Recognition and tagging literary characters in Polish language

Maria Kędzierska, Małgorzata Wachulec, Aleksandra Wichrowska

Project Scope and Goal

1. Design and implement a tool for recognition and tagging literary heroes in Polish texts.
2. Prepare datasets.
3. Test different Named Entity Recognition models (Polish and multilingual).
4. Train selected models to recognize person names on prepared data.
5. Implement algorithm for names disambiguation.
6. Test coreference models for Polish.

Example of what we wanted to achieve

Panna **Izabela** zbliżyła się do **Wokulskiego** i wskazując w jego stronę
IZABELA ŁĘCKA
STANISŁAW WOKULSKI
parasolką rzekła dobitnie:

— **Floro** bądź łaskawa zapłacić temu panu. Wracamy do domu.
FLORENTYNA

— Kasa jest tu — odezwał się **Rzecki** podbiegając do panny **Florentyny**.
IGNACY RZECKI
FLORENTYNA

Wziął od niej pieniądze i oboje cofnęli się w głąb sklepu.

Panna **Izabela** z wolna podsunęła się tuż do kantorka, za którym siedział
IZABELA ŁĘCKA

Wokulski Była bardzo blada. Zdawało się, że widok tego człowieka
STANISŁAW WOKULSKI

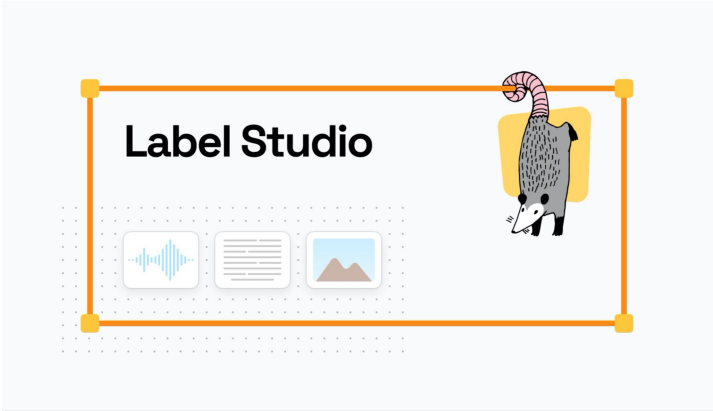
wywiera na nią wpływ magnetyczny.

Dataset

Table 1: Summary of books chosen for annotation.

Title	**Author**	**# fragments**	**# sentences**	**# heroes**
Lalka	Bolesław Prus	81	138	18
Krzyżacy	Henryk Sienkiewicz	70	111	9
Mały Książę	Antoine de Saint-Exupéry	89	144	10
Przedwiośnie	Stefan Żeromski	27	100	9
W Pustyni i w Puszczy	Henryk Sienkiewicz	24	110	9
Księga Dżungli	Rudyard Kipling	10	100	15
Robinson Crusoe	Daniel Defoe	9	225	3
Nad Niemnem	Eliza Orzeszkowa	10	192	16
Hrabia Monte Christo	Aleksander Dumas	5	378	31

Annotation Process for NER models



PERSON 2

Panna Izabela zbliżyła się do Wokulskiego i wskazując w jego stronę parasolką rzekła dobitnie:

- [ ] Stanisław Wokulski[3]
- [x] Izabela Łęcka[4]
- [ ] Ignacy Rzecki[5]
- [ ] Julian Ochocki[6]
- [ ] Tomasz Łęcki[7]
- [ ] baron Krzeszowski[8]
- [ ] baronowa Krzeszowska[9]
- [ ] Marianna[0]
- [ ] Węgiełek[q]
- [ ] Geist[w]
- [ ] Mraczewski[e]
- [ ] Jan Mincel[t]
- [ ] Franz Mincel[a]
- [ ] Małgorzata Minclowa[s]
- [ ] August Katz[d]

Final Tool Pipeline

1. Detection of all character occurrences

   Finding all *person* type entities in the text based on the NER model.

2. Disambiguation of character occurrences

   Assigning a corresponding protagonist name from the list of main characters to each person entity found, based on the similarity of strings.

3. Extending mentions of characters with pronouns and nominals

   Applying Coreference Resolution model and assigning pronouns such as *he* or *her* to proper protagonists.

NER models

Polish models

- Spacy
- PolDeepNer2

Multilingual models

- Flair
- Polyglot-NER
- Bert-base models

Initial NER results

Table 3: Performance of different NER models on annotated corpus.

Model	Precision	Recall	F-measure	Time
spacy pl_core_news_sm	0.841	0.664	0.742	18s
spacy pl_core_news_md	0.866	0.72	0.786	20s
spacy pl_core_news_lg	0.883	0.847	0.865	20s
PolDeepNer cen-n82-base	**0.928**	0.787	0.852	11m 43s
PolDeepNer cen-n82-large	0.915	0.783	0.844	34m 59s
PolDeepNer kpwr-n82-base	0.913	**0.867**	**0.89**	11m 22s
PolDeepNer nkjp-base	0.882	0.849	0.865	11m 41s
PolDeepNer nkjp-base-sq	0.905	0.847	0.875	11m 27s
spacy xx_ent_wiki_sm	0.674	0.432	0.527	5s
flair ner-multi-fast	0.736	0.603	0.663	5m 0s
flair ner-multi	0.734	0.751	0.742	15m 43s
POLYGLOT-NER	0.832	0.662	0.737	2m 53s
bert-base multilingual	**0.851**	**0.803**	**0.826**	1m 34s
distilbert-base multilingual	0.809	0.78	0.794	1m 0s

Fine-tuned NER Models

```
Model: pl_core_news_lg
Iteration 0, Losses{'ner': 629.5366953036838}
Iteration 1, Losses{'ner': 300.6115595964211}
Iteration 2, Losses{'ner': 244.62630068183725}
Iteration 3, Losses{'ner': 212.49833131660503}
Iteration 4, Losses{'ner': 164.68816731173462}
Iteration 5, Losses{'ner': 220.03374603932957}
Iteration 6, Losses{'ner': 144.87891197119836}
Iteration 7, Losses{'ner': 162.04671940227394}
Iteration 8, Losses{'ner': 125.07547913064748}
Iteration 9, Losses{'ner': 127.81112915900931}
```

Table 6: Performance of fine-tuned NER models on test set from annotated corpus.

Model	Precision	Recall	F-measure
spacy pl_core_news_sm	0.934	0.786	0.854
spacy pl_core_news_sm_finetuned_3_epochs	0.696	0.553	0.617
spacy pl_core_news_sm_finetuned_10_epochs	0.664	0.46	0.544
spacy pl_core_news_md	0.931	0.814	0.868
spacy pl_core_news_md_finetuned_3_epochs	0.678	0.628	0.652
spacy pl_core_news_md_finetuned_10_epochs	0.714	0.628	0.668
spacy pl_core_news_lg	0.933	0.777	0.848
spacy pl_core_news_lg_finetuned_3_epochs	0.685	0.628	0.655
spacy pl_core_news_lg_finetuned_10_epochs	0.702	0.623	0.66
spacy xx_ent_wiki_sm	0.712	0.437	0.542
spacy xx_ent_wiki_sm_finetuned_3_epochs	0.779	0.526	0.628
spacy xx_ent_wiki_sm_finetuned_10_epochs	0.703	0.572	0.631
bert-base multilingual	0.903	0.823	0.861
bert-base multilingual_finetuned	**0.932**	**0.833**	**0.88**

Best performing model - NER results

Table 4: Performance of *kpwr-n82-base* model from `PolDeepNer` lilbrary on annotated corpus.

Novel title	Precision	Recall	F-measure	Support
Hrabia_Monte_Christo	0.826	0.92	0.871	212
Ksiega_dzungli	0.932	0.687	0.791	99
Maly_Ksiaze	0.8	0.3	0.436	40
Przedwiosnie	0.982	0.966	0.974	116
W_pustyni_i_w_puszczy	0.922	0.881	0.902	135
Krzyzacy	0.88	0.82	0.849	89
Lalka	0.989	0.968	0.978	93
Nad_Niemnem	0.962	0.962	0.963	80
Robinson_Crusoe	0.97	0.915	0.942	71
** overall results ***	0.913	0.867	0.89	935

Best performing model - Protagonist Tagger results

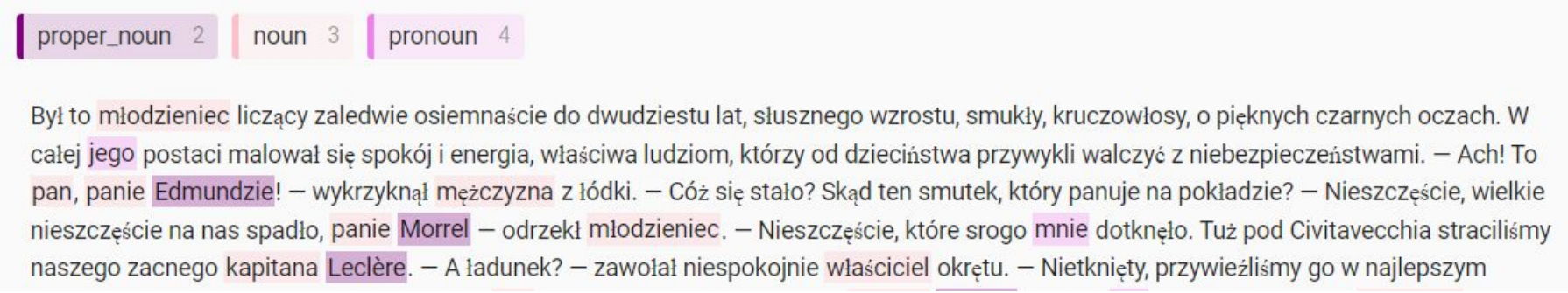
Table 7: Performance of protagonistTagger based on *kpwr-n82-base* model from `PolDeepNer` library on annotated corpus.

Novel title	Precision	Recall	F-measure
Hrabia_Monte_Christo	0.818	0.91	0.862
Ksiega_dzungli	0.904	0.667	0.767
Maly_Ksiaze	0.733	0.275	0.4
Przedwiosnie	0.825	0.81	0.817
W_pustyni_i_w_puszczy	0.682	0.652	0.667
Krzyzacy	0.819	0.764	0.791
Lalka	0.956	0.935	0.946
Nad_Niemnem	0.862	0.862	0.862
Robinson_Crusoe	0.97	0.915	0.942
** overall results ***	0.834	0.793	0.813

Annotation Process for Coreference

Additional annotation for two novels:

1. The Count of Monte Christo - longer chunks of continuous text
2. The Jungle Book - shorter chunks and animals as heroes



Testing Coreferee model from Spacy

1. Coreferee still has a lot to improve
2. For now, this is not a reliable method of identifying coreferences for the Polish language.

Table 8: Statistics for coreference testing.

Title	# true annotations	# model annotations	# missing annotations	# correctly annotated	# annotations with wrong labels	# completely wrong annotations
Księga Dżungli	190	164	136	41	13	110
Hrabia Monte Christo	285	285	166	72	47	166

Testing Coreferee model from Spacy

Percent of correctly annotated parts of speech