NLP Project 1 Report

Literary Heroes Detections in Polish Texts

Maria Kedzierska, 305704
Małgorzata Wachulec, 293790
Aleksandra Wichrowska, 291140

3 April 2022

1 Project Scope and Goal

This project aims to design and implement a tool for recognition and tagging persons in natural language texts - literary heroes in novels or other long texts. The project focuses on Polish texts. This work comprises preparing datasets (preferably in an automatic way), testing different Named Entity Recognition (NER) models for Polish and then training a selected machine learning model to recognize and disambiguate person names on prepared data. We also would like to compare our results to these obtained for English texts [1], [2].

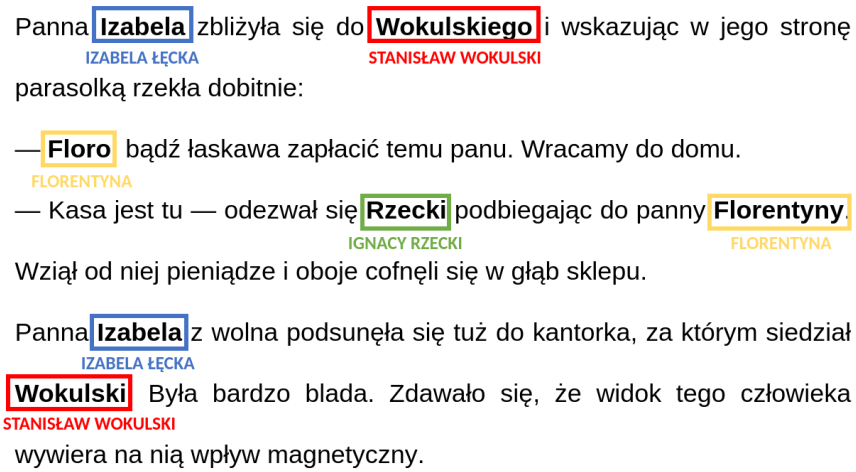

Figure 1: Expected results of the project.

2 Literature Review

2.1 Named Entity Recognition Models for Polish

On popular NLP progress tracking websites such as `http://nlpprogress.com` we can see that not many of newest NLP tasks are developed for Polish language. This is also true for the Named Entity Recognition and linking (or disambiguation) models. Some model, e.g. `flair`, developers were planing to pretrain NER models on Polish texts, but ended up delivering pre-trained embeddings instead [3]. In fact, there are many available embeddings for Polish words pretrained using continuous bag of words or skipgram methods on the National Corpus of Polish Langugage (`http://nkjp.pl`), which can be found on the website of the Institute of Computer Science of the Polish Academy of Sciences: `http://dsmodels.nlp.ipipan.waw.pl/w2v.html`.

So far we have found the following NER models for the Polish language:

- models from `spacy` library, i.e. *pl_core_news_lg*
- PolDeepNer (`https://github.com/CLARIN-PL/PolDeepNer`) - model trained during PolEval 2018 competition. It is able to recognized person names, surnames and additional names separately.
- PolDeepNer2 - an improved version of the previous model
- `https://github.com/applicaai/poleval-2018` - another model from PolEval 2018

The authors of the English counterpart of this project annotated news texts and claimed good results were achieved even without the fine tuning the model [2]. They underline, however, that news are characterized by simpler and more straightforward sentence structures which improved their annotations.

2.2 Detecting Characters in Literary Texts

The majority of entity linking models concentrate on news texts and are not suitable for literary texts such as novels with veritable sentences [4]. One of the projects focused specifically on literary texts, called GutenTag [5],[6], is able to filter the corpus of novels available through Gutenberg project [7] and generate statistics regarding these novels. The functionalities of this tool include part-of-speech annotation and also annotation of main characters specified by a user. The annotations obtained in such way are used solely for retrieval of statistical measures, served as an output of the application.

There are two main approaches to entity linking, both applied after named entity recognition has already been performed. One is based on character clustering [8], [9], where entities found by NER model are clustered into groups referring to the same person. The results presented in these articles were qualitative, rather than quantitative. The second one is based on graphs [10], where entities found by NER model are represented by the nodes and if these entities are likely to describe the same person they are connected with an edge. This creates a hazard of joining nodes corresponding to two characters with an edge and hence not being able to distinguish between the two. This is why this approach includes a second step which ensures

that for each 2 nodes for which a path exists e.g. the gender of the entities represented by these nodes is the same. Thanks to this the graph approach achieves up to 75% of accuracy in character detection.

Finally, the English counterpart of this project [1], [2] introduces the protagonistTagger tool for recognition and disambiguation of person entities (entity linkage) in literary texts and an entity linkage benchmark dataset that was manually annotated, on which the protagonistTagger achieved over 83% of both precision and recall. This exceeds the aforementioned approaches, but is not suitable for entity recognition and linking for Polish literary texts, a topic which we chose to explore.

3 Dataset Preparation

According to our knowledge, there is no dataset containing literary texts in polish with annotated names of protagonists. So we decided to prepare such dataset by ourselves. We've chosen 9 books from `wolnelektury.pl`. For each of them we conducted following steps:

1. downloading full novel text from the website
2. selection of interesting parts of text to annotate - some paragraphs with heroes and actions descriptions, as well as dialogues between protagonists
3. preparing lists of most important heroes in given novel
4. using existing version of *protagonistTagger* tool to pre-annotate chosen sentences; as model we use *pl_core_news_lg* from `spacy` library
5. correcting annotations in `LabelStudio` project

In the Table 1 there is info about chosen books - title, author, number of chosen fragments, number of sentences and number of defined heroes.

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 Ksiaże	Antoine de Saint-Exupéry	89	144	10
Przedwiośnie	Stefan Żeromski	27	100	9
W Pustyni i w Puszczy	Henryk Sienkiewicz	24	110	9
Ksiega 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

Table 2: Statistics of errors in pre-annotated data

Title	# pre-annotations	# final annotations	# correct annotations	# missing annotations	# annotations with wrong hero assigned	# annotations with wrong boundaries	# completely wrong annotations
Lalka	95	93	87	2	3	1	4
Krzyżacy	101	93	69	10	5	9	18
Mały Ksiaże	12	41	11	30	0	0	1
Przedwiośnie	120	116	95	2	16	3	6
W Pustyni i w Puszczy	128	135	87	16	32	0	9
Ksiega Dżungli	61	99	56	40	2	1	2
Robinson Crusoe	78	71	64	7	0	0	14
Nad Niemnem	85	80	69	5	6	0	10
Hrabia Monte Christo	151	145	122	21	2	0	27

4 Exploratory Data Analysis

In the Table 2 there are statistics about errors in pre-annotated data which were corrected in LabelStudio.

Table 2 shows the quantitative analysis of our dataset, but for future project development it is crucial to understand why certain entities were linked or disambiguated falsely. For this reason we also include a qualitative analysis of this dataset. While annotating the novels we saw the following sources of errors:

1. Some names of African tribes weren't recognized as PERSON in the *In Desert and Wilderness* (Polish: W pustyni i w puszczy) although they play the same function as surnames for other characters.

2. Model doesn't recognize diminutives as proper names hence fails to annotate them. This is true for both English names, e.g. Nel was annotated correctly, but Nelly wasn't, and for Polish names, e.g. Cezary was annotated but Czaruś wasn't.

3. For characters having the same surnames often wrong hero was assigned.

4. Animal names in novels were animals are NOT the main characters were classified as PERSON, e.g. dog named Saba would be recognized as PERSON. We consider this a mistake.

5. In novels were animals are the majority of main characters or have human-like attributes, e.g. talk, the model failed to annotate them. In *The Jungle Book* (Polish: Ksiega dżungli) characters are called Mother Wolf (Polish: Matka Wilczyca), Grey Brother (Polish: Szary Brat). This is problematic as standard NER models won't annotate mother or brother as character's name or surname, however in this case they would need to do that to annotate these characters correctly. Fixing this issue and the previous one might be challenging as they are somewhat contradictory.

6. Sometimes at the beginning of a sentence, due to the capital letter the model treated referencing to protagonist by marital status or by occupation as PERSON e.g. Widow, Captain, which is generally correct. However in this project we aim to narrow the scope of recognizing as PERSON to reference by first name or surname.

7. Similar situation to the aforementioned occurs with reference by nationality or ethnicity e.g. Catalan(Polish: Katalonka) is classified as a PERSON, which is an unwanted situation.

These are the main challenges we will face while developing this project. The ideas on how to improve the initial performance of the model will be described in next section about solution concept.

5 Solution Concept

Similarly to the English counterpart of our project [1], [2], our solution will consist of a NER model that filters out PERSON entities from the text and an annotator that links the found PERSON entities to main characters that will be passed as arguments. It is possible that no character will be assigned to the entity if this is a side-character. We will measure the quality of prediction for both entity recognition and entity linking stage, using common metrics such as precision, recall and F-score.

Our intention is to test different Polish and multilingual models to compare their performance. As underlined in the literature review, many of recent NLP tools are not trained specifically on Polish texts, so we are forced to use the multilingual ones. By comparing Polish-specific (such as spacy pl_core_news_lg model trained on news texts) to multilingual models (such as flair model) we will see how much using multilingual models downgrades our results. This topic is interesting from research point of view as seeing that language-specific models outperform multilingual ones is a great motivation for training more language-specific models in the future.

To overcome the second limitation listed in the previous section we consider using lists of name diminutives (short versions of names) from websites like `https://www.ksiegaimion.com`. The English counterpart of this project used lists of gender specific titles such as duke, miss, governor, etc. to better disambiguate characters of 19th century British novels. Polish novels don't use so many titles, but we believe using lists of names diminutives could improve the model significantly as Polish language is particularly rich in those, one name can have several short versions not all of which will be found by approximate text matching.

The English counterpart of this project used sentences drawn at random from novels. This means that every sentence had to be considered separately and no information regarding characters in that sentences could be inferred from previous phrases. We decided to add entire passages of text into our dataset, without changing the order of sentences. Thanks to that we can experiment with how the context of previous phrases

improves the quality of annotation. This could be a factor improving the annotations for characters with same surname.

The main aim of this project is to accurately recognize and link (disambiguate) characters in longer Polish texts. The ideas described above are directions we will explore rather than strict guidelines. We believe working on this project will be an iterative result-driven process where we test new approaches and decide whether they are worth further exploration or whether we should follow another path instead. Nevertheless, in the future reports we will include summaries of our exploration and rationale behind our decisions.

References

[1] W. Łajewska and A. Wróblewska, "Protagonists' tagger in literary domain – new datasets and a method for person entity linkage," 2021.

[2] W. Łajewska and A. Wróblewska, "ProtagonistTagger – a tool for entity linkage of persons in texts from various languages and domains," 2022.

[3] "Polish language support," 2018.

[4] T. Stanislawek, A. Wróblewska, A. Wójcicka, D. Ziembicki, and P. Biecek, "Named entity recognition - is there a glass ceiling?," 2019.

[5] J. Brooke, A. Hammond, and G. Hirst, "Gutentag: an nlp-driven tool for digital humanities research in the project gutenberg corpus," 2015.

[6] A. Hammond and J. Brooke, "Gutentag: A user-friendly, open-access, open-source system for reproducible large-scale computational literary.," 2017.

[7] M. Hart, "The history and philosophy of project gutenberg," 1992.

[8] D. Elson, N. Dames, and K. McKeown, "Extracting social networks from literary fiction.," 2010.

[9] D. Bamman, T. Underwood, and N. A. Smith, "A bayesian mixed effects model of literary character," 2014.

[10] H. Vala, D. Jurgens, A. Piper, and D. Ruths, "Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts.," 2015.