	0%	0/1	0%	0/28	0.6%	3/470

Spark NLP for Polish: A complete failure

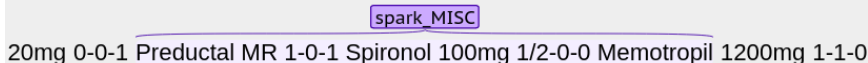
Olicard retard 40 mg 1-0-0 Acard 75 mg 1-0-0 Zocor 20 mg 0-0-1 Prinivil 20 mg 1-0-0 Milurit 100 mg 1-0-0 Diabrezide 1-0-1



Wyw.zawodowy bez znaczenia Używki neguje papierosy, Wyw. rodzinny - bez znaczenia



20mg 0-0-1 Preductal MR 1-0-1 Spironol 100mg 1/2-0-0 Memotropil 1200mg 1-1-0



Sortis 20 mg 0-0-1 Heminervin 300 mg 1-1-1 Risperidone 1 mg 1/2-0-1/2 kalipoz 1-0-0

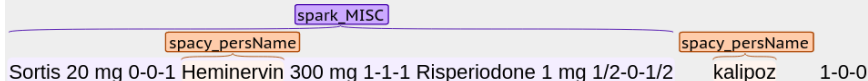


Figure: Extremely long matches

Spark NLP for Polish: A complete failure

spark_LOC

Budowa ciała prawidłowa.

spark_LOC

Skóra czysta, bez wykwitów patologicznych i blizn.

spark_LOC

Tkanka tłuszczowa podskórna prawidłowo rozwinięta, obwodowe węzły chłonne niepowiększone, bez obrzęków obwodowych.

spark_LOC

Czaszka wysklepiona symetrycznie.

spark_LOC

Gałki oczne ustawione symetrycznie, o zachowanej ruchomości.

spark_LOC

Żrenice okrągłe, symetryczne.

spark_LOC

Nos, przewody słuchowe zewnętrzne drożne, bez wydzieliny.

spark_LOC

Szyja o zachowanej ruchomości.

spark_LOC

Gruzoł tarczowy niepowiększony, gładki, ruchomy połykowo.

spark_LOC

Klatka piersiowa wysklepiona symetrycznie, ruchoma oddechowo, palpacyjnie niebolesna.

Figure: There are sections in which every sentence beginning is identified as a location

Spark NLP for Polish: A complete failure

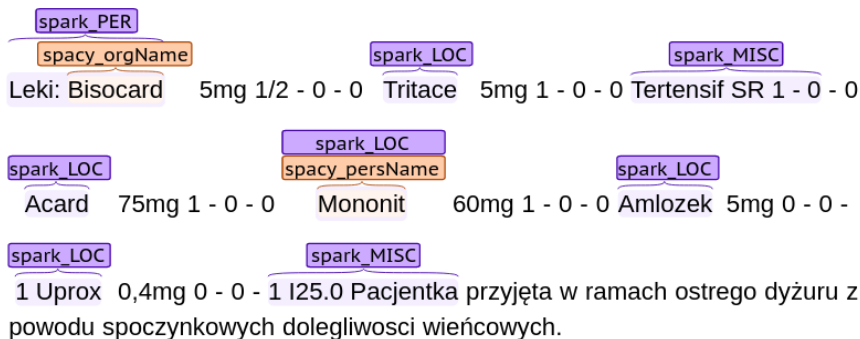


Figure: Spark NLP does seem to match some medicine names, but it is because it matches almost everything

spaCy: Better, but still lots of noise

spark_MISC
spacy_persName

Drżenie głosowe zachowane, symetryczne, ODGŁOS OPUKOWY jawny, nieco symetryczny.

spark_MISC
spacy_placeName

GRANICE DOLNE PLUC: po stronie prawej i lewej symetryczne, na prawidłowej wysokości.

spacy_persName

· leczona ablacją ?, Używki: papierosy - od 6 m-cy nie pali, paliła 5/dz/7 lat Uczulenia - nie podaje

spark_LOC
spacy_orgName

spark_LOC
spacy_placeName

Spojówki: wilgotne, ukrwione prawidłowo.

Figure: False positives undermine the usability of spaCy's Polish model

spaCy: Date/time functionality

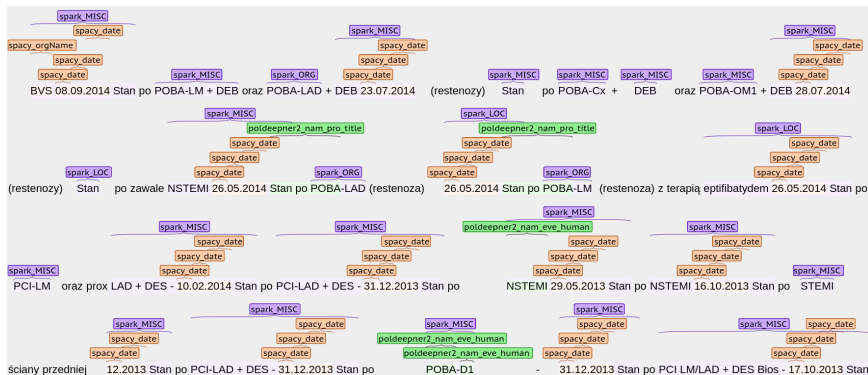


Figure: Most date/time expressions spaCy finds are language-independent, but it is a useful feature

Shared error tendencies

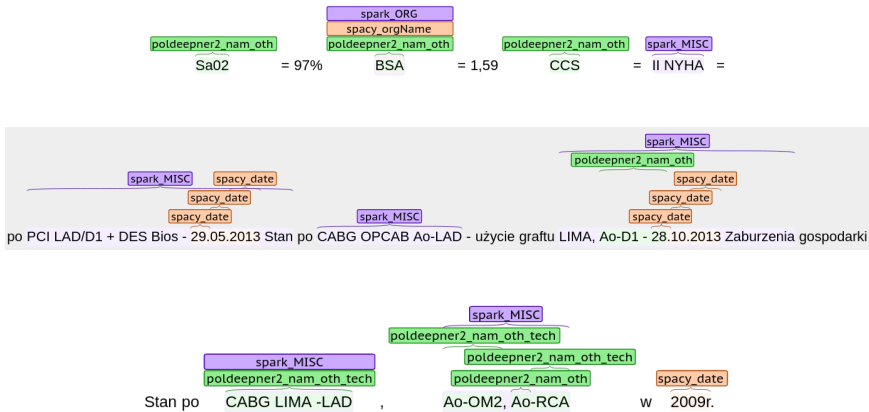


Figure: All the models feel that abbreviations are important, but without vocabularies or syntactical context, they are difficult to classify properly

PolDeepNer2: Good at personal names

Obrony mięśniowej, objawu spark_PER
spacy_persName
poldeepner2_nam_liv_person Blumberga nie stwierdza się.

spark_ORG
WATROBA nie powiększona.

spark_LOC
Śledziona, nerki niewyczuwalne.

spark_MISC
spacy_persName
poldeepner2_nam_liv_person OBJAWY Chelmonskiego nb, spark_PER
spacy_persName
poldeepner2_nam_liv_person Goldflamma nb.

Figure: Most names of people in the corpus were the names of examination signs named after inventors

PolDeepNer2: Decent at organizations

spark_MISC
poldeepner2_nam_org_organization
POChP- od ok 50 lat (w młodości praca w hucie cynku)

spark_LOC
spacy_orgName
poldeepner2_nam_org_institution
z zaleceniem systematycznej kontroli w Poradni Kardiologicznej.

spacy_orgName
poldeepner2_nam_org_institution
spark_PER spark_LOC
Pacjentka została przyjęta do Szpitala w Tychach, gdzie rozpoznano
spark_PER
spacy_orgName
poldeepner2_nam_org_institution
tutejszej Kliniki.

spark_LOC
spacy_orgName
poldeepner2_nam_org_institution
Pacjentka została przyjęta do Kliniki Chirurgii Ogólnej i Naczyni.

Figure: Names of departments and hospitals are relevant for medicine, albeit marginally

What can we expect from NER as a concept?

- General NER is looking for entities that are only partially relevant for medical science:
 - Personal names (useful for deidentification)
 - Organization names (irrelevant except for the occasional hospital department name)
 - Location names (irrelevant)
 - Miscellaneous categories depending on the model
- What medicine needs often fails to satisfy the definition of a NE:
 - Diagnosis names
 - Procedure names
 - Symptom names
 - Medicine names
 - Body part/function names
 - Measurement values

KPWr n82 model (PolDeepNer2): Fine-grained categories

- nam_liv: Living
- nam_org: Organization
- nam_fac: Facility
- nam_loc: Location
- nam_adj: Adjective
- nam_num: Numerical expression
- nam_eve: Event
- nam_pro: Product
- nam_oth: Other

PolDeepNer2: Tiny glimpses of hope

spark_LOC
spacy_orgName
Bisocard 5 mg 1/2-0-0 tabl.

spark_LOC
Inhibace 5mg 1-0-0 tabl.

spark_LOC
spacy_orgName
Tertensif SR 1,5 mg 1-0-0 tabl.

spark_LOC
poldeepner2_nam_pro_brand
Encorton 5 mg 1/2-0-0 tabl.

spark_MISC
poldeepner2_nam_pro_brand
Theovent 300 mg 1-0-1 tabl.

spark_MISC spark_MISC spark_LOC
Pulmicort 100 2 x 2 wziewy Oxis 2 x 2 wziewy Sortis 20 mg 0-0-1 tabl.

spark_PER
poldeepner2_nam_org_group_band spark_MISC
Insulina Mixtard 30 35j.-0-10j.

Figure: PolDeepNer2 identifying some medicine names as products is impressive, but it only finds a tiny fraction of the total

PolDeepNer2: Tiny glimpses of hope

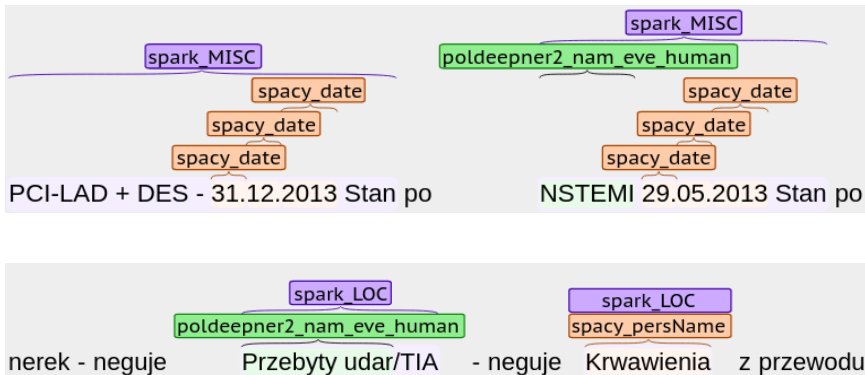


Figure: In a couple of unique cases, PolDeepNer2 identified heart attacks as events - however, recall is close to zero

Conclusions

- PolDeepNer2 could be used for de-identifying health records, but it would have to be prevented from deleting surnames in the names of medical concepts
- spaCy does identify most names but the other 50% of found entities are noise, rendering its name recognition unusable
- spaCy's identification of date/time expressions might be useful for timestamping events
- The authors of Spark NLP's Polish pipeline might need to further verify it as nothing should perform as bad as this

A valuable take-home message

Some of PolDeepNer2's categories (product, event) are a step in the right direction and suggest the possibility of tuning the existing RoBERTa model and thus creating PolDeepNer2's own MER (medical entity recognition) model.

Bibliography I

- [1] M. Honnibal et al. *spaCy: Industrial-strength Natural Language Processing in Python*. 2020. DOI: <https://doi.org/10.5281/zenodo.1212303>.
- [2] Veysel Kocaman and David Talby. “Spark NLP: Natural language understanding at scale”. In: *Software Impacts* (2021), p. 100058. ISSN: 2665-9638. DOI: <https://doi.org/10.1016/j.simpa.2021.100058>.
- [3] Michał Marcińczuk. *KPWr n82 NER model (on Polish RoBERTa base)*. 2020. URL: <http://hdl.handle.net/11321/743>.
- [4] Michał Marcińczuk and Jarema Radom. “A Single-run Recognition of Nested Named Entities with Transformers”. In: *Procedia Computer Science* 192 (2021), pp. 291–297.

Thank you for your attention!

MUNI

FACULTY

OF INFORMATICS