NLP Project 1 Report

Literary Heroes Detections in Polish Texts

Maria Kędzierska, 305704

Małgorzata Wachulec, 293790

Aleksandra Wichrowska, 291140

Anna Wróblewska

6 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 (in n semi-automatic way), testing different Named Entity Recognition (NER) models for Polish, training a selected model to recognize person names on prepared data and implementing algorithm for names disambiguation. We also would like to compare our results to these obtained for English texts [1], [2].

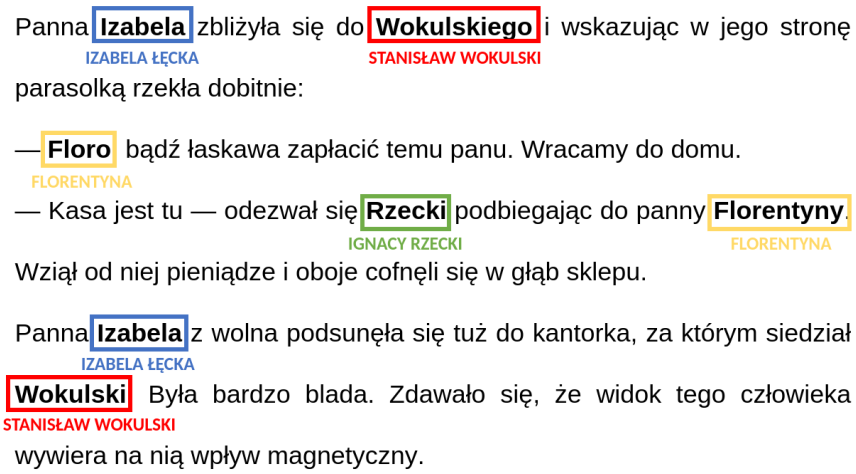

Figure 1: Expected results of the project.

2 Literature Review

2.1 Named Entity Recognition Models for Polish

On NLP progress tracking websites such as `http://nlpprogress.com` there are not many entries for Polish language. Most of the newest NLP tasks are not developed in Polish. 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`.

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.

There are few available NER models for Polish language. Below, we go into details of most interesting of them - what is the architecture of the model, on which data it is trained, etc.

Models from `spacy` library [4]
`Spacy` is one of the most popular NLP libraries in Python. It supports many language, including Polish. There are available three models - *pl_core_news_sm*, *pl_core_news_md* and *pl_core_news_lg*. All of them were trained mostly on news data from three different sources:

- UD Polish PDB v2.8 [5]
- National Corpus of Polish [6]
- PoliMorf [7]

Models shared by `spacy` are deep convolutional neural networks with residual connections. They are using *Bloom* embeddings.

PolDeepNer [8]
This model was trained during PolEval 2018 competition and achieved second place. It is able to recognize person names, surnames and additional names separately. Model was trained on the NKJP corpus. Hierarchy of classes included in this corpus is presented in Figure 2.

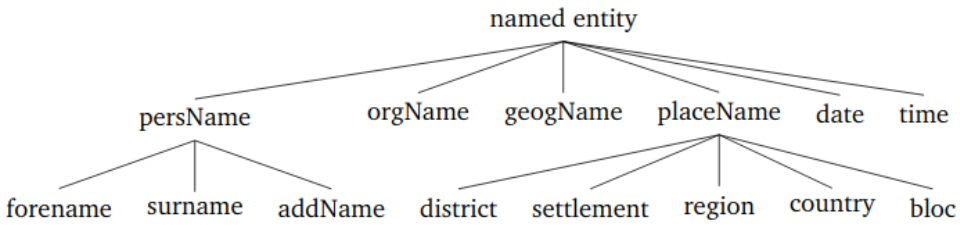

Figure 2: Hierarchy of classes annotated in NKJP corpus [6].

Main architecture of the model is an ensemble of three neural networks - one BiLSTM and two BiGRU. Ensemble uses majority voting for computing final output. Word embeddings for each model were trained using `fastText` library but with different training sets and sizes of word vectors:

- Model1 - BiGru - trained on Common Crawl [1] and Wikipedia [2]; dimension of vectors is 300.
- Model2 - BiGru - trained on KGR10 corpus [9] (data from the Internet); only words with minimum 50 occurrences in the corpus were used; dimension of vectors is 300.
- Model3 - BiLSTM - trained on KGR10; only words with minimum 5 occurrences in the corpus were used; dimension of vectors is 100.

Processing pipeline is shown in Figure 3. Each model consists of input layer, dropout layers, bidirectional GRU (or LSTM respectively), dense layer and CRF layers.
There is also available second version of this project - PolDeepNer-2 [3] which authors claim to be an improved version of the first one.

2.2 Detecting Characters in Literary Texts

Following [10], there are three main types of mentions of protagonists in the literature. There are:

- **proper nouns** - names of specific people. They usually start with a capital letter. For example *Stanisław Wokulski*, *Mały Książę* or *Staś*.
- **nominals** - anaphoric noun phrases which identify characters. For example *the boy* or *this man*.
- **pronouns** - for example *he* or *her*.

According to [11] about 75% of protagonists mentions in literature are pronouns. However, proper nouns mentions are the most important source of information about the novel, and are often sufficient. The

[1] `https://www.commoncrawl.org/`
[2] `https://www.wikipedia.org`
[3] `https://github.com/CLARIN-PL/PolDeepNer2`

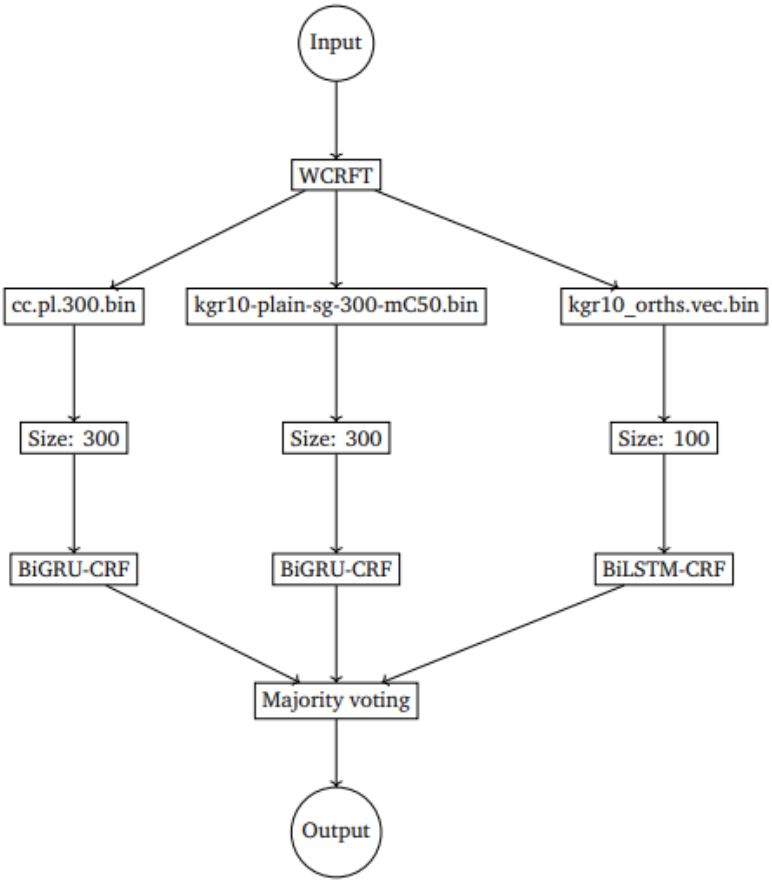

Figure 3: Processing pipeline used in PolDeepNer model [8].

detection of the place of occurrence of the names of the main characters allows, among other things, to create a social network of characters ([12]), analyze relationships between characters ([13]), detect roles ([14]) or create protagonist characteristics ([15]).

There are a lot of researches aiming to detect character mentions in the text and classify them with proper protagonist names. Most of them is focused only on detecting proper nouns - this is typically based on NER models ([16], [12]). In [17] the authors try to extend detected mentions with pronouns and nominals. Majority of researches is working on English texts, there is one paper about German novels ([18]). According to our knowledge, there is no publication about detecting protagonists in Polish literature.

Having detected PERSON entities by NER model, the next step is entity linking. The majority of entity linking models concentrate on news texts and are not suitable for literary texts such as novels with veritable sentences [19]. One of the projects focused specifically on literary texts, called GutenTag [20], [21], is able to filter the corpus of novels available through Gutenberg project [22] 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 [23], [24], 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 [25], 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. In protagonistTagger entity linking step is a rule-based algorithm. In a nutshell, found entities are compared to the characters from predefined list and string similarity is computed. Then, for the matches (characters with similarity score higher than the threshold) gender and personal title agreements is checked. Character with highest score is returned as the final result.

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`. The books have been selected to represent different genres, come from different eras, and present interesting issues in recognizing characters. For example, we expect that in *The Jungle Book* (Polish: Księga Dżungli) or *The Little Prince* (Polish: Mały Książę), the NER models will have a problem detecting the main characters (who are not human). In the *The Teutonic Knights* (Polish: Krzyżacy), we can find defining characters specific to the knighthood, for example, Zbyszko z Bogdańca.

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. [4] We assumed 100 sentences per book as the minimum size of the set.

3. Preparing lists of most important heroes in given novel. [5]

4. Using existing version of *protagonistTagger* tool to pre-annotate chosen sentences. As model we use *pl_core_news_lg* from `spacy` library. In disambugation step only string similarity was taken into account.

5. Correcting annotations in `LabelStudio` project - in Figure 4 there is an example view on annotation interface.

In the Table 1 there are quantitative details 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 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

4 Exploratory Data Analysis

Firstly, we will take a look at the general properties of our datasets in Figure 5.

In attached notebook, we have plotted the number of occurrences of each protagonist for each novel, e.g. Figure 6.

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

[4] Due to the small size of the dataset, this step was performed manually

[5] We couldn't find any source for scraping lists of main characters. Therefore, and because of the small dataset size, this step was performed manually.

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 Książę	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
Księga 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

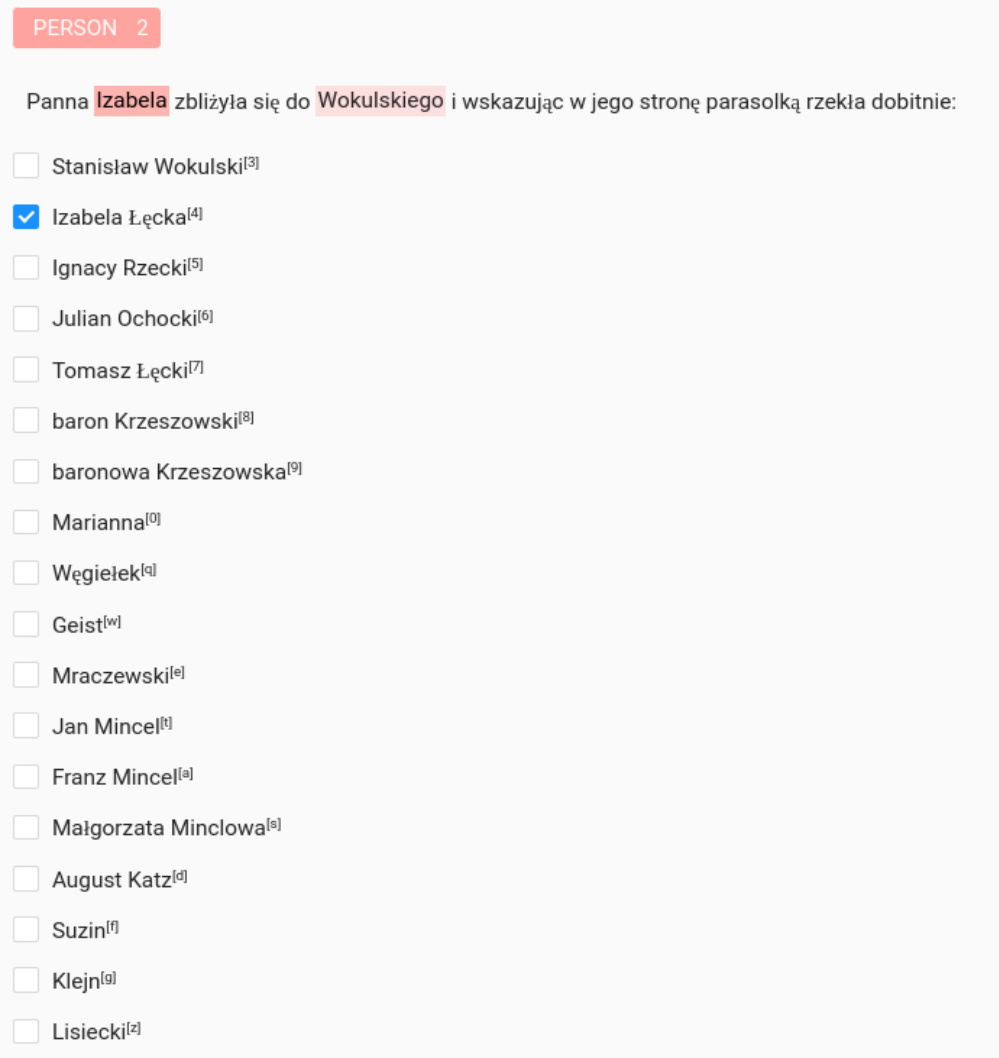

Figure 4: An example view of annotation interface.

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, as shown in Fig. 7.

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 (as shown in Fig. 8), but Nelly wasn't (as shown in Fig. 9), and for Polish names, e.g. Cezary was annotated but Czaruś wasn't.

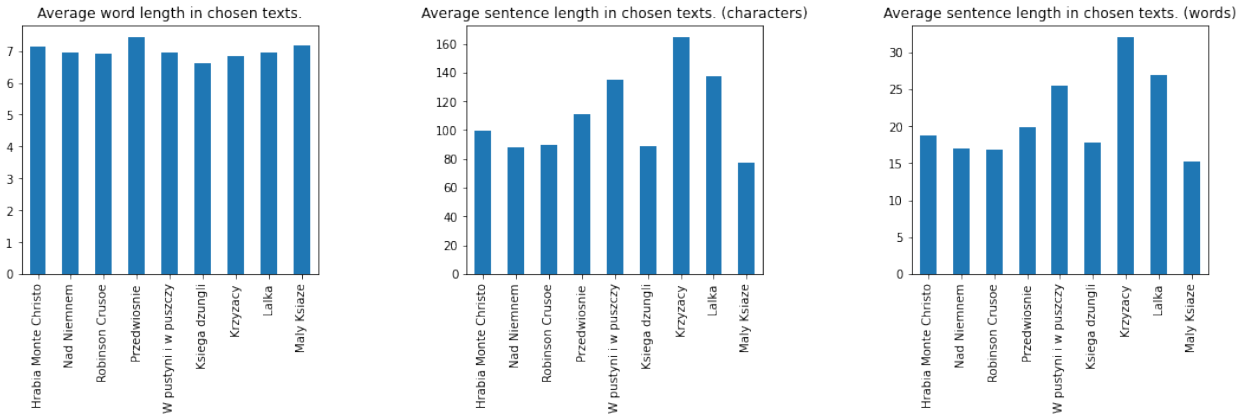

Figure 5: General statistics for all novels.

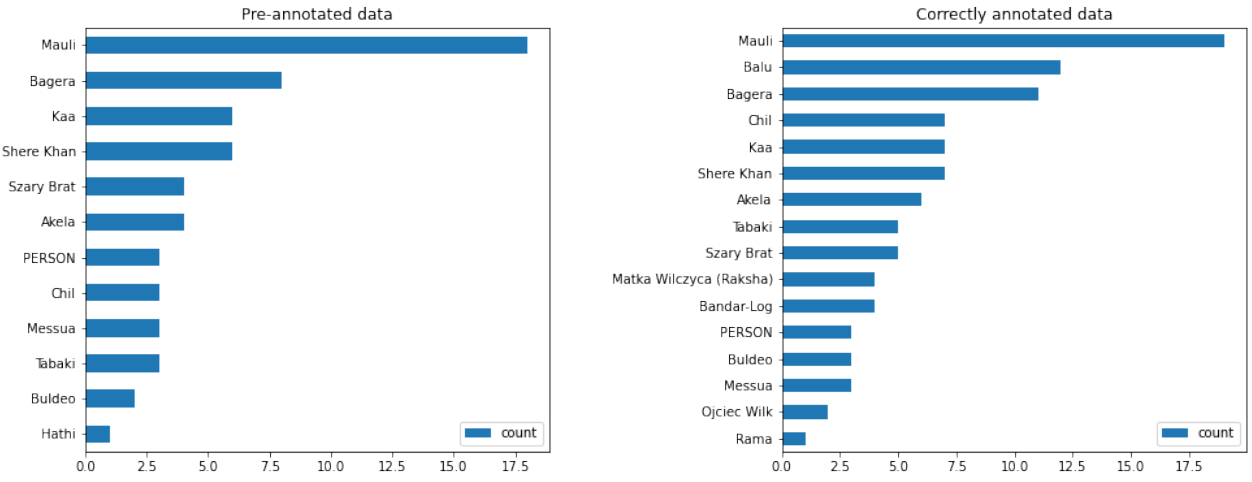

Figure 6: Number of occurrences of characters' annotations in *The Jungle Book*.

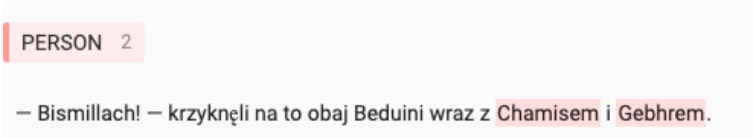

Figure 7: Beduini not found by the model.

Figure 8: Nel assigned correctly by the model.

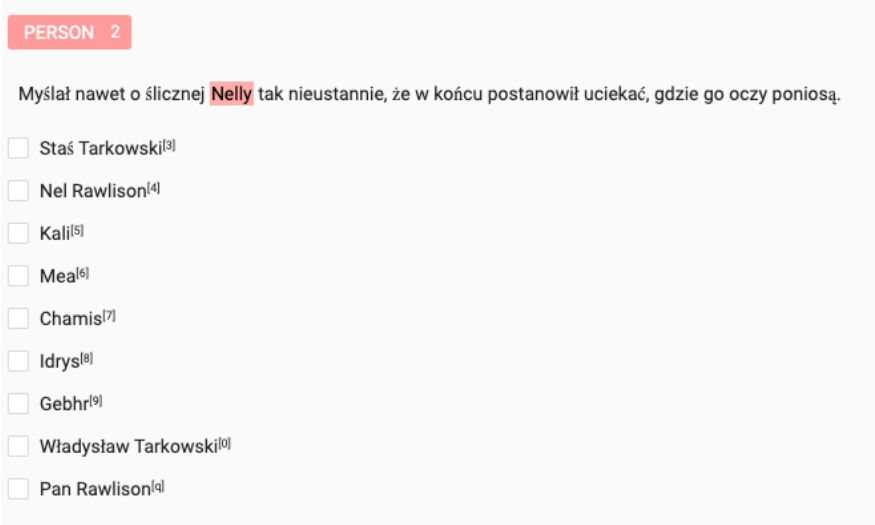

Figure 9: Nelly identified as PERSON but not assigned to the proper character.

3. For characters having the same surnames often wrong hero was assigned as shown in Fig. 10 and 11.

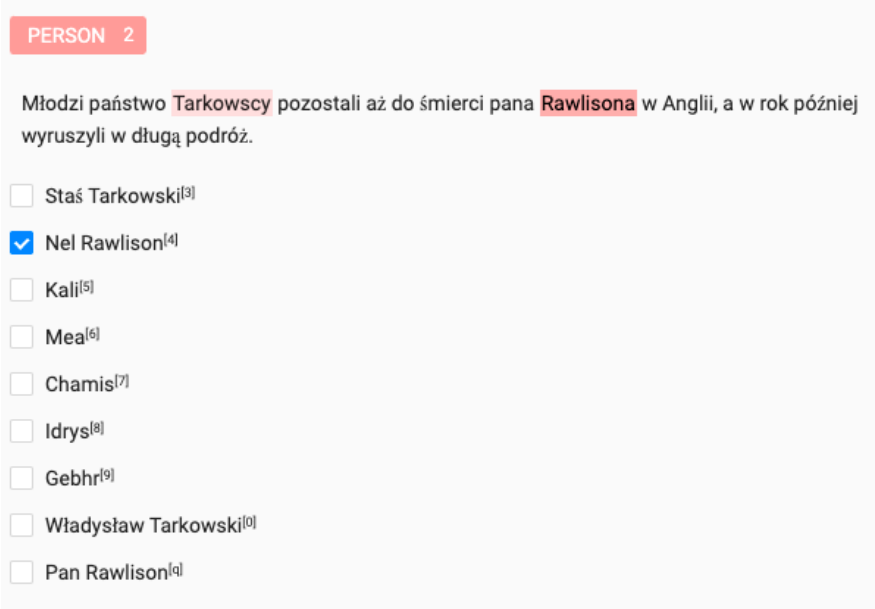

Figure 10: Mr Ravilson mistaken for Nel Rawilson.

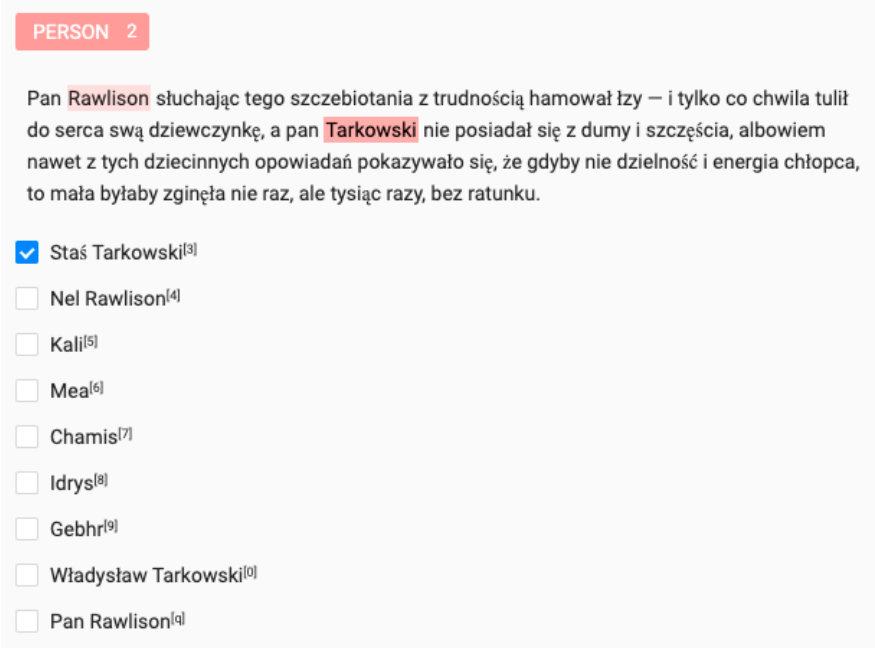

Figure 11: Mr Tarkowski mistaken for Staś Tarkowski.

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. This example is shown in Fig. 12.

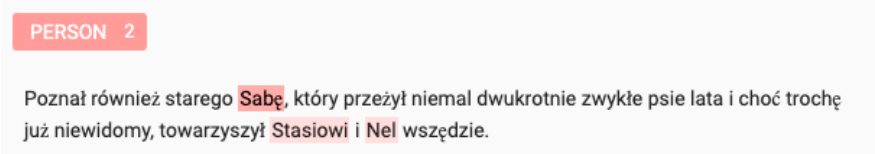

Figure 12: Saba - a dog, classified as PERSON by the NER model. False positive.

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: Księga 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, as shown in Fig. 13, 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.

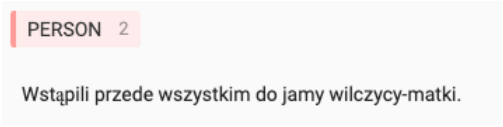

Figure 13: Mother Wolf is not recognized as porper name by the model.

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.

Preparing good, literature-oriented NER model will be very important part of our task. We would like to overcome limitations listed in section 4. This will probably involve testing many NER models, re-training selected on the annotated data (including cases where protagonist is non-human being). Probably training model on artificially prepared dataset (for example by replacing common names with rare ones) will be beneficial. If necessary, we can also test approaches with applying some additional rules, i.a. in [12], [26] difficult to detect heroes are found by looking for speech verbs or possessive forms.

Our first 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 *pl_core_news_lg* model from `spacy` 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 (probably *Ola* would not be found as string similar to *Aleksandra*).

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] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." To appear, 2017.

[5] A. Wróblewska, "Extended and enhanced polish dependency bank in universal dependencies format," in *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)* (M.-C. de Marneffe, T. Lynn, and S. Schuster, eds.), pp. 173–182, Association for Computational Linguistics, 2018.

[6] A. Savary, J. Waszczuk, and A. Przepiórkowski, "Towards the annotation of named entities in the National Corpus of Polish," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, (Valletta, Malta), European Language Resources Association (ELRA), May 2010.

[7] M. Woliński, M. Miłkowski, M. Ogrodniczuk, A. Przepiórkowski, and Szałkiewicz, "PoliMorf: a (not so) new open morphological dictionary for Polish," pp. 860–864.

[8] M. Marcińczuk, J. Kocoń, and M. Gawor, "Recognition of named entities for polish-comparison of deep learning and conditional random fields approaches," in *Proceedings of the PolEval 2018 Workshop* (M. Ogrodniczuk and Kobyliński, eds.), pp. 77–92, Institute of Computer Science, Polish Academy of Science, 2018.

[9] J. Kocon and M. Gawor, "Evaluating KGR10 polish word embeddings in the recognition of temporal expressions using bilstm-crf," *CoRR*, vol. abs/1904.04055, 2019.

[10] D. K. Elson, *Modeling Narrative Discourse*. PhD thesis, Columbia University, 2012.

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

[12] M. Coll Ardanuy and C. Sporleder, "Structure-based clustering of novels," in *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, (Gothenburg, Sweden), pp. 31–39, Association for Computational Linguistics, Apr. 2014.

[13] S. Chaturvedi, M. Iyyer, and H. D. III, "Unsupervised learning of evolving relationships between literary characters," AAAI'17, p. 3159–3165, AAAI Press, 2017.

[14] A. Groza and L. Corde, "Information retrieval in folktales using natural language processing," in *2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 59–66, 2015.

[15] M. Elsner, "Character-based kernels for novelistic plot structure," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, (Avignon, France), pp. 634–644, Association for Computational Linguistics, Apr. 2012.

[16] 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," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (Lisbon, Portugal), pp. 769–774, Association for Computational Linguistics, Sept. 2015.

[17] H. Vala, S. Dimitrov, D. Jurgens, A. Piper, and D. Ruths, "Annotating characters in literary corpora: A scheme, the CHARLES tool, and an annotated novel," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, (Portorož, Slovenia), pp. 184–189, European Language Resources Association (ELRA), May 2016.

[18] L. Hettinger, M. Becker, I. Reger, F. Jannidis, and A. Hotho, "Genre classification on german novels," pp. 249–253, 09 2015.

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

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

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

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

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

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

[25] 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.

[26] M. Trovati and J. Brady, "Towards an automated approach to extract and compare fictional networks: An initial evaluation," 12 2014.