



Protagonists' Catcher in Novels

A Dataset and A Tool

Weronika Łajewska
supervised by PhD Anna Wróblewska

Cofunded by REESA: *Machine Learning-based systems for the automation of systematic literature reviews in food safety domain*

The National Center for Research and Development (NCBiR), POLNOR 2019 call,
grant no: NOR/POLNOR/REESA/0059/2019
Project leader: prof. Radosław Pytlak

*"Great literature is simply language
charged with meaning to the utmost possible
degree."*

Ezra Pound

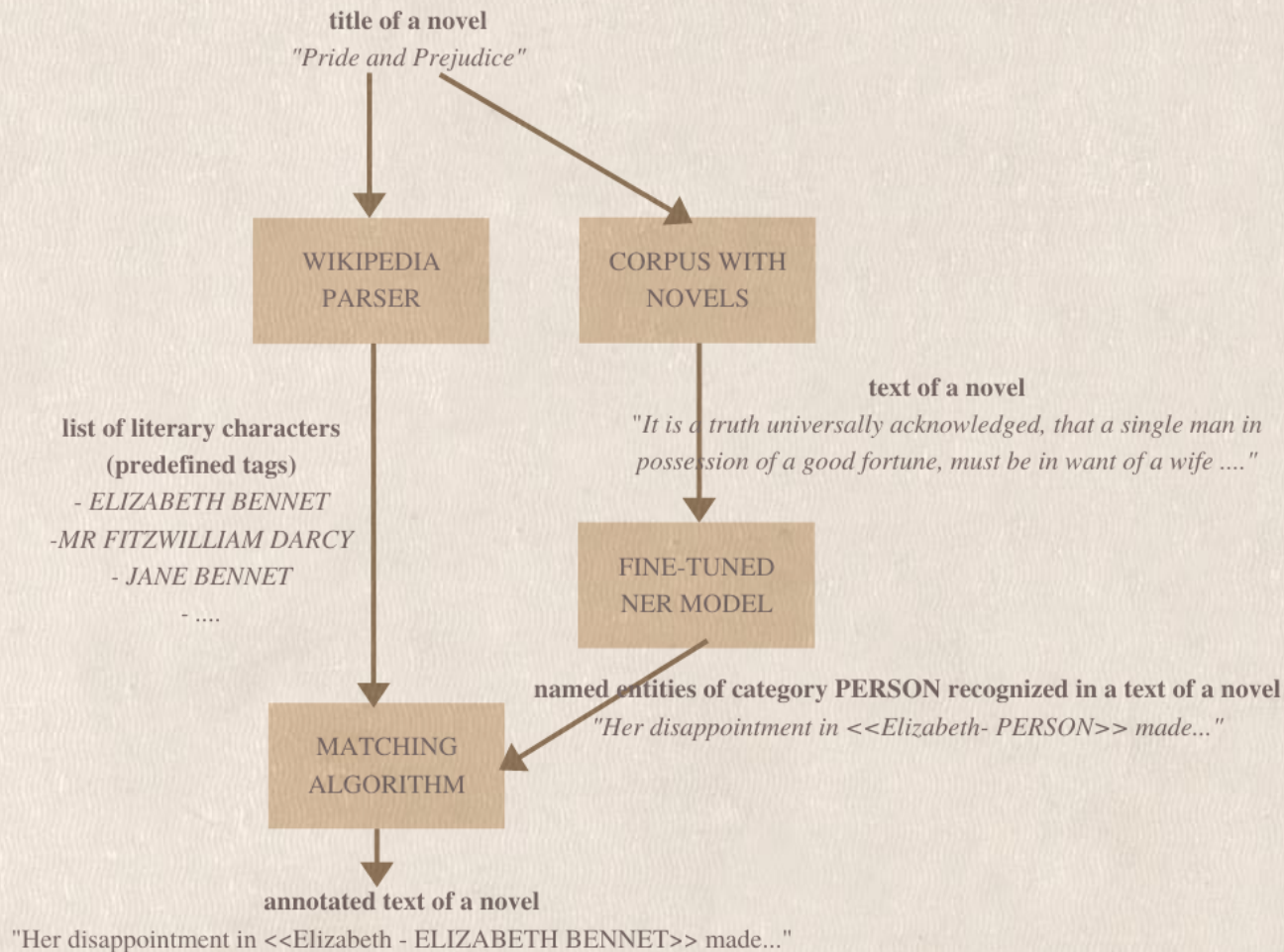
Main goal of the project

reasonably big corpus with annotated novels in which every protagonist is annotated with his proper name

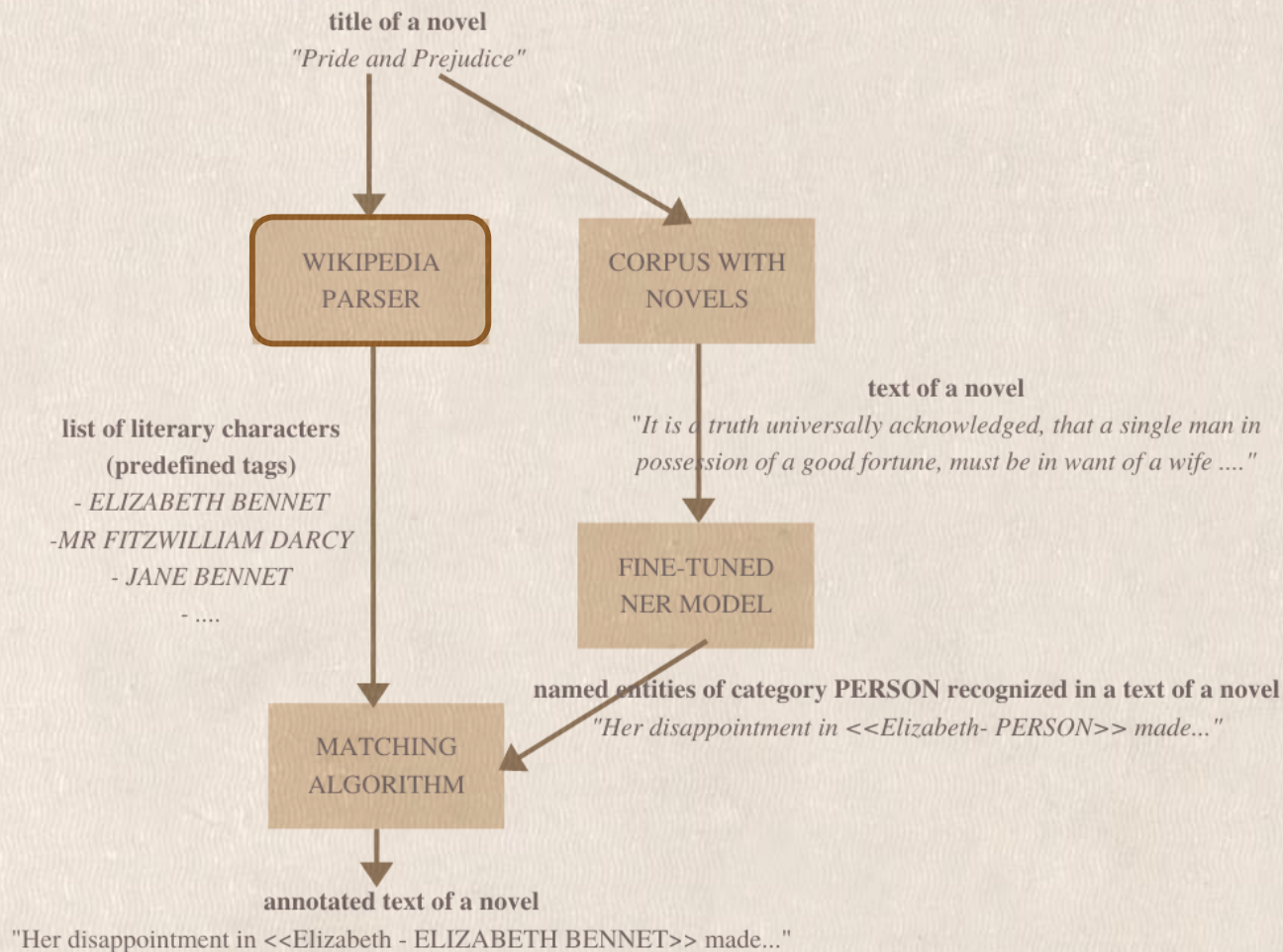
model for recognizing appearances of protagonists in a novel

Exemplary output of *ProtagonistTagger*

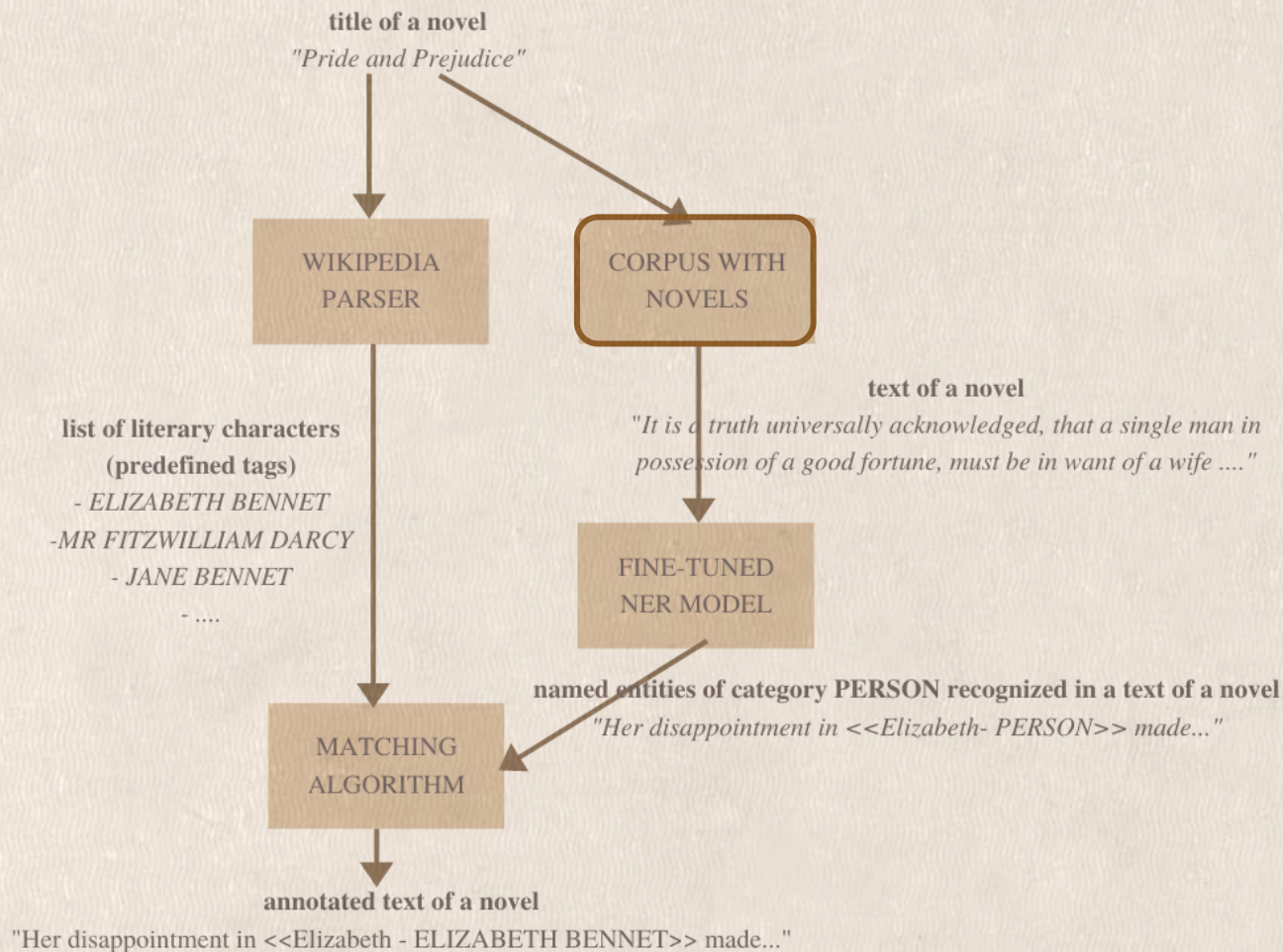
"Her disappointment in **Charlotte**«*Charlotte Lucas*» made her turn with fonder regard to her sister, of whose rectitude and delicacy she was sure her opinion could never be shaken, and for whose happiness she grew daily more anxious, as **Bingley**«*Charles Bingley*» had now been gone a week and nothing 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 twelvemonth's abode in the family might have prompted."



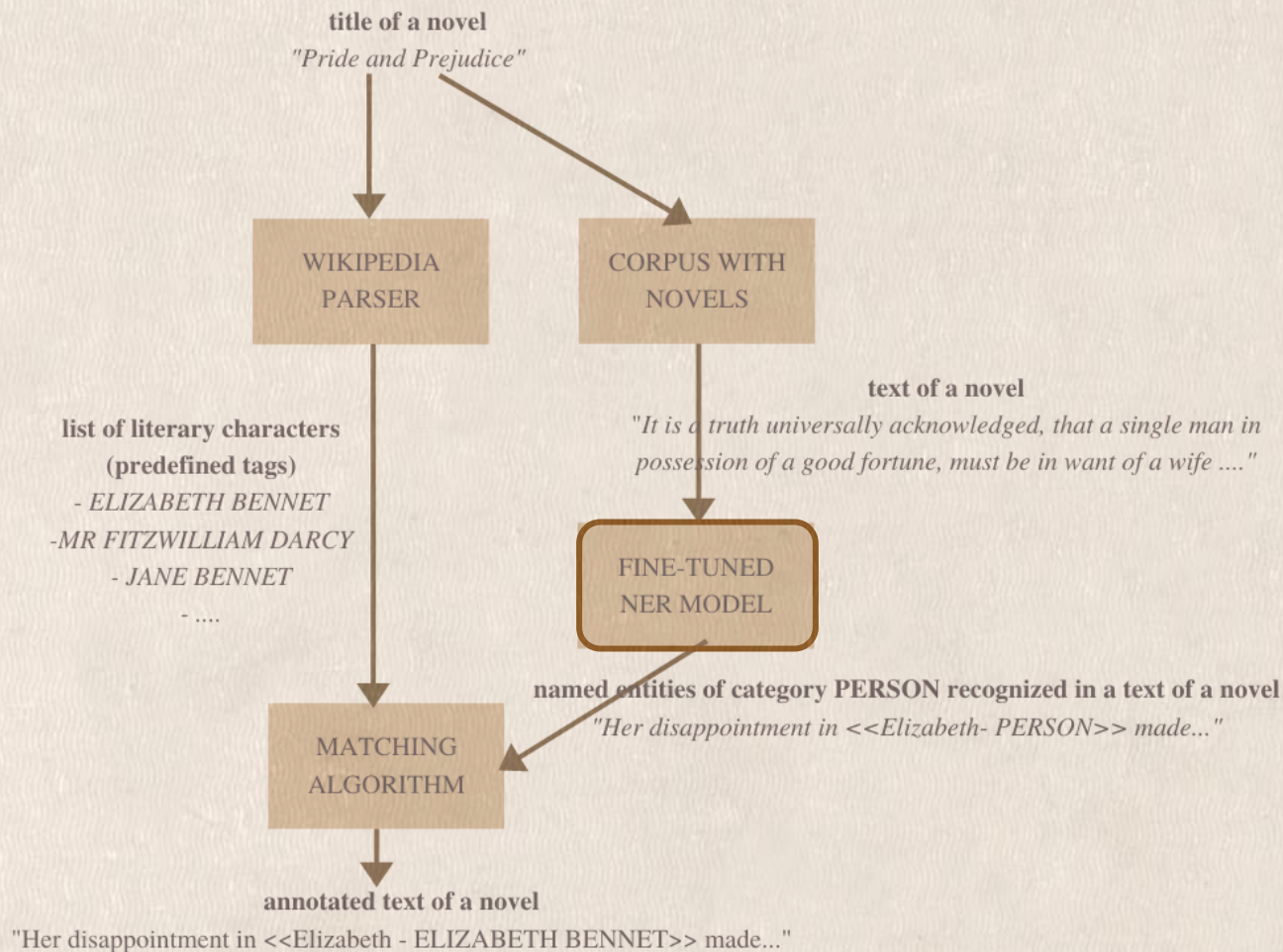
Protagonist Tagger



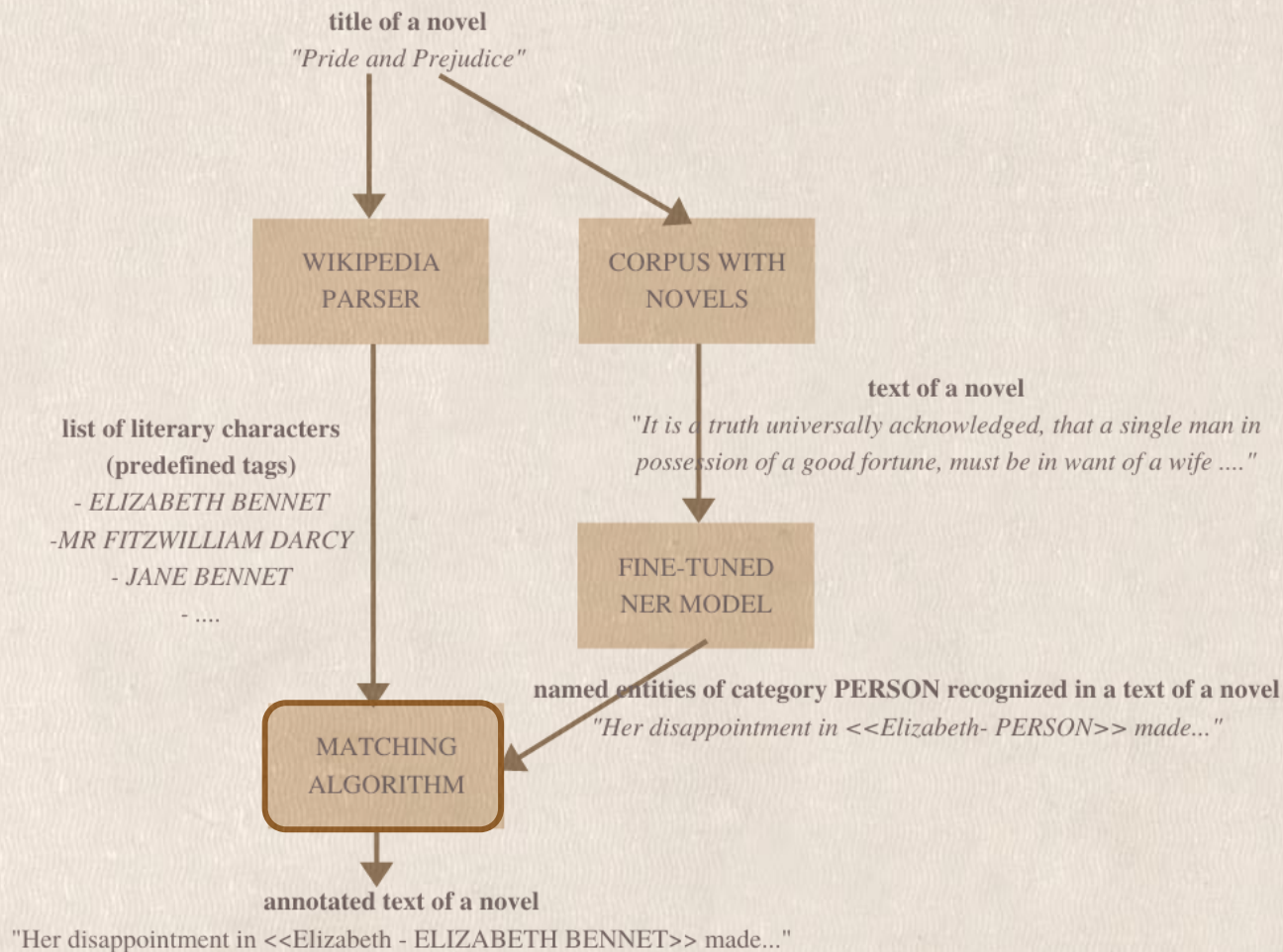
Protagonist Tagger



Protagonist Tagger

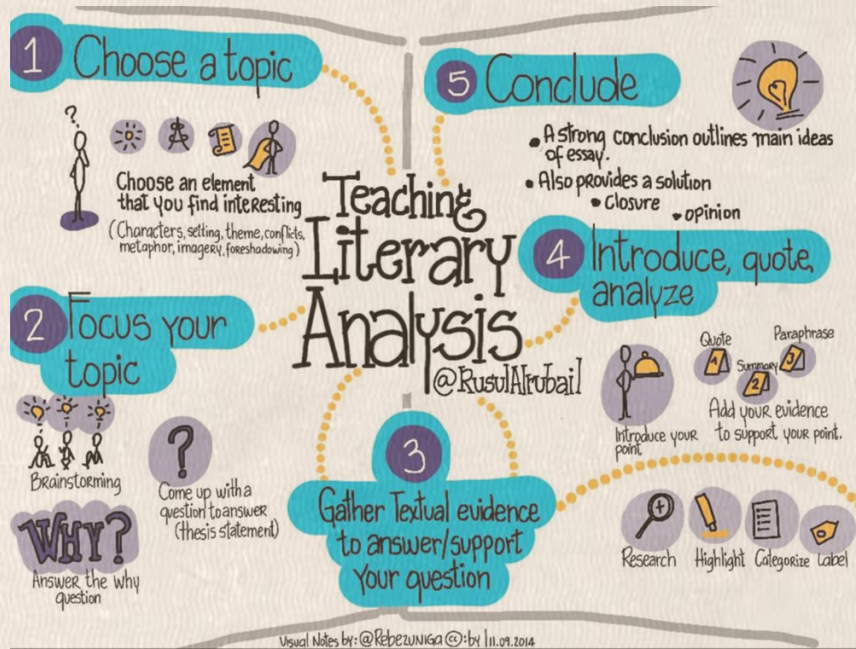


Protagonist Tagger



Protagonist Tagger

State-of-the-art Research on Literary Text Analysis



- the analysis of novels
- the problem of annotating literary characters
- the possibilities in novels analysis given by literary characters' annotations

State-of-the-art Research on Literary Text Analysis

Analysis of the Novels

- *Project Gutenberg* - huge literary texts corpus (fiction novels, collections of poetry, dramas, cookbooks, bibliographies, dictionaries)
- *GutenTag* - tool for analysis of texts in *Project Gutenberg*:
 - corpus reader
 - subcorpus filter
 - tagging

SUBCORPUS 1 (2)

Character dialogism (stylistic diversity measure): 0.1217

Percent dialogue by female characters: 0.6016

Percent of characters which are female: 0.5009

Lexical density: 0.4679

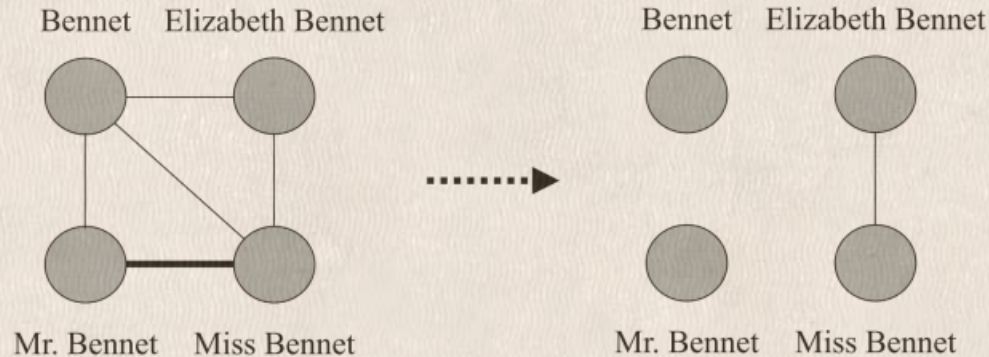
Average length of words: 3.8973

Percent of text which is dialogue: 0.4771

State-of-the-art Research on Literary Text Analysis

Problem of Annotating Literary Characters

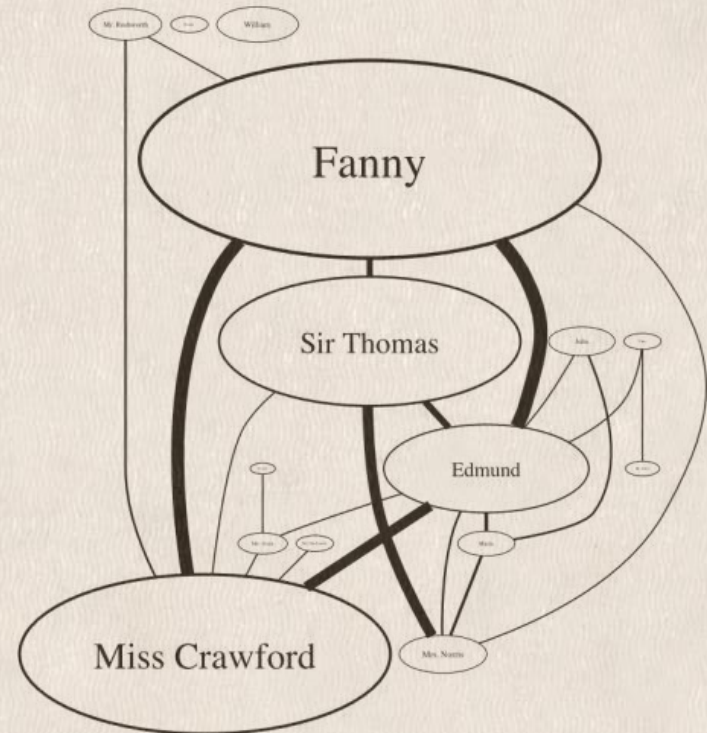
- clustering noun phrases - noun phrases are clustered into groups referring to the same person creating sets of co-referents for the same named entity
- graph representation - each node corresponds to a name found using NER and edges connects nodes referring to the same character



State-of-the-art Research on Literary Text Analysis

Further Possibilities




Detecting relationships between literary characters based on dialogue interactions
- modelling social conversations that occur between characters in a form of a network



State-of-the-art Research on Literary Text Analysis

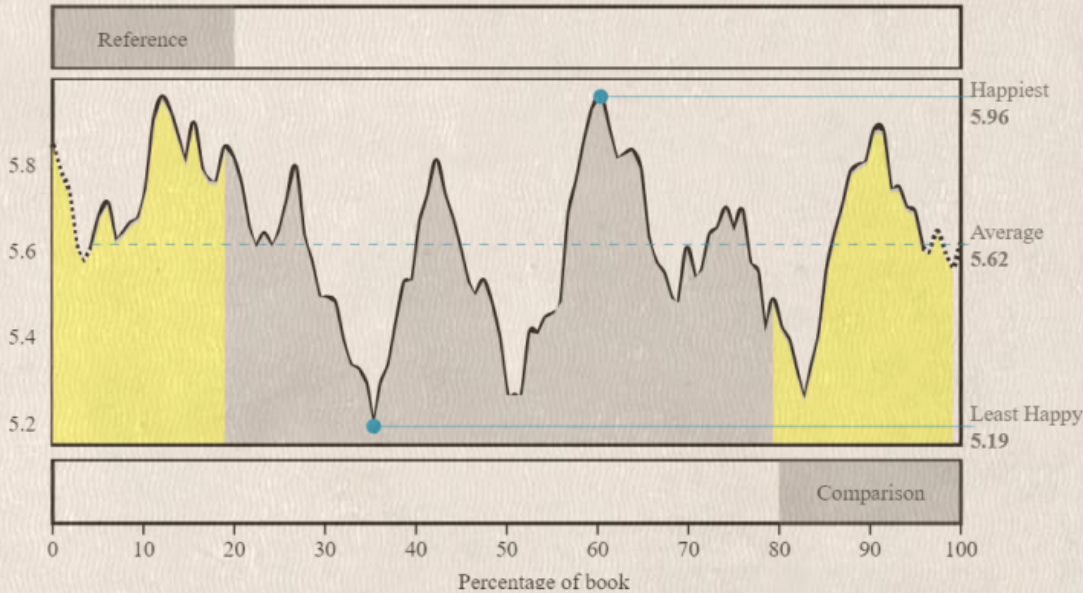
Further Possibilities

Detecting evolving relationships between literary characters - modelling dynamic relationships between pairs of characters by detecting relationships sequences in data

-  { S1: Tom falls in love with Becky Thatcher, a new girl in town, and persuades her to get “engaged” to him.
-  { S2: Their romance collapses when she learns that Tom has been “engaged” before—to a girl named Amy Lawrence.
- S3: Shortly after being shunned by Becky, Tom ...
-  { S4: ...Tom gets himself back in Becky’s favor after he nobly accepts the blame for a book that she has ripped.
- S5: Meanwhile, Tom goes on a picnic to McDougal’s Cave with Becky and their classmates.

State-of-the-art Research on Literary Text Analysis

Further Possibilities



Sentiment-based literary text classification – discovering a literary genre through sentiment analysis.

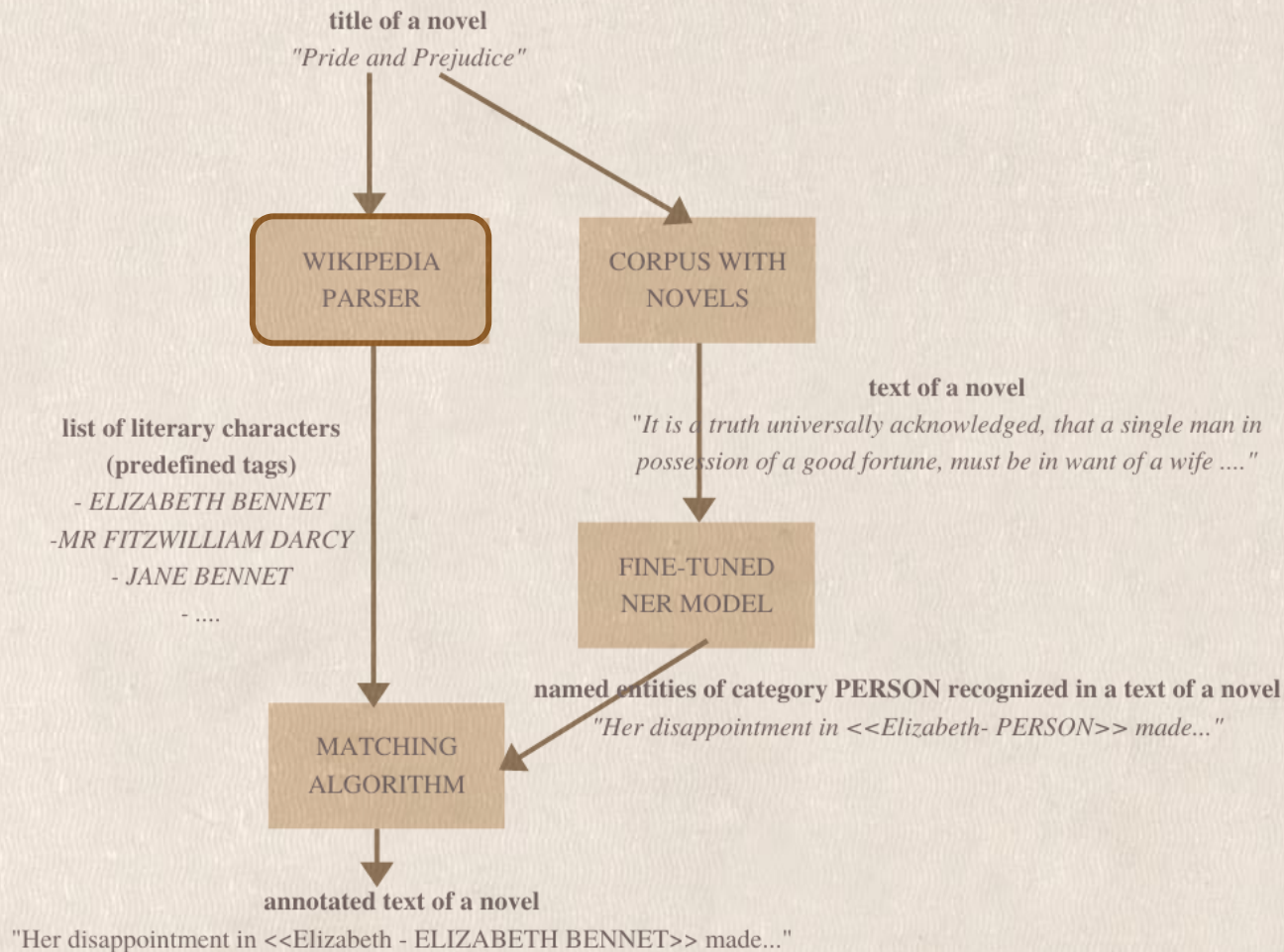
State-of-the-art Research on Literary Text Analysis

Further Possibilities

Characters' analysis – by their role in the story (eg. hero, villain, princess) or by profiling personality (extrovert or introvert) based on frequency word clouds.



List of literary characters



Protagonist Tagger

Characters [[edit](#)]

- **Elizabeth Bennet** – the second eldest of the Bennet daughters, she is twenty years old and intelligent, lively, playful, attractive, and witty – but with a tendency to form tenacious and prejudicial first impressions. As the story progresses, so does her relationship with Mr Darcy. The course of Elizabeth and Darcy's relationship is ultimately decided when Darcy overcomes his pride, and Elizabeth overcomes her prejudice, leading them both to surrender to their love for each other.
- **Mr Fitzwilliam Darcy** – Mr Bingley's friend and the wealthy, twenty-eight-year-old owner of the family estate of [Pemberley](#) in [Derbyshire](#), rumoured to be worth at least £10,000 a year (equivalent to £796,000 or \$1,045,000 in 2018).^[7] While he is handsome, tall, and intelligent, Darcy lacks ease and [social graces](#), and so others frequently mistake his initially haughty reserve and rectitude as proof of excessive pride (which, in part, it is). A new visitor to the village, he is ultimately Elizabeth Bennet's love interest.
- **Mr Bennet** – A late-middle-aged [landed gentleman](#) of a modest income of £2000 per annum, and the dryly sarcastic [patriarch](#) of the now-dwindling [Bennet family](#) (a family of [Hertfordshire](#) landed gentry), with five unmarried daughters. His estate, Longbourn, is [entailed](#) to the male line.
- **Mrs Bennet (née Gardiner)** – the middle-aged wife of her social superior, Mr Bennet, and the mother of their five daughters. Mrs Bennet is a [hypochondriac](#) who imagines herself susceptible to attacks of tremors and palpitations (her "poor nerves"), whenever things are not going her way. Her main ambition in life is to marry her daughters off to wealthy men. Whether or not any such matches will give her daughters happiness is of little concern to her.
- **Jane Bennet** – the eldest Bennet sister. Twenty-two years old when the novel begins, she is considered the most beautiful young lady in the neighbourhood and is [inclined to see only the good in others](#) (but can be persuaded otherwise on sufficient evidence). She falls in love with Charles Bingley, a rich young gentleman recently moved to Hertfordshire and a close friend of Mr Darcy.
- **Mary Bennet** – the middle Bennet sister, and the plainest of her siblings. Mary has a serious disposition and mostly reads and plays music, although she is often impatient to display her accomplishments and is rather vain about them. She frequently moralises to her family. According to James Edward Austen-Leigh's *[A Memoir of Jane Austen](#)*, Mary ended up marrying one of her Uncle Philips' law clerks and moving into Meryton with him.
- **Catherine "Kitty" Bennet** – the fourth Bennet daughter at 17 years old. Though older than Lydia, she is her shadow and follows her in her pursuit of the officers of the militia. She is often portrayed as envious of Lydia and is described a "silly" young woman. However, it is said that she improved when removed from Lydia's influence. According to James Edward Austen-Leigh's *[A Memoir of Jane Austen](#)*, Kitty later married a clergyman who lived near Pemberley.
- **Lydia Bennet** – the youngest Bennet sister, aged 15 when the novel begins. She is frivolous and headstrong. Her main activity in life is socializing, especially flirting with the officers of the militia. This leads to her running off with George Wickham, although he has no intention of marrying her. Lydia shows no regard for the moral code of her society; as Ashley Tauchert says, she "feels without reasoning."^[8]
- **Charles Bingley** – a handsome, amiable, wealthy young gentleman from the north of England (possibly [Yorkshire](#), as [Scarborough](#) is mentioned, and there is, in fact, a real-life town called [Bingley](#) in [West Yorkshire](#)), who leases Netherfield Park, an estate three miles from Longbourn, with the hopes of purchasing it. He is contrasted with Mr Darcy for having more generally pleasing manners, although he is reliant on his more experienced friend for advice. An example of this is the prevention of Bingley and Jane's romance because of Bingley's undeniable dependence on Darcy's opinion.^[9] He lacks resolve and is easily influenced by others; his two sisters, Miss Caroline Bingley and Mrs Louisa Hurst, both disapprove of Bingley's growing affection for Miss Jane Bennet.

Character genealogy

[\[show\]](#)

Pride and Prejudice by Jane Austen

Elizabeth Bennet

Mr Fitzwilliam Darcy

Mr Bennet

Mrs Bennet (née Gardiner)

Jane Bennet

Mary Bennet

Catherine "Kitty" Bennet

Lydia Bennet

Charles Bingley

Caroline Bingley

George Wickham

Mr William Collins

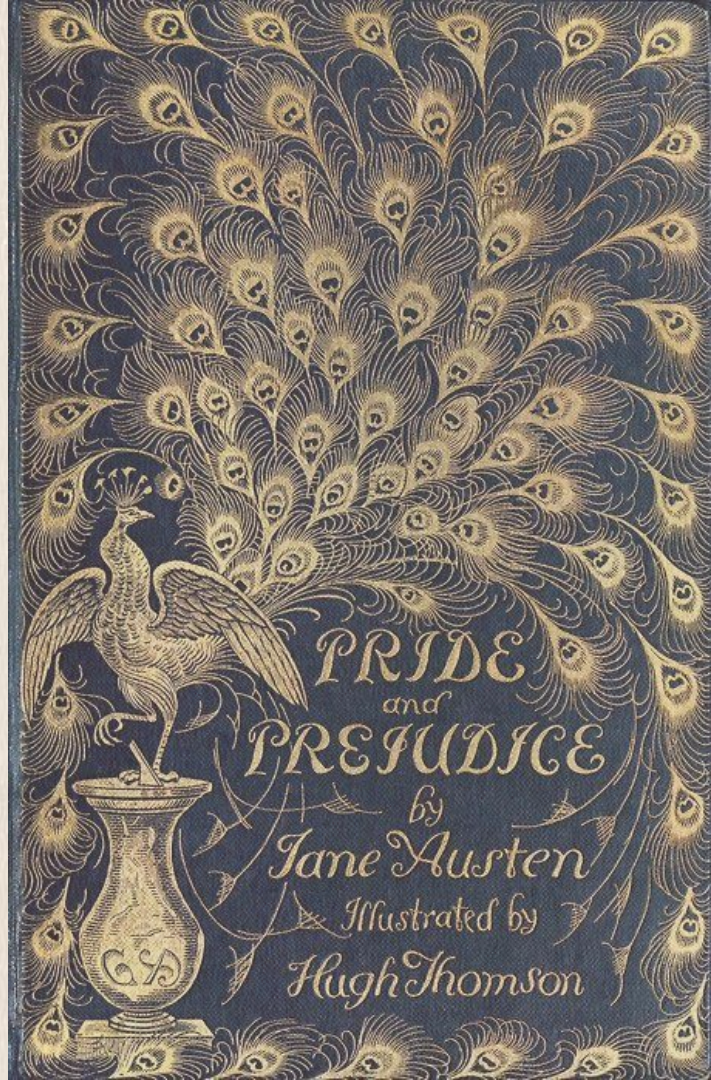
Lady Catherine de Bourgh

Mr Edward and Mrs M Gardiner

Georgiana Darcy

Charlotte Lucas

Colonel Fitzwilliam



NER models in general

How can we approach NER?

Handcrafted rules

Handwritten finite state patterns to recognize names, dates, etc.

How can we approach NER?

Knowledge-based systems

Relying on domain specific resources such as lexicons or dictionaries.

Handcrafted rules

Handwritten finite state patterns to recognize names, dates, etc.

How can we approach NER?

Unsupervised and bootstrapped systems

Training classifier on unlabelled data. Reducing supervision to 7 *seed* rules.

Knowledge-based systems

Relying on domain specific resources such as lexicons or dictionaries.

Handcrafted rules

Handwritten finite state patterns to recognize names, dates, etc.

How can we approach NER?

..., says **Mr. Cooper**, a vice **president** of ...

Unsupervised and bootstrapped systems

Training classifier on unlabelled data. Reducing supervision to 7 *seed* rules.

Knowledge-based systems

Relying on domain specific resources such as lexicons or dictionaries.

Handcrafted rules

Handwritten finite state patterns to recognize names, dates, etc.

How can we approach NER?

Feature-engineered supervised systems

System based on hidden Markov Model that learns to recognize and classify named entities.

Unsupervised and bootstrapped systems

Training classifier on unlabelled data. Reducing supervision to 7 *seed* rules.

Knowledge-based systems

Relying on domain specific resources such as lexicons or dictionaries.

Handcrafted rules

Handwritten finite state patterns to recognize names, dates, etc.

How can we approach NER?

Feature-inferring neural networks

Avoiding feature-engineering thanks to word embeddings.

Feature-engineered supervised systems

System based on hidden Markov Model that learns to recognize and classify named entities.

Unsupervised and bootstrapped systems

Training classifier on unlabelled data. Reducing supervision to 7 *seed* rules.

Knowledge-based systems

Relying on domain specific resources such as lexicons or dictionaries.

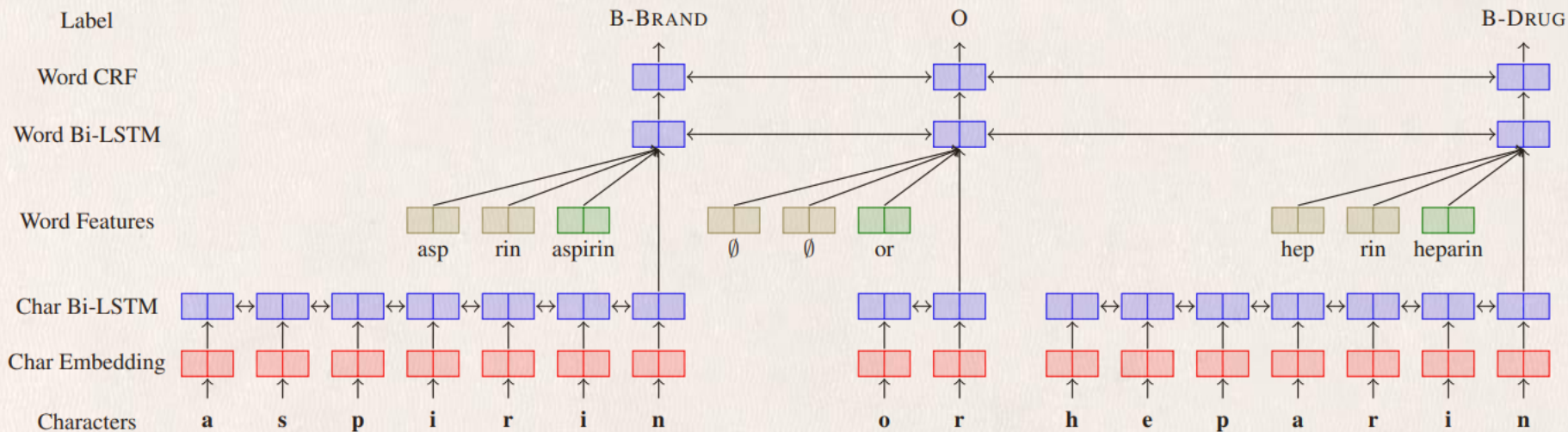
Handcrafted rules

Handwritten finite state patterns to recognize names, dates, etc.

Neural network NER systems

- word level architectures
- character + word level architectures
 - word embedding + convolution over word's characters
 - word embedding + LSTM over word's characters
- character + word + affix level architectures

Character + word + affix level architecture in neural network NER system

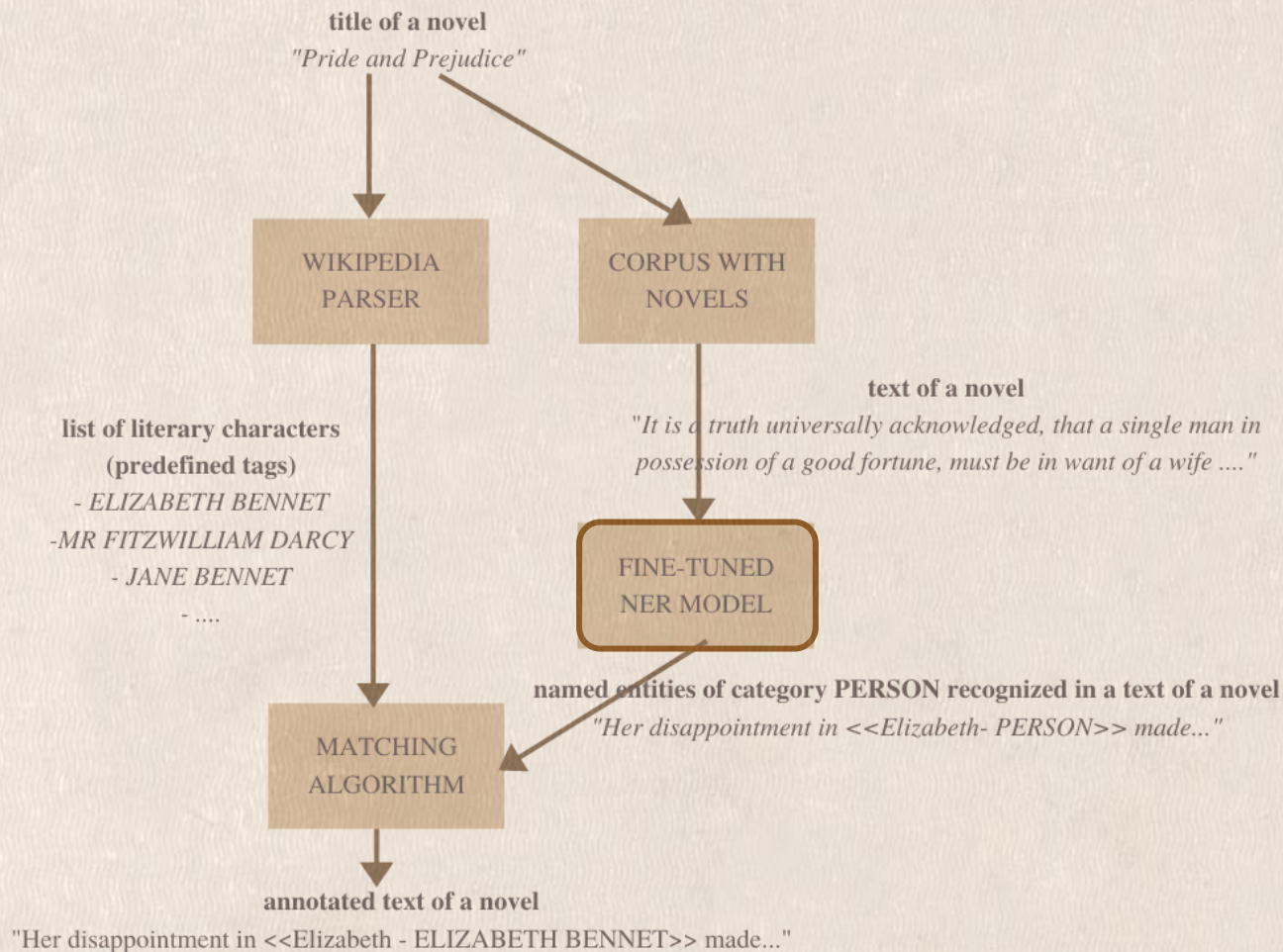


Vikas Yadav, Rebecca Sharp, and Steven Bethard. Deep affix features improve neural named entity recognizers. In Proceedings of the 7th Joint Conference on Lexical and Computational Semantics, pages 167–172, 2018

Summing up NER models degression

- feature-inferring NN systems outperform all the others
- word + character hybrid models are generally better than both word-based and character-based models
- incorporating affix representation into word + character hybrid models improved NER systems performance in some domains (Spanish, Dutch, German, MedLine)
- word embeddings are crucial in avoiding vector feature engineering
- NER task is not solved - there is still room for improvement

NER model performance



Fine-tuning NER

Imperfections of NER in novels

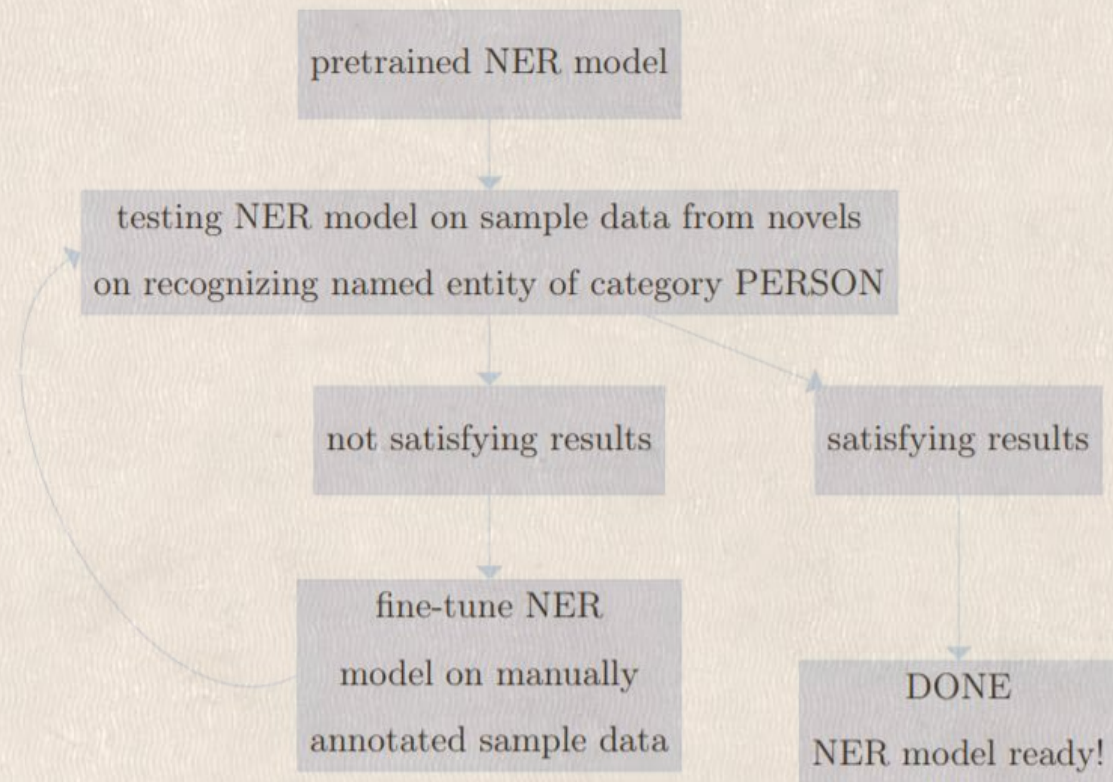
Novel title	precision	recall	F-measure	support
The Picture of Dorian Gray	0.69	0.41	0.51	90
Frankenstein	0.91	0.62	0.74	93
Treasure Island	0.75	0.66	0.7	97
Emma	0.84	0.77	0.81	115
Jane Eyre	0.86	0.78	0.82	97
Wuthering Heights	0.95	0.87	0.91	108
Pride and Prejudice	0.85	0.87	0.86	124
Dracula	0.86	0.94	0.9	97
Anne of Green Gables	0.91	0.96	0.94	114
Adventures of Huckleberry Finn	0.71	0.99	0.83	86
*** Overall results ***	0.84	0.8	0.82	1021

NORP Not recognized named entities

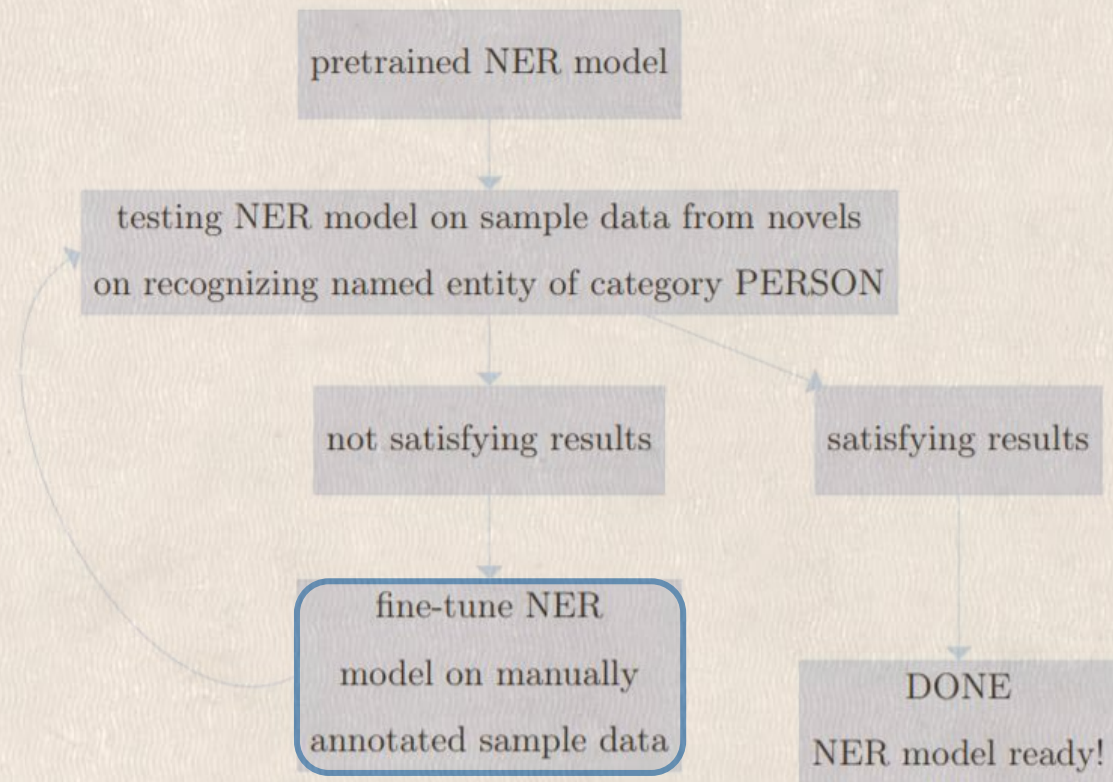
Novel title	Named entities of category <i>PERSON</i> not recognized by NER
The Picture of Dorian Gray	Dorian/Dorian Gray, Sibyl Vane, Hallward/Basil/Basil Hallward
Treasure Island	Flint/Cap'n Flint, Silver (however John Silver is recognized), Black Dog, Gray, Trelawney, Billy Bones, Hawkins, Arrow (however Mr. Arrow is recognized), Pew
Frankenstein	Safie, Victor, Felix, Walton, Justine, creature/monster, Clerval, De Lacey
Emma	Emma/Miss Woodhouse, Harriet
Jane Eyre	Blanche/Blanche Ingram/Miss Ingram, Bessie, Leah, Miss Eyre, Helen, Georgiana, Rosamond, Fairfax Rochester, Rivers, Madam Reed, Miss Temple, Grace
Wuthering Heights	Nelly (however Ellen is recognized), Linton, Hindley, Hareton, Isabella, Heathcliff
Pride and Prejudice	Charlotte, Bingley, Wickham, Lydia, Gardiners, Georgiana, Kitty

Not recognized named entities

Novel title	Named entities of category <i>PERSON</i> not recognized by NER
The Picture of Dorian Gray	Dorian/Dorian Gray, Sibyl Vane, Hallward/Basil/Basil Hallward
Treasue Island	Flint/Cap'n Flint, Silver (however John Silver is recognized), Black Dog, Gray, Trelawney, Billy Bones, Hawkins, Arrow (however Mr. Arrow is recognized), Pew
Frankenstein	Safie, Victor, Felix, Walton, Justine, creature/monster, Clerval, De Lacey
Emma	Emma/Miss Woodhouse, Harriet
Jane Eyre	Blanche/Blanche Ingram/Miss Ingram, Bessie, Leah, Miss Eyre, Helen, Georgiana, Rosamond, Fairfax Rochester, Rivers, Madam Reed, Miss Temple, Grace
Wuthering Heights	Nelly (however Ellen is reconigzed), Linton, Hindley, Hareton, Isabella, Heathcliff
Pride and Prejudice	Charlotte, Bingley, Wickham, Lydia, Gardiners, Georgiana, Kitty



Fine-tuning NER



Fine-tuning NER

Training sets for fine-tuning NER

Sentences with not
recognized named
entities of category
person

Novel title	Named entities of category <i>PERSON</i> not recognized by NER
The Picture of Dorian Gray	Dorian/Dorian Gray, Sibyl Vane, Hallward/Basil/Basil Hallward
Treasue Island	Flint/Cap'n Flint, Silver (however John Silver is recognized), Black Dog, Gray, Trelawney, Billy Bones, Hawkins, Arrow (however Mr. Arrow is recognized), Pew
Frankenstein	Safie, Victor, Felix, Walton, Justine, creature/monster, Clerval, De Lacey
Emma	Emma/Miss Woodhouse, Harriet
Jane Eyre	Blanche/Blanche Ingram/Miss Ingram, Bessie, Leah, Miss Eyre, Helen, Georgiana, Rosamond, Fairfax Rochester, Rivers, Madam Reed, Miss Temple, Grace
Wuthering Heights	Nelly (however Ellen is reconigzed), Linton, Hindley, Hareton, Isabella, Heathcliff
Pride and Prejudice	Charlotte, Bingley, Wickham, Lydia, Gardiners, Georgiana, Kitty

Exemplary sentences:

- “I found her a fine woman, in the style of <<**Blanche Ingram - PERSON**>>: tall, dark, and majestic.”
- “But tell me, what did she say about <<**Mr. Dorian Gray - PERSON**>>?”

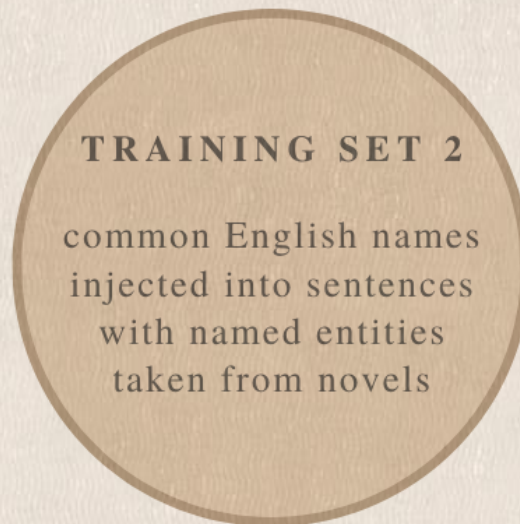
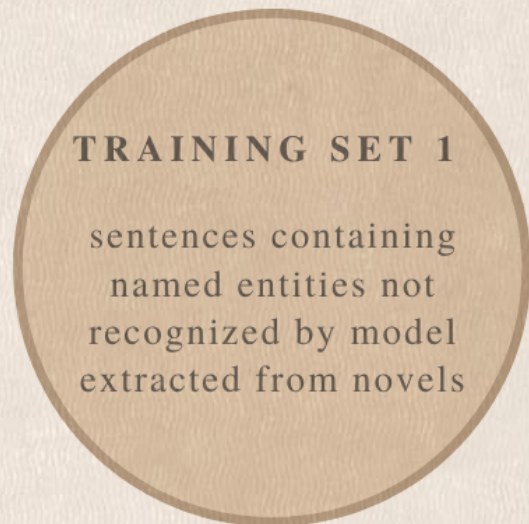
Training sets for fine-tuning NER

Sentences from novels
with injected common
English names

*"**Jane**'s delicate sense of honour would not allow her to speak to **Elizabeth** privately of what **Lydia** had let fall; **Elizabeth** was glad of it; till it appeared whether her inquiries would receive any satisfaction, she had rather be without a confidante."*



*"**Deborah**'s delicate sense of honour would not allow her to speak to **Harvey** privately of what **Lydia** had let fall; **Harvey** was glad of it; till it appeared whether her inquiries would receive any satisfaction, she had rather be without a confidante."*



Tested
NER
models

Fine-tuned NER models performance

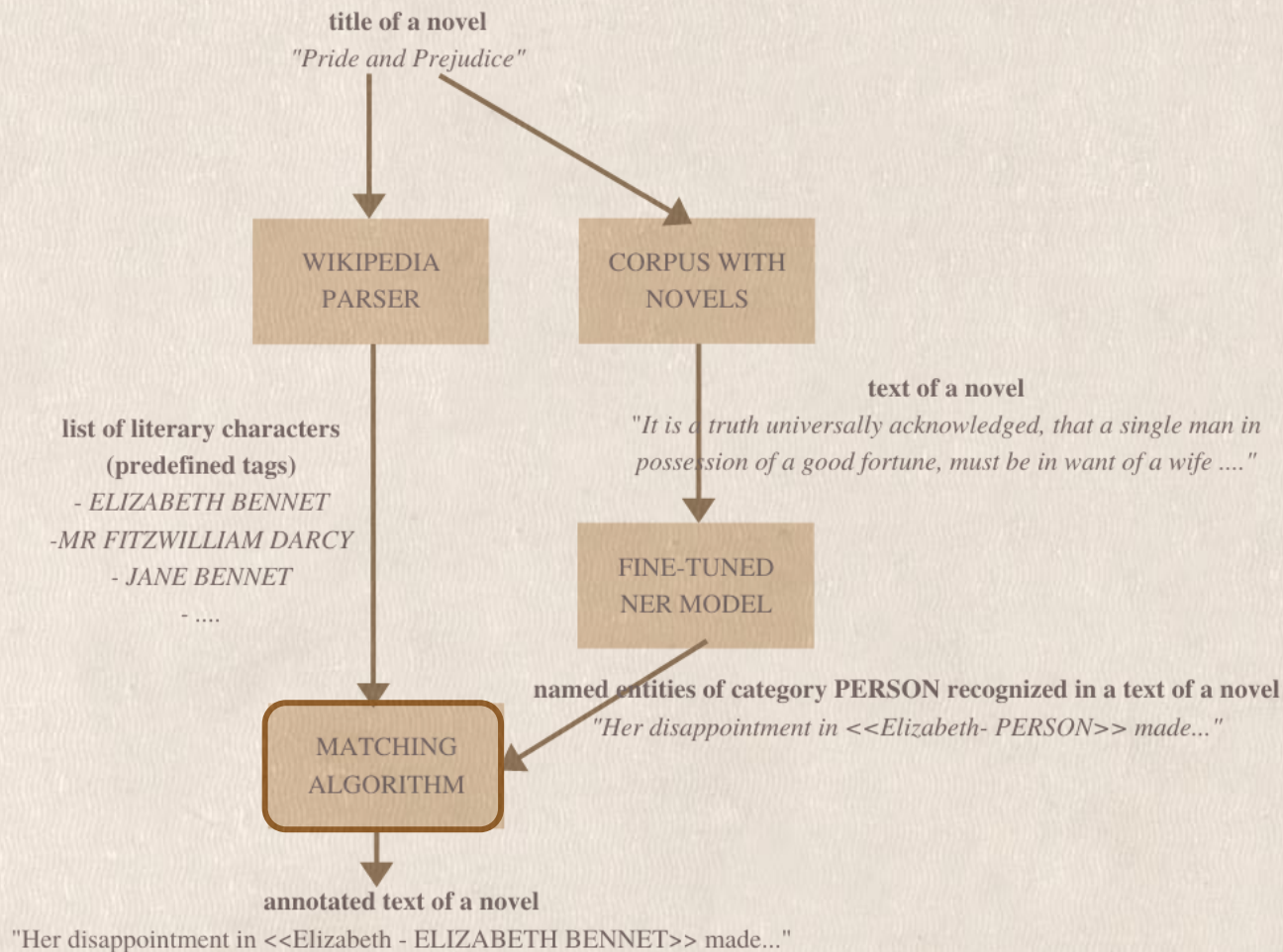
Large testing set →

NER model	precision	recall	F-measure
standard	0.84	0.8	0.82
fine-tuned	0.77	0.99	0.87

*Small testing set
(totally new novels)* →

standard	0.78	0.79	0.78
fine-tuned	0.69	0.95	0.8

Matching algorithm



Assigning
found named
entities to
specific
literary
character

*matching
algorithm*

Approximate text matching

Given a long text of length n and a comparatively short pattern of length m , both sequences over an alphabet find the text positions that match the pattern with at most k "errors".

Levenshtein distance

single-character operations required to
change one sequence of characters to the
other

The considered operations are insertion, deletion and substitution. Formally speaking the distance $d(x,y)$ between two strings x and y is the minimum number of such errors needed to convert one into the other.

- `distance("William Cohen", "Willliam Cohon")`

[illegible]

Example of approximate text matching in practice

Pattern	String	Regular string similarity	Partial string similarity
Elizabeth	Elizabeth Bennet	72%	100%
Lizzy	Elizabeth Bennet	19%	40%
Lizzy	Mr Fitzgerald Darcy	24%	40%

Known by
many names

Entity	Appearances
Elizabeth	635
Lizzy	96
Miss Bennet	72
Miss Elizabeth	12
Elizabeth Bennet	8

appearances of the references to
Elizabeth Bennet in the novel in different
configurations

Example of approximate text matching in practice

Pattern	String	Regular string similarity	Partial string similarity
Elizabeth	Elizabeth Bennet	72%	100%
Lizzy	Elizabeth Bennet	19%	40%
Lizzy	Mr Fitzgerald Darcy	24%	40%

Handling diminutives

aaron,erin,ronnie,ron
abbie,abby,abigail
abe,abraham,abram
abednego,bedney
abel,ebbie,ab,abe,eb
abiel,ab
abigail,nabby,abby,gail
abijah,ab,bige
abner,ab
abraham,ab,abe
abram,ab
absalom,app,ab,abbie

ada,addy
adaline,delia,lena,dell,addy,ada
adam,edie,ade
addy,adele
adela,della
adelaide,heidi,adele,dell,addy,della
adelbert,del,albert,delbert,bert
adele,dell
adeline,delia,lena,dell,addy,ada
adelphia,philly,delphia,adele,dell,ad
dy
adolphus,dolph,ado,adolph
adrian,rian

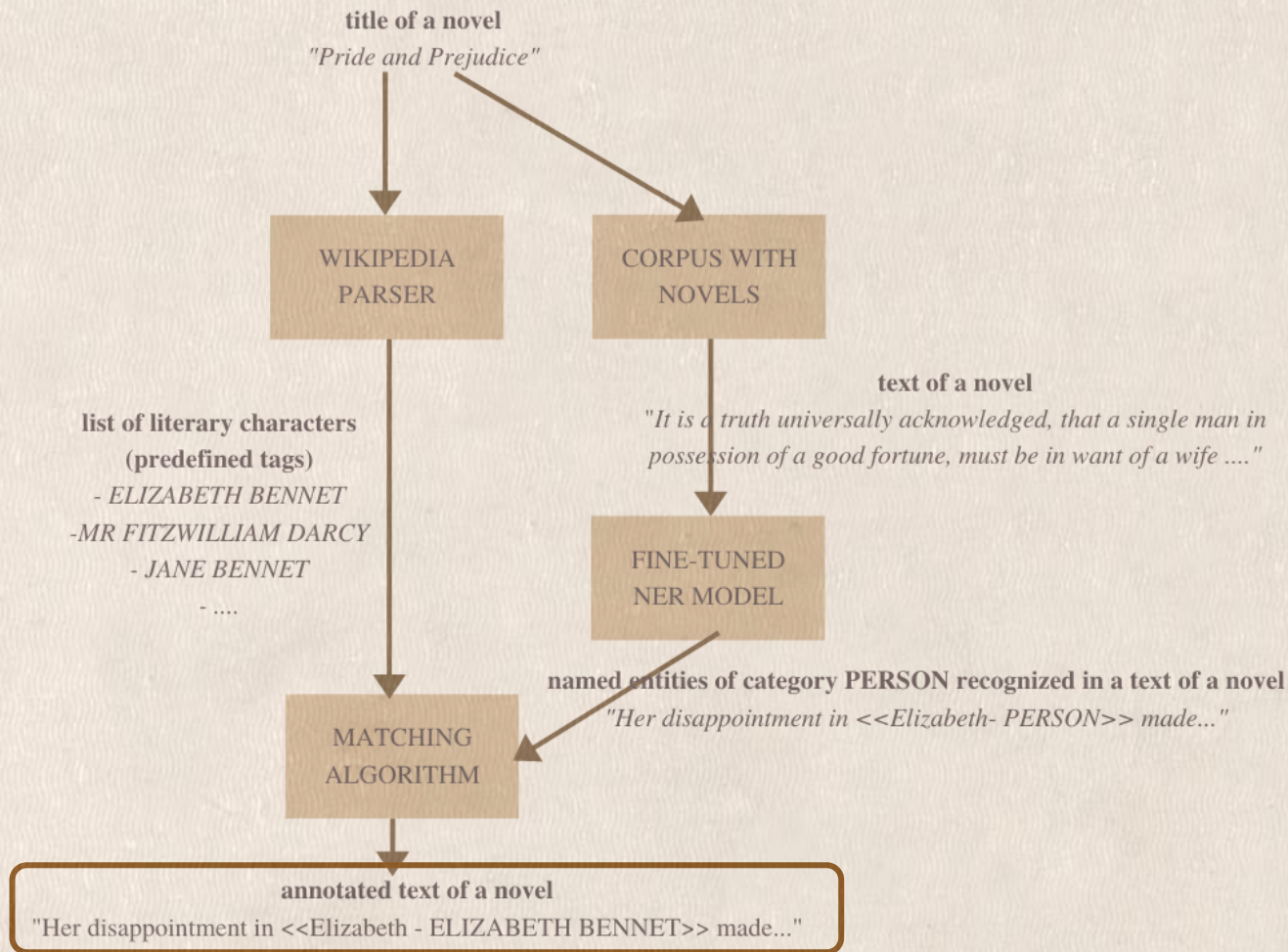
adrienne,adrian
agatha,aggy
agnes,inez,aggy,nessa
aileen,lena,allie
al,albert,bert,alex
alan,al
alanson,al,lanson
alastair,al
alazama,ali
albert,bert,al
alberta,bert,allie,bertie
aldo,al

One name,
many protagonists

Entity	Appearances
Bennet	323
Mrs. Bennet	153
Mr. Bennet	89
Miss Bennet	72

appearances of the entity **Bennet** in the
novel in different configurations

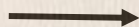
Analysis of *ProtagonistTagger* results



Protagonist Tagger results

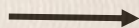
ProtagonistTagger results

Large testing set



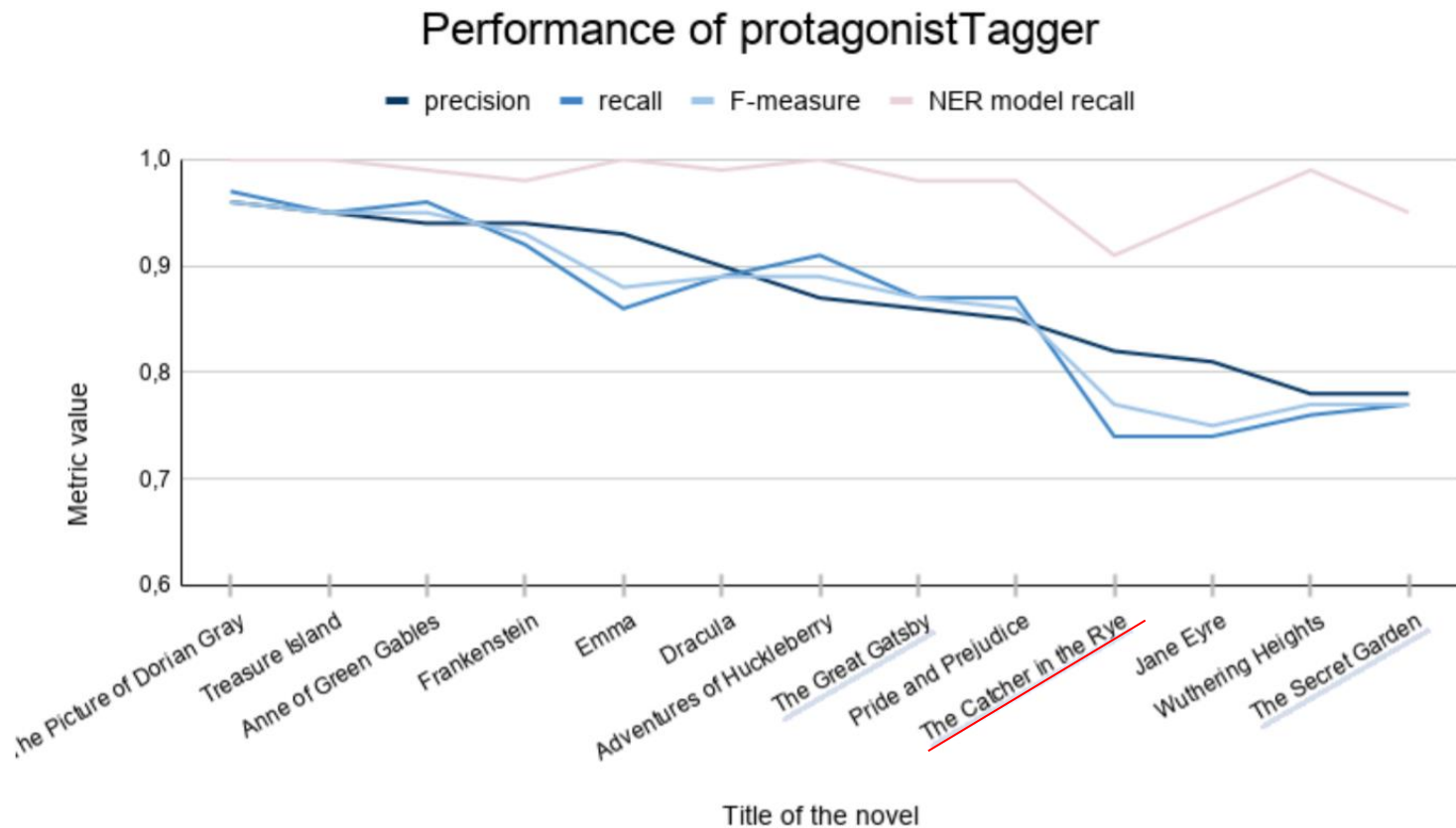
Novel title	precision	recall	F-measure
Pride and Prejudice	0.85	0.87	0.86
The Picture of Dorian Gray	0.96	0.97	0.96
Anne of Green Gables	0.94	0.96	0.95
Wuthering Heights	0.78	0.76	0.77
Jane Eyre	0.81	0.74	0.75
Frankenstein	0.94	0.92	0.93
Treasure Island	0.95	0.95	0.95
Adventures of Huckleberry Finn	0.87	0.91	0.89
Emma	0.93	0.86	0.88
Dracula	0.9	0.89	0.89
*** Overall results ***	0.89	0.88	0.88

*Small testing set
(totally new novels)*

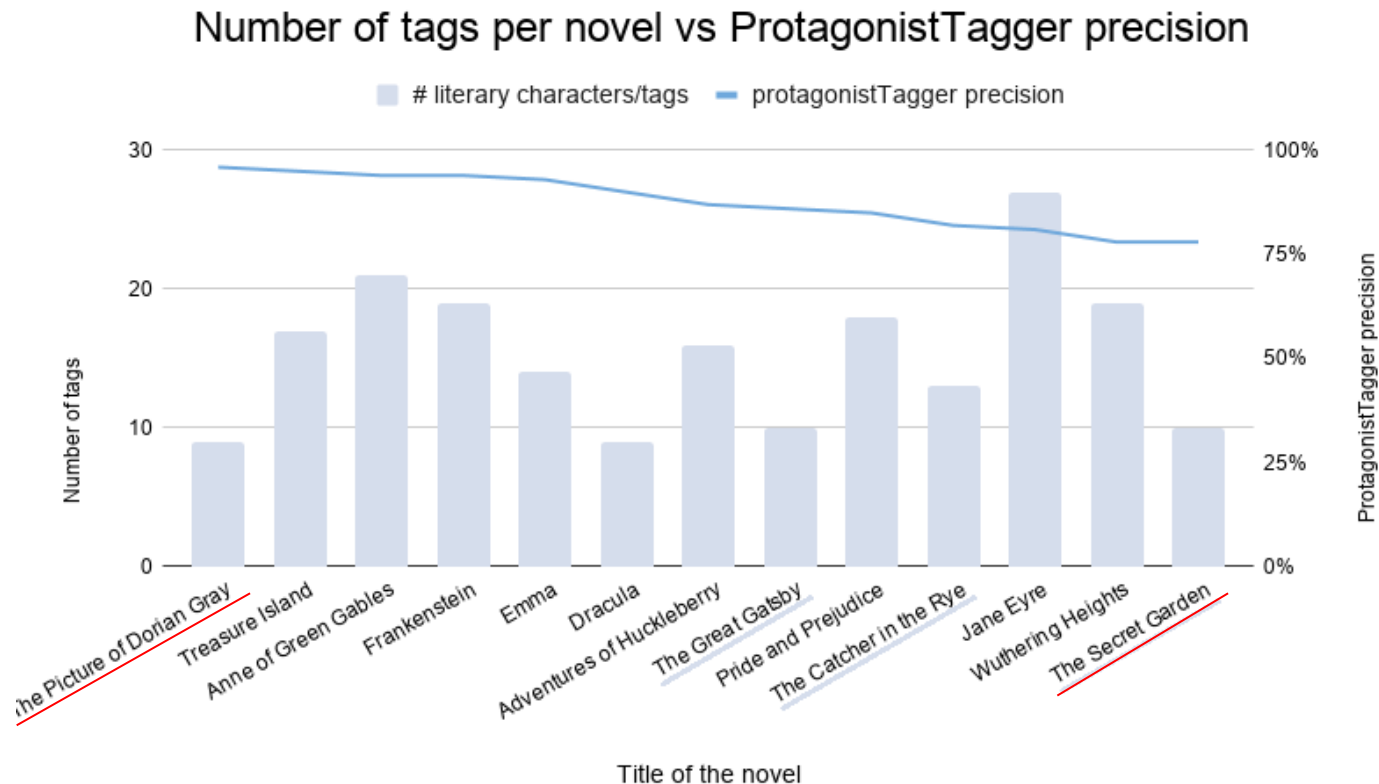


Novel title	precision	recall	F-measure
The Catcher in the Rye	0.82	0.74	0.77
The Great Gatsby	0.86	0.87	0.87
The Secret Garden	0.78	0.77	0.77
*** Overall results ***	0.82	0.8	0.81

ProtagonistTagger performance vs NER performance



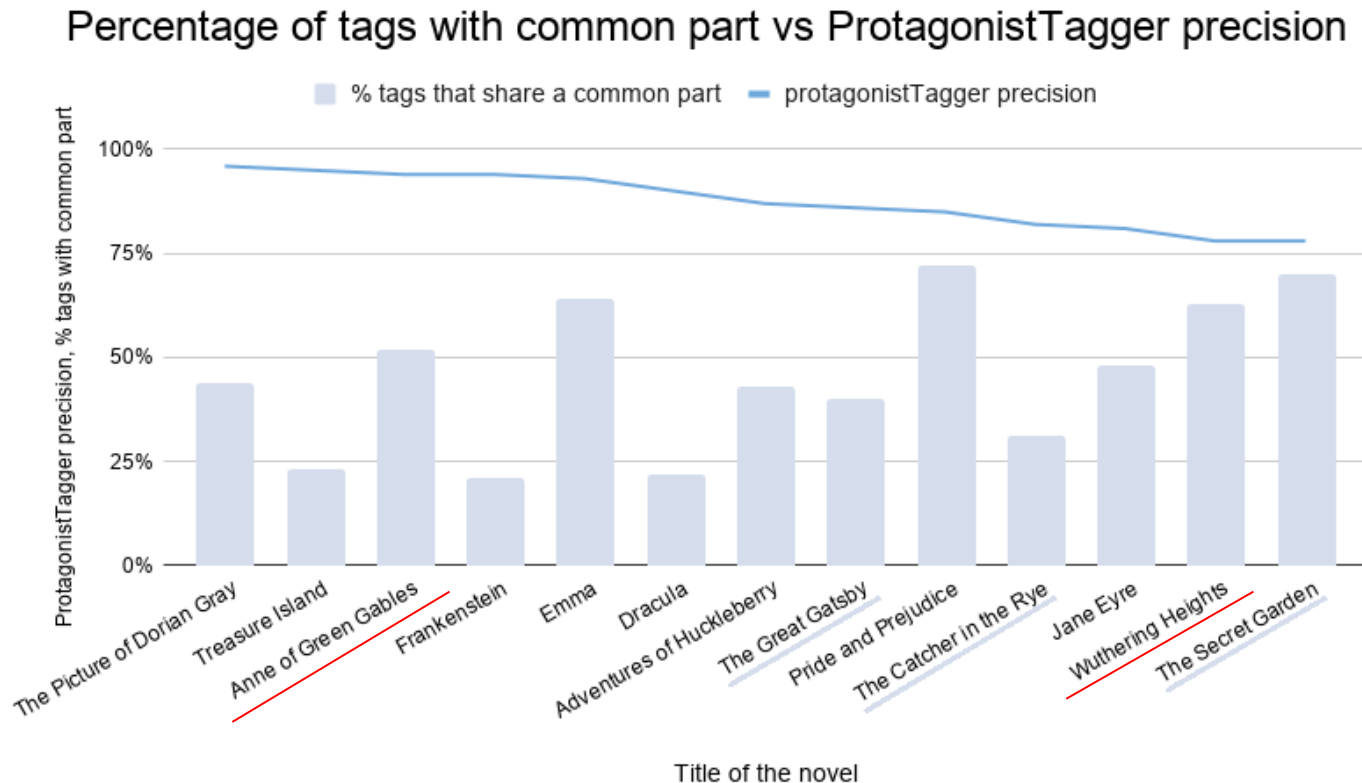
ProtagonistTagger performance vs number of tags

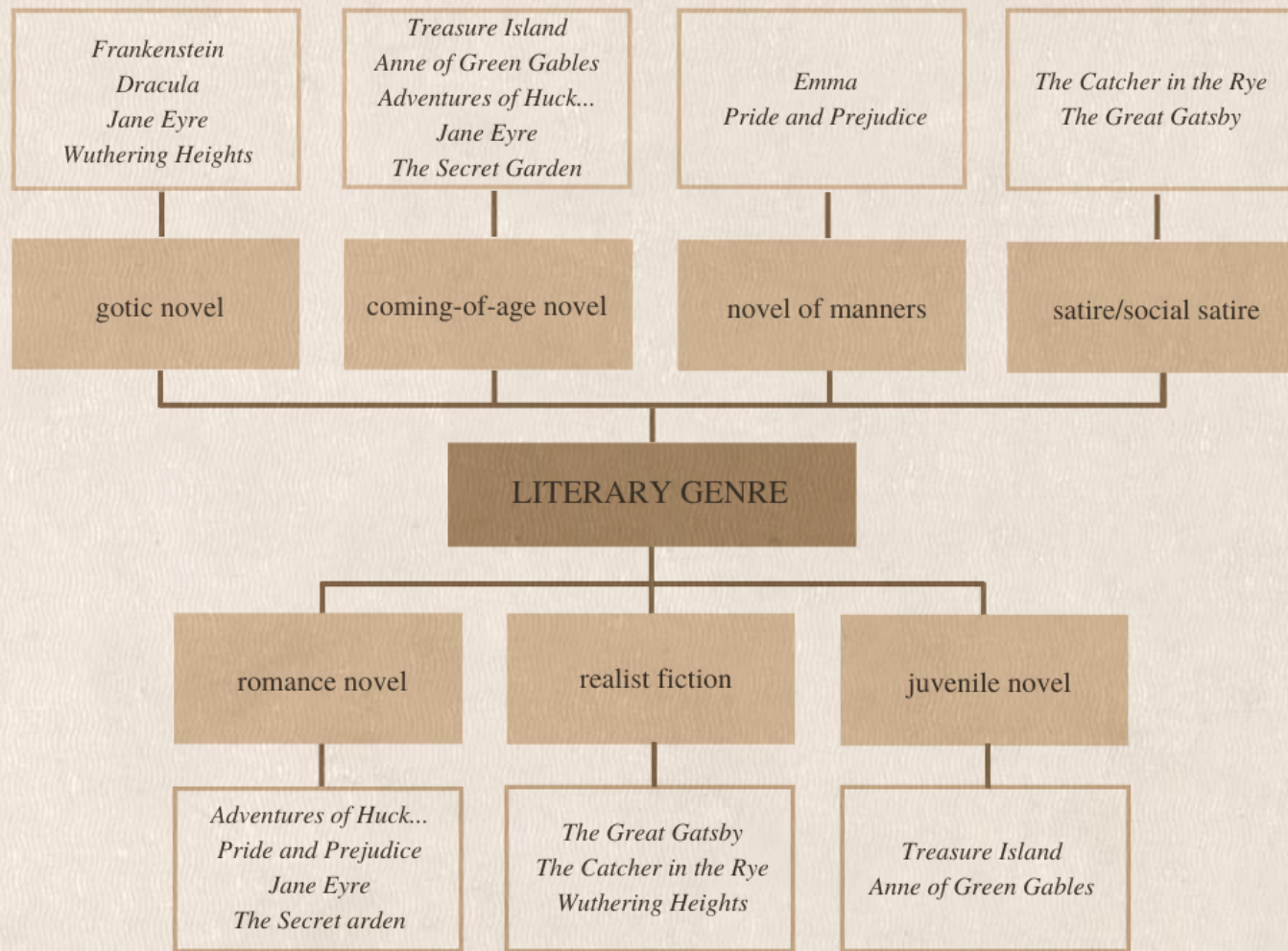


Title of the novel	# literary characters/tags	# tags that share a common part	% tags that share a common part
Pride and Prejudice	18	13	72%
The Picture of Dorian Gray	9	4	44%
Anne of Green Gables	21	11	52%
Wuthering Heights	19	12	63%
Jane Eyre	27	13	48%
Frankenstein	19	4	21%
Treasure Island	17	4	23%
Adventures of Huckleberry Finn	16	7	43%
Emma	14	9	64%
Dracula	9	2	22%
The Catcher in the Rye	13	4	31%
The Great Gatsby	10	4	40%
The Secret Garden	10	7	70%

Tags that
share a
common
part

ProtagonistTagger performance vs tags with common part

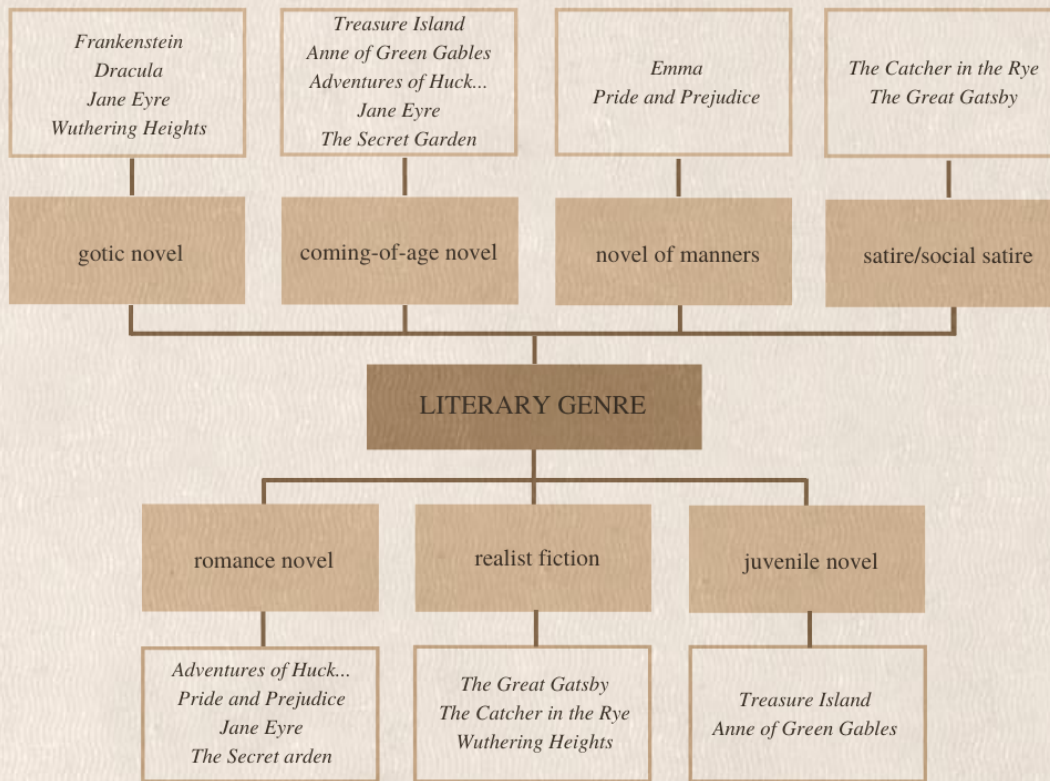




Literary
genre of
novels in
testing sets

Title of the novel	Literary genre	Style of the novel
The Picture of Dorian Gray	gothic novel, novel of manners, comedy	kept in dark mood; supernatural motifs; sardonic but comedic
Treasure Island	adventure story, coming-of-age novel, juvenile novel	revealing little emotions; pirate style of speech; focused on reporting events
Anne of Green Gables	coming-of-age novel, juvenile novel	poetic, descriptive, focused on emotions and inner life of the protagonists, as well as on the world around them
Frankenstein	gothic novel, science fiction	formal, elevated; using complex vocabulary
Emma	novel of manners, comedy	subtle; simple but direct; witty, sharp, epigrammatic, abstract; focused on dialogues
Dracula	gothic fiction, horror	epistolary (novel is composed of diary entries, letters, etc.); straightforward
Adventures of Huckleberry Finn	picaresque novel, coming-of-age novel, romance novel	written in the vernacular of the characters; casual, sometimes even incorrect way of speaking in dialogues
Pride and Prejudice	romance novel, novel of manners, comedy	exaggerated, ironic and witty; focused on dialogues
Jane Eyre	romance novel, gothic novel, coming-of-age novel	descriptive and formal; long sentences; verbiage and lengthy syntax
Wuthering Heights	tragedy, gothic novel, realist fiction	designed to horrify and fascinate; incorporating supernatural elements; novel kept in dark, foreboding atmosphere
The Secret Garden	coming-of-age novel, romance, novel of ideas	flowery and rich; descriptive; using multiple adjectives
The Catcher in the Rye	coming-of-age novel, realist fiction, satire	vernacular style with slang and curse words; hyperbolic; using generalizations
The Great Gatsby	tragedy, realism, modernism, social satire	sophisticated, elegiac, wry; including extended metaphors and poetic language; incorporating sharp and sardonic humor

Literary genre and style of novels in testing sets



Literary genre vs *protagonistTagger* performance

Novel title	precision	recall	F-measure
Pride and Prejudice	0.85	0.87	0.86
The Picture of Dorian Gray	0.96	0.97	0.96
Anne of Green Gables	0.94	0.96	0.95
Wuthering Heights	0.78	0.76	0.77
Jane Eyre	0.81	0.74	0.75
Frankenstein	0.94	0.92	0.93
Treasure Island	0.95	0.95	0.95
Adventures of Huckleberry Finn	0.87	0.91	0.89
Emma	0.93	0.86	0.88
Dracula	0.9	0.89	0.89
*** Overall results ***	0.89	0.88	0.88

Novel title	precision	recall	F-measure
The Catcher in the Rye	0.82	0.74	0.77
The Great Gatsby	0.86	0.87	0.87
The Secret Garden	0.78	0.77	0.77
*** Overall results ***	0.82	0.8	0.81

Conclusions

- high complexity of the names appearing in the novels
- relatively low performance of the standard NER models on novels
- two different methods for creating training set for fine-tuning NER were required
- recall above 95% for fine-tuned NER
- **the precision and the recall of the *ProtagonistTagger* above 80% in case of almost all analyzed novels**
- tool's performance depends on the type of the novel and the NER model's performance
- from linguistic point of view the number of literary characters, the percentage of the literary characters with the same name or surname and the literary genre of the novels influence the tool's performance

Main contributions in the research

- the in-depth analysis of named entities of category person appearing in literary texts,
- the *protagonistTagger* tool for ontology population and annotating people in long, complex texts,
- manually annotated extract from the novels to measure the performance of *protagonistTagger*,
- NER benchmark datasets for entities (instances) of category person, annotated also with full names of literary characters,
- a corpus of novels annotated with *protagonistTagger*.

Future work

- applying the created corpus for more detailed analysis of the novels
 - detection of relationships between literary characters
 - sentiment-based analysis of literary characters
- *use case* in non-literary domain (eg. social media)
 - investigating human opinions
 - sentiment analysis

Thank you for your attention

Bibliography

- Apoorv Agarwal et al. "Social network analysis of Alice in Wonderland". In: Proceedings of the NAACL-HLT 2012 Workshop on computational linguistics for literature. 2012, pp. 88– 96.
- Alan Akbik, Tanja Bergmann, and Roland Vollgraf. "Pooled contextualized embeddings for named entity recognition". In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019, pp. 724–728.
- David Bamman, Ted Underwood, and Noah A Smith. "A bayesian mixed effects model of literary character". In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014, pp. 370–379.
- Daniel M Bikel, Richard Schwartz, and Ralph M Weischedel. "An algorithm that learns what's in a name". In: Machine learning. 1999, pp. 211–231.
- Julian Brooke, Adam Hammond, and Graeme Hirst. "GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus". In: Proceedings of the Fourth Workshop on Computational Linguistics for Literature. 2015, pp. 42–47.
- Snigdha Chaturvedi, Mohit Iyyer, and Hal Daumé III. "Unsupervised Learning of Evolving Relationships Between Literary Characters." In: AAAI. 2017, pp. 3159–3165.
- Snigdha Chaturvedi et al. "Modeling evolving relationships between characters in literary novels". In: Thirtieth AAAI Conference on Artificial Intelligence. 2016.
- Davide Chicco and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: BMC genomics. Vol. 21. 1. Springer, 2020, p. 6.
- Jason PC Chiu and Eric Nichols. "Named entity recognition with bidirectional LSTM CNNs". In: Transactions of the Association for Computational Linguistics. 2016, pp. 357– 370
- Prabhakar Raghavan Christopher D. Manning and Hinrich Schütze. Introduction to information retrieval. 2010.
- Adam Cohen. "FuzzyWuzzy: Fuzzy string matching in python". In: ChairNerd Blog. Vol. 22. 2011.
- Michael Collins and Yoram Singer. "Unsupervised models for named entity classification". In: Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 1999.
- Ronan Collobert et al. "Natural language processing (almost) from scratch". In: Journal of machine learning research. 2011, pp. 2493–2537.
- Paul T Costa Jr and Robert R McCrae. The Revised NEO Personality Inventory (NEOPI-R). Sage Publications, Inc, 2008.
- David Elson, Nicholas Dames, and Kathleen McKeown. "Extracting social networks from literary fiction". In: Proceedings of the 48th annual meeting of the association for computational linguistics. 2010, pp. 138–147.
- Lucie Flekova and Iryna Gurevych. "Personality profiling of fictional characters using sense-level links between lexical resources". In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015, pp. 1805–1816.
- Filip Graliski et al. "GEval: Tool for Debugging NLP Datasets and Models". In: Proceedings of the 2019 ACL Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP. Florence, Italy: Association for Computational Linguistics, 2019, pp. 254–262.
- Adrian Groza and Lidia Corde. "Information retrieval in folk tales using natural language processing". In: 2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP). IEEE. 2015, pp. 59–66.
- GutenTag - web application. <https://gutentag.sdsu.edu/>.
- Adam Hammond and Julian Brooke. GutenTag: A User-Friendly, Open-Access, OpenSource System for Reproducible Large-Scale Computational Literary. 2017.
- Michael Hart. "The history and philosophy of Project Gutenberg". In: Project Gutenberg. 1992.
- Nancy Ide and James Pustejovsky. Handbook of linguistic annotation. Springer, 2017.
- Ridong Jiang, Rafael E Banchs, and Haizhou Li. "Evaluating and combining name entity recognition systems". In: Proceedings of the Sixth Named Entity Workshop. 2016, pp. 21–27.

Bibliography

- Evgeny Kim and Roman Klinger. "A survey on sentiment and emotion analysis for computational literary studies". In: arXiv preprint arXiv:1808.03137. 2018.
- Guillaume Lample et al. "Neural Architectures for Named Entity Recognition". In: Proceedings of the the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016, pp. 260–270.
- Gonzalo Navarro. "A guided tour to approximate string matching". In: ACM computing surveys (CSUR). Vol. 33. 1. ACM New York, NY, USA, 2001, pp. 31–88.
- Gonzalo Navarro et al. "Indexing methods for approximate string matching". In: IEEE Data Eng. Bull. Vol. 24. 4. Citeseer, 2001, pp. 19–27.
- Andrew J Reagan et al. "The emotional arcs of stories are dominated by six basic shapes". In: EPJ Data Science. Vol. 5. 1. SpringerOpen, 2016, pp. 1–12.
- Tim Rocktäschel, Michael Weidlich, and Ulf Leser. "ChemSpot: a hybrid system for chemical named entity recognition". In: Bioinformatics. Oxford University Press, 2012, pp. 1633–1640.
- Xavier Schmitt et al. "A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate". In: 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE. 2019, pp. 338–343.
- Isabel Segura-Bedmar, Paloma Martnez, and Mara Herrero-Zazo. "SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)". In: 2nd Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013). 2013, pp. 341–350.
- Mohammad Golam Sohrab and Makoto Miwa. "Deep Exhaustive Model for Nested Named Entity Recognition". In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018.
- SpaCy NER Annotator. <https://manivannanmurugavel.github.io/annotating-tool/spacy-ner-annotator/>.
- Bhargav Srinivasa-Desikan. Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing Ltd, 2018.
- Tomasz Stanislawek et al. "Named Entity Recognition - Is There a Glass Ceiling?" In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). Association for Computational Linguistics, 2019, pp. 624–633.
- Hardik Vala et al. "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. 2015, pp. 769–774.
- Ralph Weischedel, Martha Palmer, and Mitchell Marcus. OntoNotes Release 5.0 - Linguistic Data Consortium. 2013.
- Vikas Yadav and Steven Bethard. "A Survey on Recent Advances in Named Entity Recognition from Deep Learning models". In: Proceedings of the 27th International Conference on Computational Linguistics. 2018, pp. 2145–2158.
- Vikas Yadav, Rebecca Sharp, and Steven Bethard. "Deep affix features improve neural named entity recognizers". In: Proceedings of the 7th Joint Conference on Lexical and Computational Semantics. 2018, pp. 167–172.
- Xiaodong Yu et al. "On the Strength of Character Language Models for Multilingual Named Entity Recognition". In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018.
- Albin Zehe et al. "Prediction of happy endings in German novels based on sentiment information". In: 3rd Workshop on Interactions between Data Mining and Natural Language Processing, Riva del Garda, Italy. 2016.