

FedCSIS_2023_paper_5154.pdf

ProtagonistTagger – Person Entity Linkage Tool

⁷ Weronika Łajewska
Warsaw University of Technology
Warsaw, Poland
Email: weronikalajewska@gmail.com

Anna Wróblewska
⁷ 0000-0002-3407-7570
Warsaw University of Technology
Warsaw, Poland
Email: anna.wroblewska1@pw.edu.pl

Abstract—Named entity linkage (NEL), which is a combination of named entity recognition (NER) and disambiguation (NED), can add semantic context to the recognized named entities in texts. It links the entities mentioned in unstructured texts and individual instances of real-world objects. This paper presents a tool – *protagonistTagger* – for *person* NER and NED in texts. The tool was tested on diversified texts extracted from classic English novels and Polish Internet news. The tool's performance (F1-score) ranges between 78% and even 89%, depending on the dataset.

¹ INTRODUCTION

EXTRACTING, integrating and matching instances of named entities referring to the same real-world objects still remains a big challenge in Natural Language Processing and the Semantic Web. Nevertheless, this task is necessary to achieve the understanding and integrity of various resources. One of the most basic and standard categories of named entities appearing in almost every type of text is referring to people. Named entity recognition (NER) models can provide persons' mentions, which are undifferentiated annotations tagged with general label *person*. To analyze text semantics or find shared mentions between various texts, we need contradistinction between the recognized *person* named entities. The most desired way is to have a unique identifier for each person appearing in the considered dataset. Ideally, each person's mention (given in different forms, e.g. diminutive, only surname or first name) in the dataset should be linked with a tag containing the full name of the corresponding person.

Researchers often concentrate only on one aspect of identifying the person mentions. There are available NER models, particularly for news and common language texts. Search engines can search through particular names and match distinct names each time we query them. However, it is complicated to recognize and match mentions of specific people with named entities in an arbitrary text using a single tool that combines the recognition and identification of persons in various domains.

The general workflow of the created tool – called *protagonistTagger* – is presented in Figure 1. It works automatically with a list of peoples' full names (as predefined tags) and a text to be annotated and given as inputs. The process of *person* entity linkage implemented in *protagonistTagger* is divided into two main phases:

- 1) named entity recognition (NER) of mentions of people in a text.

- 2) named entity disambiguation (NED) providing links between the entities recognized in unstructured text and identifiers referring to real-world objects (our predefined tags given as a list of persons' full names).

Table I shows an example of an annotated text – the final result made with *protagonistTagger*. The tool's good performance, verified on two different testing sets, proves that *protagonistTagger* can be used successfully to recognise and identify people in complex texts. The tool and the corpus can help detect and analyse the relationships between persons protagonists in novels, known personalities in public news) and create a social semantic network. Furthermore, the fine-tuned NER model and the whole tool can be applied to texts from a non-literary domain containing named entities of the category *person*, as long as a list of full names to be linked with them is available.

Summing up, this research's main contributions include:

- verification of the standard NER model performance in the literary domain and implementing a method for preparing training sets for model fine-tuning,
- the *protagonistTagger* tool for recognition and disambiguation of *person* entities (entity linkage) in literary English texts and an example of its adaptation to a different domain and language (Polish Internet news),
- entity linkage (NER & NED) benchmark testing datasets for entities of category *person* manually annotated also with full names of literary characters (1,300 sentences each), and also revised dataset for entity linkage in Polish Internet news domain (about 1,000 sentences annotated with full names of Polish personalities, e.g. politicians, actors),
- a vast corpus of novels annotated with *protagonistTagger* (13 novels, more than 50,000 sentences and more than 35,000 mentions of literary characters).

In the following, we provide an overview of the related works (Section II), then a general description of our approach in the English literary domain (Section III), and an adaptation to a different domain – Internet news with texts in Polish (Section IV).¹ The prepared dataset and our approach performance are described in Section V. The paper completes with conclusions and future works (Section VI).

¹The *protagonistTagger* tool, benchmark datasets from literary domain and annotated corpus are available, along with documentation and manual, at <https://doi.org/10.5281/zenodo.4699418>, and the adaptation – tool along with the new datasets are at <https://zenodo.org/record/5060232>.

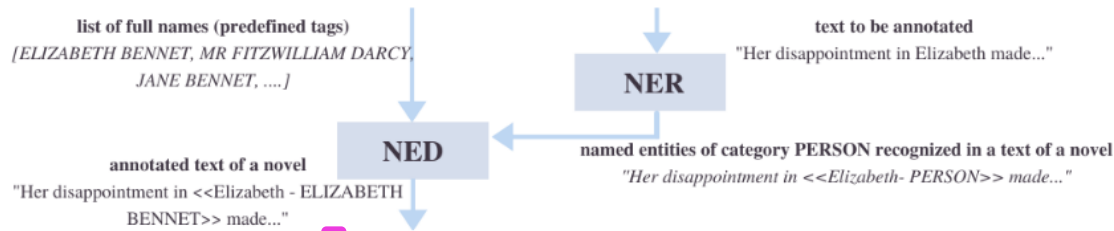


Fig. 1. Simplified process employed in our tool – *protagonistTagger*.

"Her disappointment in **Charlotte**_{Charlotte Lucas} made her turn with fonder regard to her sister, of whose rectitude and delicacy she was sure her opinion could never be shaken, and for whose happiness she grew daily more anxious, as **Bingley**_{Charles Bingley} had now been gone a week and nothing more was heard of his return. **Jane**_{Jane Bennet} had sent **Caroline**_{Caroline Bingley} an early answer to her letter and was counting the days till she might reasonably hope to hear again. The promised letter of thanks from **Mr. Collins**_{Mr. William Collins} arrived on Tuesday, addressed to their father, and written with all the solemnity of gratitude which a twelve month's abode in the family might have prompted."

TABLE I

AN EXAMPLE OF TEXT EXTRACTED FROM NOVEL *Pride and Prejudice* BY JANE AUSTEN. EACH ENTITY RECOGNIZED as *person* IS LINKED TO THE PROPER FULL NAME (TAGS FROM THE INPUT LIST); THEY ARE WRITTEN IN A SUBSCRIPT.

II. RELATED WORK

Named entity linkage (NEL) is a process of recognition (NER) and disambiguation (NED) of named entities in a text. The general conclusion can be drawn that standard NER and NED models can be used successfully on dissected datasets. However, they may have lower performance working in the pipeline and on different text domains, e.g. on novels containing veritable sentences [1].

NER can be approached in several different ways [2]. The most recent approaches are designed using mainly neural networks. They do not require domain-specific resources or feature engineering, thanks to word embeddings used as feature vectors [3]. Neural network-based NER systems can be classified into several groups depending on their representation of the words in a sentence. Such representations can be based on words [4], characters [5], [6], [7], sub-word units different than characters [8] or any combination of these. This last approach, originating from feature engineering, where affixes play a significant role, offers an exhaustive way of creating word representations utilizing the semantics of morphemes.

An up-to-date list of entity linkage techniques and datasets is given in [?], [?]. However, these resources are concentrated on texts such as news, and most models are not prepared to face particular challenges in other domains, e.g. literary texts.

The most popular library of texts of various kinds is *Project Gutenberg* [9]. It is accompanied by GutenTag [10], [11] – a software tool offering NLP techniques for analysing literary texts. It contains an automatic corpus reader, subcorpus filters and a tagging functionality based on different tags specified by a user, performing statistical analysis of the selected subcorpus. Even though the NER is used to identify the main literary characters, the information gained is used only for statistical measures. There is no advanced mechanism of recognition and disambiguation of literary characters.

One of the possible approaches to detecting and matching characters in literary texts, based on characters clustering [12], [13], assumes that the noun phrases recognised in the text by the NER model can be clustered into groups referring to the same person. In the case of both papers, the presented results are more qualitative than numeric, based on verifying preregistered hypotheses. Another method [14] is based on representing characters using a graph, where each node corresponds to a name found using NER, and edges connect nodes referring to the same character. Furthermore, this method attempts to identify entities the NER has not recognized by uncovering prototypical characters' behaviours. Accuracy of character detection using this method varies between 45% and 75%, depending on the novel and the challenges appearing in its texts.

Even though numerous papers devoted to recognising literary characters are quoted here, in most cases, there are no available datasets to compare the results or models' parameters to recreate the experiments [13]. Moreover, some proposed methods apply only to narrow sets of texts and require a manual contribution. A good quality, general-purpose way for long, complex texts such as novels and benchmark datasets would be of great applicability.

III. *ProtagonistTagger*: ENGLISH LITERATURE USE CASE ²

The idea for creating a tool for named entity linkage dedicated to novels arose from the complexity of texts in the literary domain and the weak performance of standard methods on such complex texts [14]. The novel is a particular type of text in terms of writing style, the links between sentences, the plot's complexity, the number of characters, etc.

²The *protagonistTagger* tool, benchmark datasets from literary domain and annotated corpus are available at <https://zenodo.org/record/4699418>

TABLE II
METRICS COMPUTED FOR THE STANDARD NER MODELS (*en_core_web_sm* OR *pl_core_news_sm* AVAILABLE AT [HTTPS://SPACY.IO/MODELS](https://spacy.io/models)) AND THE FINE-TUNED NER MODEL FOR ANNOTATIONS WITH GENERAL LABEL *person*.

Testing set/NER model	Precision	Recall	F-measure	Mentions
<i>Test_large_person</i>				
standard	0.84	0.8	0.82	1021
fine-tuned	0.77	0.99	0.87	1021
<i>Test_small_person</i>				
standard	0.78	0.79	0.78	273
fine-tuned	0.69	0.95	0.8	273
Internet news	0.84	0.94	0.89	324

TABLE III
PERFORMANCE OF THE *protagonistTagger* ON VARIOUS DATASETS.

Testing set	Precision	Recall	F-measure
Novels - <i>Test_large_names</i>	0.88	0.87	0.87
Novels - <i>Test_small_names</i>	0.83	0.83	0.83
Internet news	0.8	0.78	0.78

The NER phase of the linkage process in the literary domain uses a pretrained standard NER model fine-tuned with data from the literary domain annotated with general tag *person* in a semi-automatic way [15]. It was necessary due to the relatively low performance of standard models trained primarily on web data such as blogs, news, and comments [16], [17]. In the NER phase, we want to find as many potential *person* entities to be matched in the NED phase as possible. Thus we need to achieve the highest possible recall metric. The phase of NED aims at linking each mention of a protagonist in a given text with a proper tag (i.e., the full name of this protagonist). The list of available tags, i.e. the list of protagonists' proper names predefined for each novel, is given as input to this algorithm. The matching method is mainly based on *approximate text matching*. The algorithm also incorporates a set of hand-crafted rules (lexico-semantic patterns) for distinguishing between different instances of concept *person*, and it uses several external dictionaries. This way, entities' similarities are considered on the semantic and syntactic levels. The *protagonistTagger* addresses the problem of diminutives of basic forms of names and disambiguation of people with the same surname (by analyzing personal title preceding surname in a text).

The tool was used to prepare a corpus of 13 novels (more than 50,000 sentences) and more than 35,000 annotations of literary characters. The corpus was used for the analysis of sentiment and relationships in novels.

V. *ProtagonistTagger*: INTERNET NEWS USE CASE ³

To verify the usability of the created tool in a new domain, we investigated its performance on a set made from internet news written in Polish [18]. The *protagonistTagger* turned out to be universal and easily adapted for new data. The only required modification is changing the language model from English to Polish. In the NER phase, we skipped the step of fine-tuning the standard NER model. This decision was caused by the relatively high performance of the standard NER model on the news compared with literary texts. Since external resources, such as dictionaries and syntactic rules,

³The adapted tool along with the new datasets: <https://zenodo.org/record/5060232>

the language-dependent, they were disabled in this experiment. The NED phase was primarily based on the *approximate string matching* part of the *matching algorithm*.

V. EVALUATION AND DATASETS

Testing sets for the literary domain contain sentences chosen randomly from 13 novels differing in style and genre (*large* – testing sets from 10 novels, and *small* – from distinct 3 novels). The testing sets used for *protagonistTagger* contain together 1,300 sentences (100 sentences from each novel) annotated manually with a general tag *person* for NER testing (see Table II) and full names of the mentioned people for NED testing, as well as for the testing of the entire tool.

The *protagonistTagger* tool achieves high results on all the tested novels – precision and recall above 83% (see Table III). The tool's performance on *Test_small_names* shows that it can successfully be used for new distinct novels to create a larger corpus of annotated texts. The best proof of novels' diversity in the test sets is the tool's performance, whose precision varies from 79% to even 96% for different novels. Even though the performance is tested on various texts, the precision of the annotations remains high, proving the applicability of the proposed method in the literary domain.

The new dataset with internet news written in Polish contains around 1,000 sentences [18]. It is annotated with 10 identifiers (full names of popular Polish individuals, e.g. politicians, actors, researchers, etc.). The overall quality of the annotations is high (both recall and precision above 78%). Even though we applied only a few simple rules and no additional resources (e.g. dictionaries) in the NED phase.

VI. CONCLUSIONS

This paper proposes a tool for *person* entity linkage in various domains – *protagonistTagger*. We also gathered datasets to express the problem of individuals' matching with text mentions. The method uses pretrained NER models and various techniques for NED. The initial tests were performed in the English literary domain. Most recent experiments prove the adaptability and effectiveness of the proposed approach for application to another language and domain of Internet news. The tool achieved high results. The only precondition of

using the tool is access to the predefined tags defining the full names (persons' identifiers) to be linked with named entities appearing in a text.

Future applications of *protagonistTagger* may concern fully automatically annotating texts from social media and using these annotations to investigate human opinions and analysing sentiments towards specific persons. Even in the literary domain, constructing the representation and interpretation of narratives, extracting social networks from novels, and modelling social conversations that occur between characters in the form of a network [13] are promising directions when a corpus with recognised protagonists is available. Besides, literary characters and their relationships evolve with the novel's progress. Modelling dynamic relationships between pairs of characters by detecting relationship sequences in data much more adequate in the case of long texts [19], [20]. Another aspect the interpretation of narratives is sentiment analysis [21]. It includes, among others, a classification of literary texts by literary genre based on emotions they convey [22], [23] and emotion-based character analysis [24], [25].

REFERENCES

- [1] T. Stanislawek and A. Wróblewska, "Named entity recognition - is there a glass ceiling?" in *CoNLL*. ACL, 2019, pp. 624–633.
- [2] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," in *27th COLING*, 2018, pp. 2145–2158.
- [3] A. Akbik and T. Bergmann, "Pooled contextualized embeddings for named entity recognition," in *NAACL-HLT: Volume 1*, 2019, pp. 724–728.
- [4] R. Collobert and J. Weston, "Natural language processing (almost) from scratch," in *Journal of machine learning research*, 2011, pp. 2493–2537.
- [5] X. Yu and S. Mayhew, "On the strength of character language models for multilingual named entity recognition," in *EMNLP*. ACL, 2018.
- [6] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional lstm-cnns," in *Transactions of the ACL*, 2016, pp. 357–370.
- [7] G. Lample and M. Ballesteros, "Neural architectures for named entity recognition," in *15th NAACL-HLT*, 2016, pp. 260–270.
- [8] V. Yadav and R. Sharp, "Deep affix features improve neural named entity recognizers," in *7th *SEM*, 2018, pp. 167–172.
- [9] M. Hart, "The history and philosophy of project gutenberg," in *Project Gutenberg*, 1992.
- [10] J. Brooke and A. Hammond, "Gutentag: an nlp-driven tool for digital humanities research in the project gutenberg corpus," in *4th Workshop Computational Linguistics for Literature*, 2015, pp. 42–47.
- [11] A. Hammond and J. Brooke, "Gutentag: A user-friendly, open-access, open-source system for reproducible large-scale computational literary," 2017.
- [12] D. Bamman and T. Underwood, "A bayesian mixed effects model of literary character," in *52nd ACL*, 2014, pp. 370–379.
- [13] D. Elson and N. Dames, "Extracting social networks from literary fiction," in *48th ACL*, 2010, pp. 138–147.
- [14] H. Vala and D. Jurgens, "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 *EMNLP*, 2015, pp. 769–774.
- [15] J. Kim, Y. Ko, and J. Seo, "Construction of machine-labeled data for improving named entity recognition by transfer learning," *IEEE Access*, vol. 8, pp. 59 684–59 693, 2020.
- [16] R. Jiang and R. E. Banchs, "Evaluating and combining name entity recognition systems," in *6th Named Entity Workshop*, 2016, pp. 21–27.
- [17] X. Schmitt and S. Kubler, "A replicable comparison study of ner software: Stanfornlp, nltk, opennlp, spacy, gate," in *6th SNAMS*. IEEE, 2019, pp. 338–343.
- [18] M. Pichocki and A. Wróblewska, "Categorization of persons based on the mentions in polish news texts," *JAMRIS*, vol. 14, 2020.
- [19] S. Chaturvedi and S. Srivastava, "Modeling evolving relationships between characters in literary novels," in *13th AAAI Conference on Artificial Intelligence*, 2016.
- [20] S. Chaturvedi and M. Iyyer, "Unsupervised learning of evolving relationships between literary characters," in *AAAI*, 2017, pp. 3159–3165.
- [21] E. Kim and R. Klinger, "A survey on sentiment and emotion analysis for computational literary studies," in *arXiv preprint arXiv:1808.03137*, 2018.
- [22] A. J. Reagan and L. Mitchell, "The emotional arcs of stories are dominated by six basic shapes," in *EPJ Data Science*, vol. 5. SpringerOpen, 2016, pp. 1–12.
- [23] A. Zehe and M. Becker, "Prediction of happy endings in german novels based on sentiment information," in *DMNLP*, 2016.
- [24] A. Groza and L. Corde, "Information retrieval in folktales using natural language processing," in *ICCP*, 2015, pp. 59–66.
- [25] L. Flekova and I. Gurevych, "Personality profiling of fictional characters using sense-level links between lexical resources," in *EMNLP*, 2015, pp. 1805–1816.

76%

SIMILARITY INDEX

PRIMARY SOURCES

- | | | |
|--|---|------------------|
| <div style="background-color: red; color: white; padding: 5px; display: inline-block; width: 30px; height: 30px; text-align: center; line-height: 30px;">1</div> | deepai.org
<small>Internet</small> | 1378 words — 43% |
| <hr/> | | |
| <div style="background-color: magenta; color: white; padding: 5px; display: inline-block; width: 30px; height: 30px; text-align: center; line-height: 30px;">2</div> | www.arxiv-vanity.com
<small>Internet</small> | 837 words — 26% |
| <hr/> | | |
| <div style="background-color: purple; color: white; padding: 5px; display: inline-block; width: 30px; height: 30px; text-align: center; line-height: 30px;">3</div> | Tingming Lu, Yaocheng Gui, Zhiqiang Gao. "Learning Document-Level Label Propagation and Instance Selection by Deep Q-Network for Interactive Named Entity Annotation", IEEE Access, 2021
<small>Crossref</small> | 41 words — 1% |
| <hr/> | | |
| <div style="background-color: teal; color: white; padding: 5px; display: inline-block; width: 30px; height: 30px; text-align: center; line-height: 30px;">4</div> | Qianwen Wang, Mizuho Iwaihara. "Deep Neural Architectures for Joint Named Entity Recognition and Disambiguation", 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), 2019
<small>Crossref</small> | 34 words — 1% |
| <hr/> | | |
| <div style="background-color: green; color: white; padding: 5px; display: inline-block; width: 30px; height: 30px; text-align: center; line-height: 30px;">5</div> | S Thara, Prabakaran Poornachandran. "Corpus Creation and Transformer based Language Identification for Code-Mixed Indian Language", IEEE Access, 2021
<small>Crossref</small> | 30 words — 1% |
| <hr/> | | |
| <div style="background-color: orange; color: white; padding: 5px; display: inline-block; width: 30px; height: 30px; text-align: center; line-height: 30px;">6</div> | hal.archives-ouvertes.fr
<small>Internet</small> | 27 words — 1% |
| <hr/> | | |
| <div style="background-color: brown; color: white; padding: 5px; display: inline-block; width: 30px; height: 30px; text-align: center; line-height: 30px;">7</div> | web.archive.org
<small>Internet</small> | |

20 words — 1%

8 insightsociety.org
Internet

16 words — < 1%

9 peerj.com
Internet

14 words — < 1%

10 [Lecture Notes in Computer Science, 2013.](#)
Crossref

12 words — < 1%

11 www.medrxiv.org
Internet

10 words — < 1%

EXCLUDE QUOTES OFF
EXCLUDE BIBLIOGRAPHY OFF

EXCLUDE SOURCES OFF
EXCLUDE MATCHES OFF