Automatic preparation of sets for training models for recognizing names of the characters on the example of a novel

Weronika Łajewska
supervised by PhD Anna Wróblewska

Main goal of the project

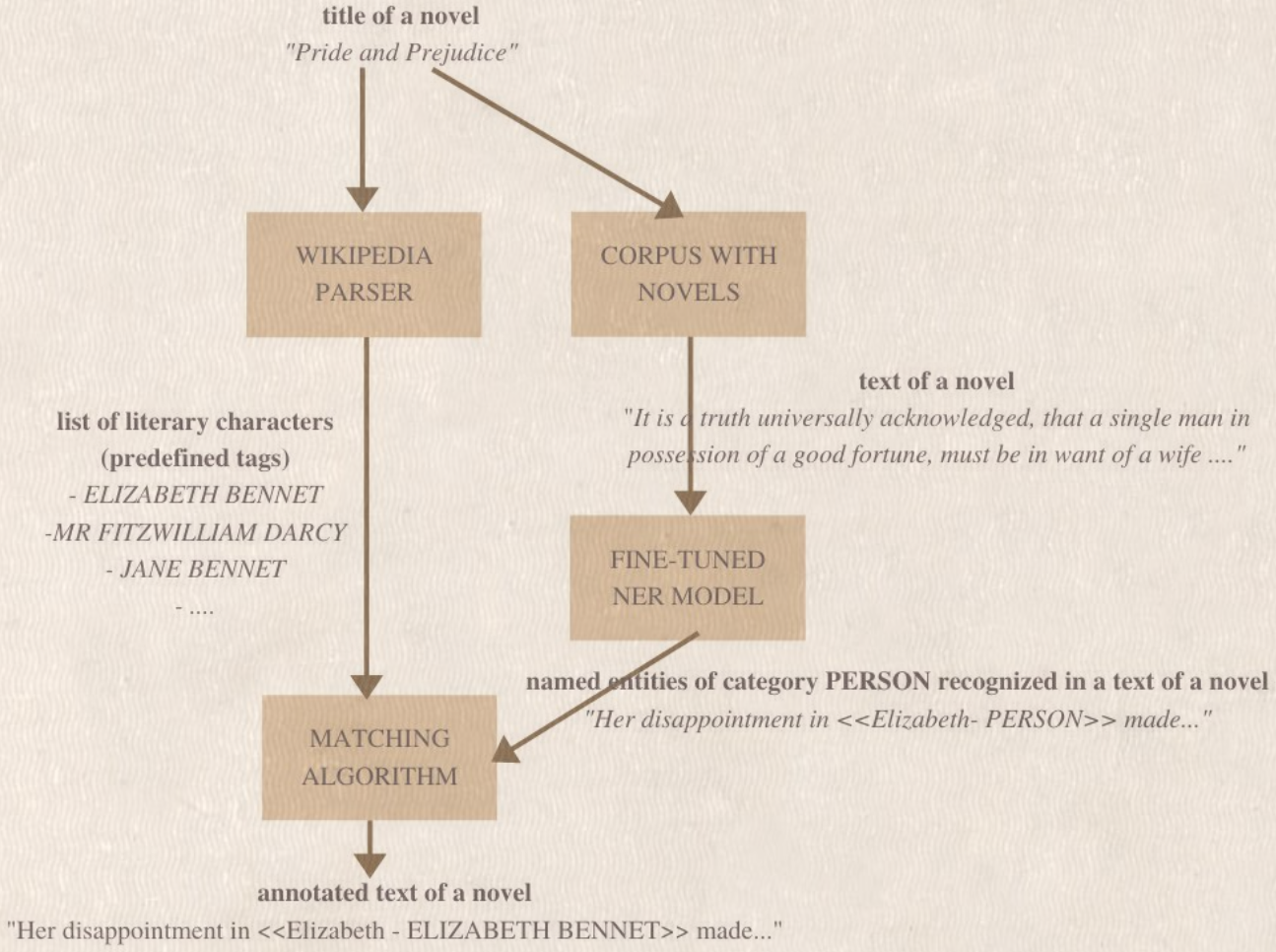
reasonably big corpus with annotated novels in which every protagonist is annotated with his proper name

model for recognizing appearances of protagonists in a novel

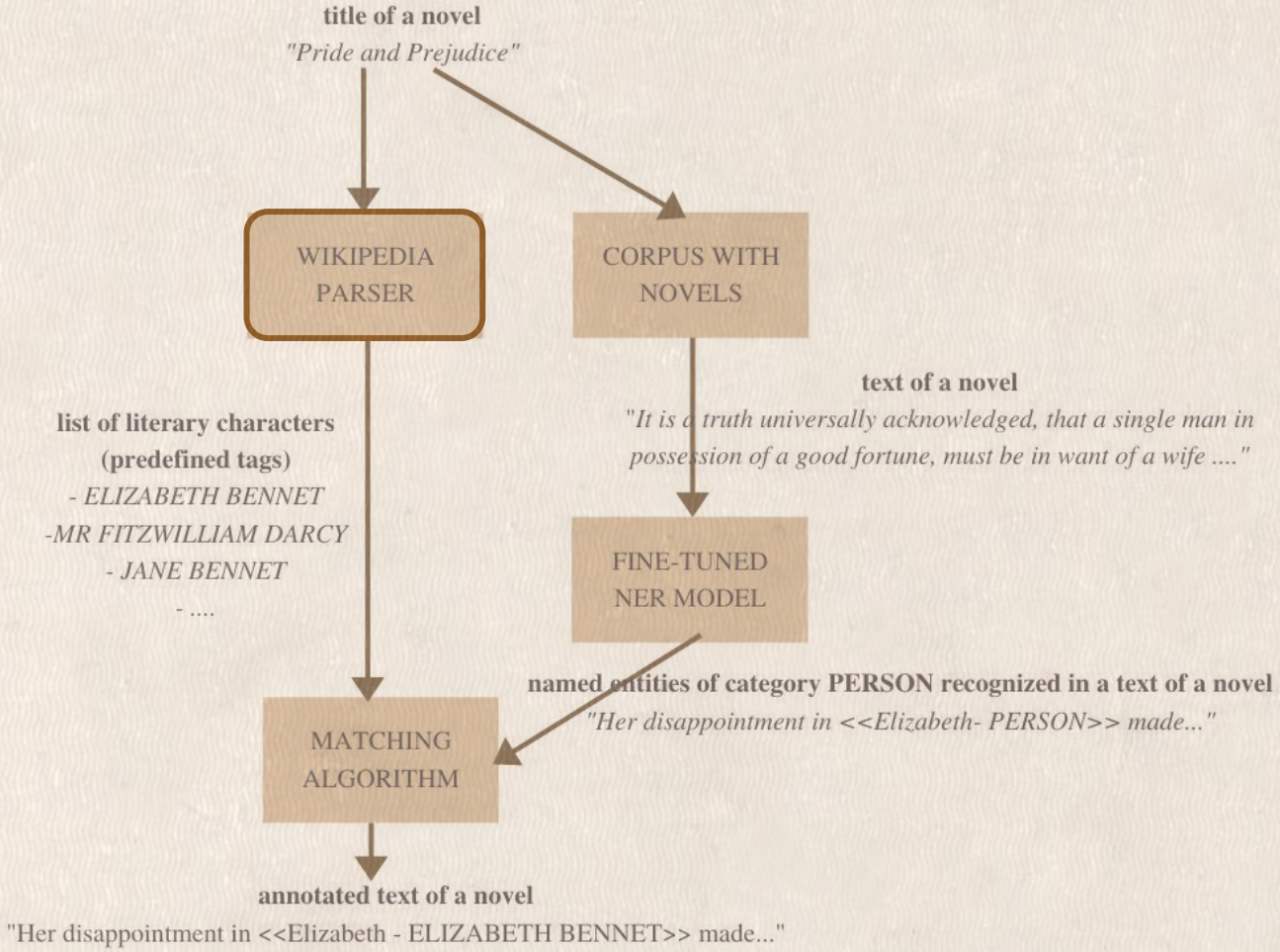
Exemplary output of *ProtagonistTagger*

"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."*

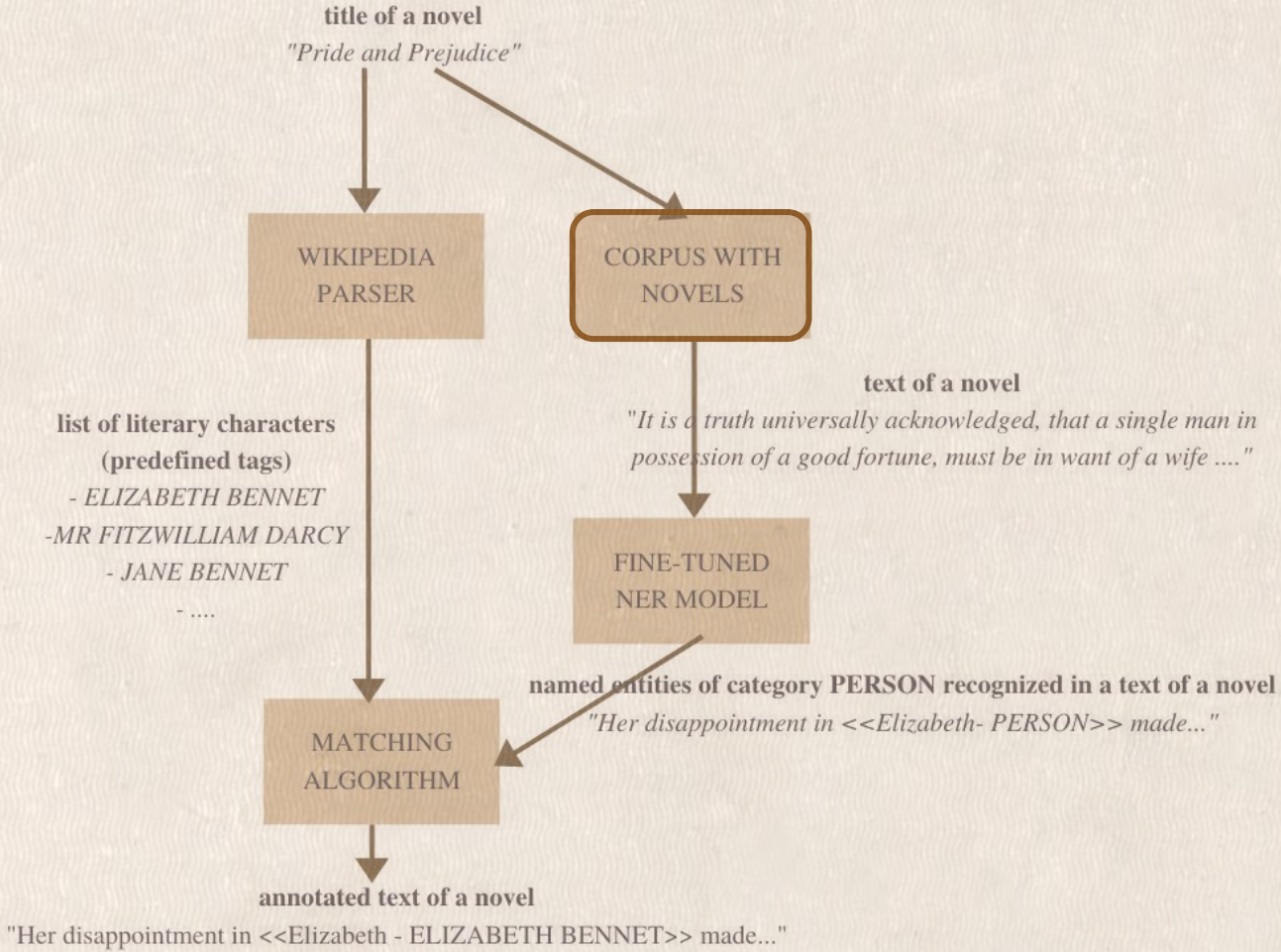
Protagonist Tagger



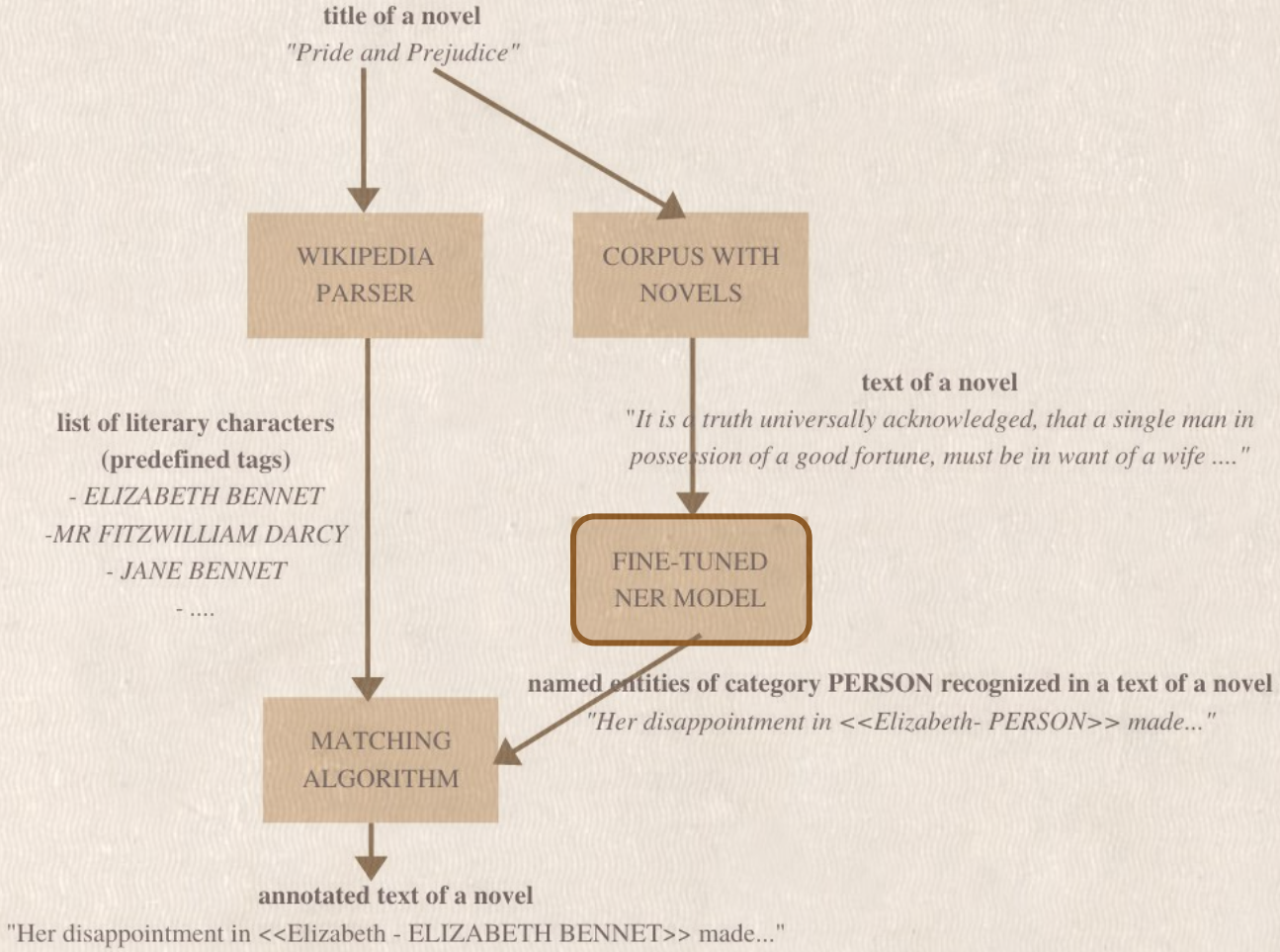
Protagonist Tagger



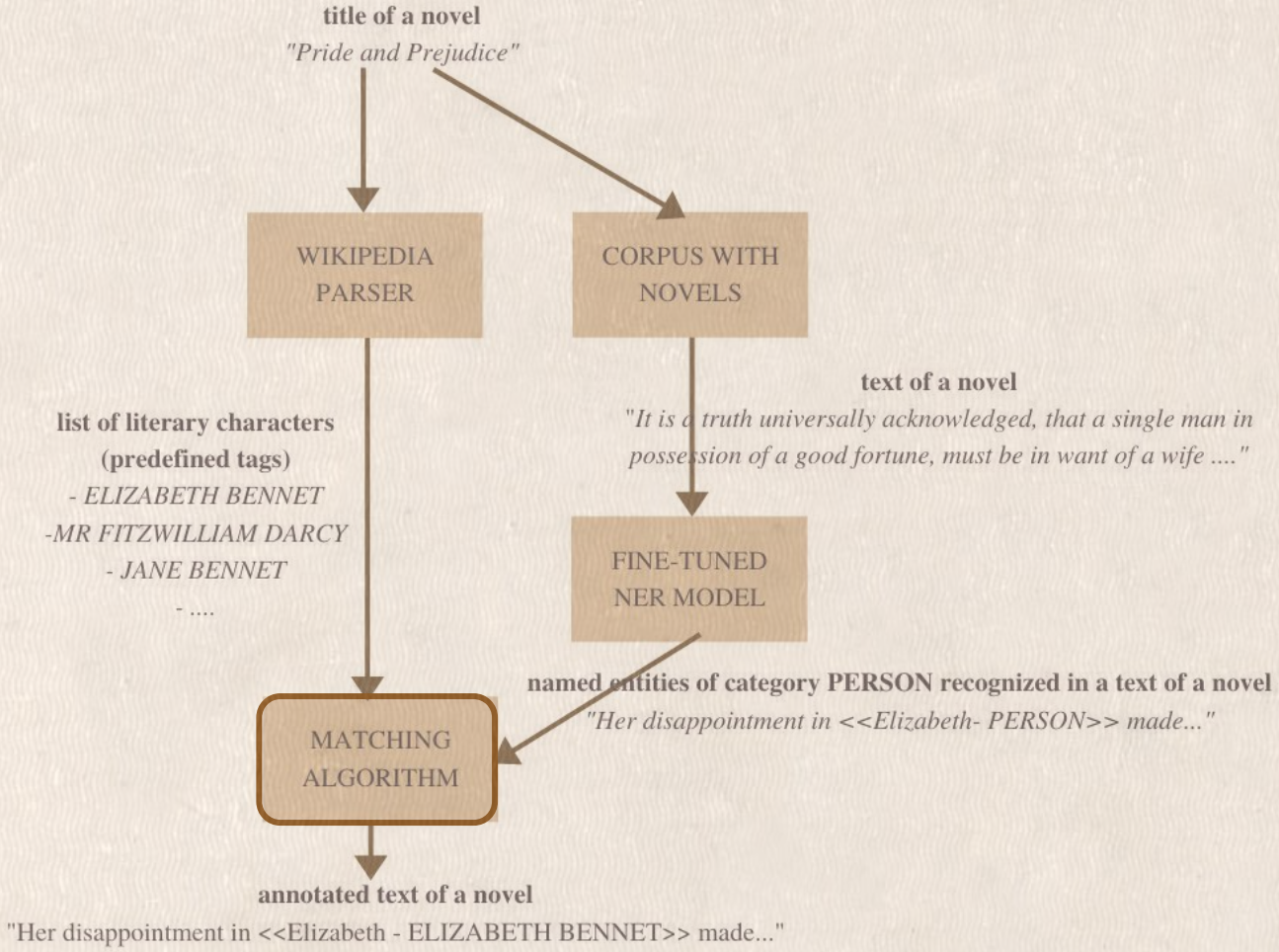
Protagonist Tagger

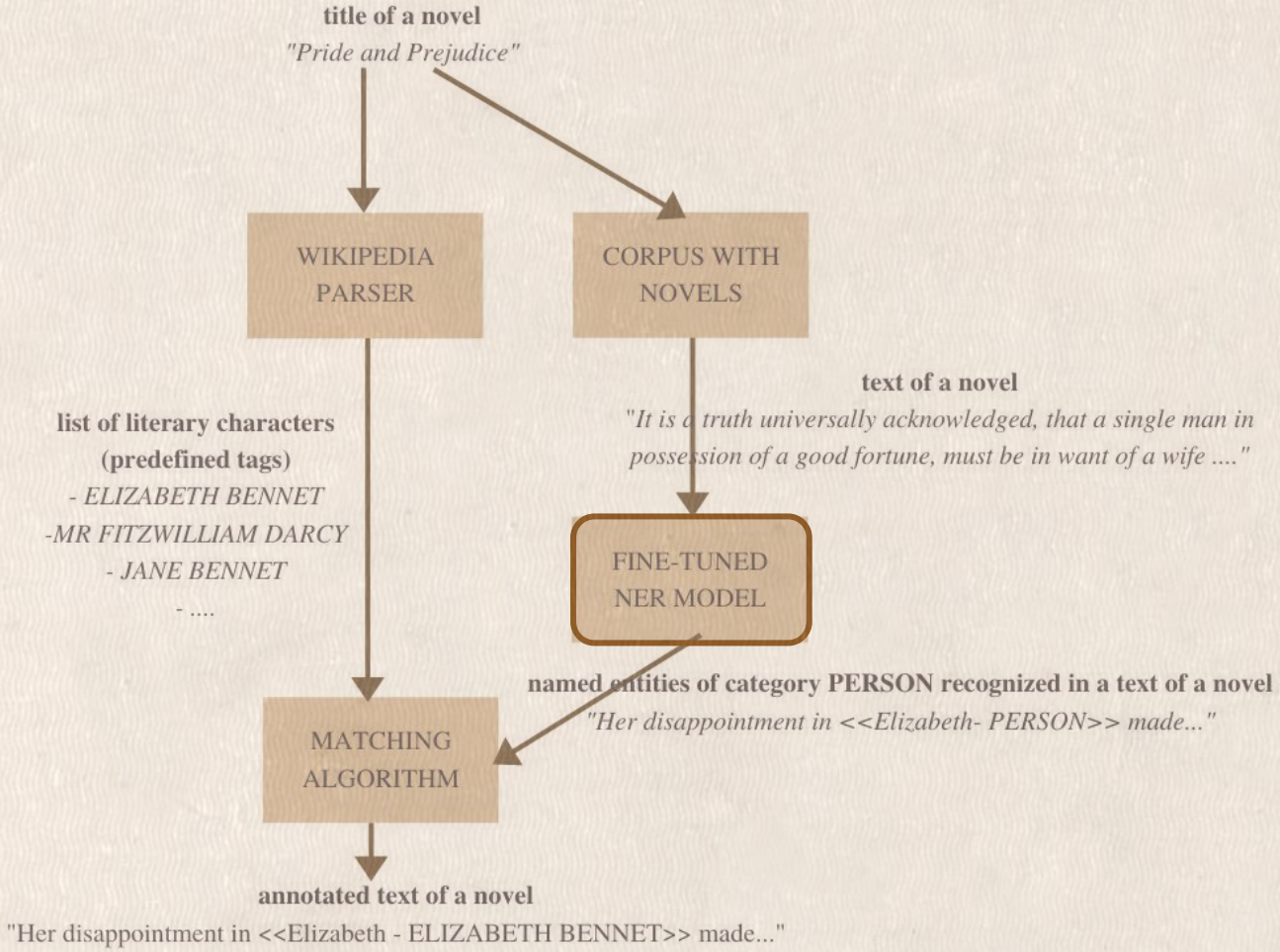


Protagonist Tagger



Protagonist Tagger





Fine-tuning NER

Imperfections of NER in novels

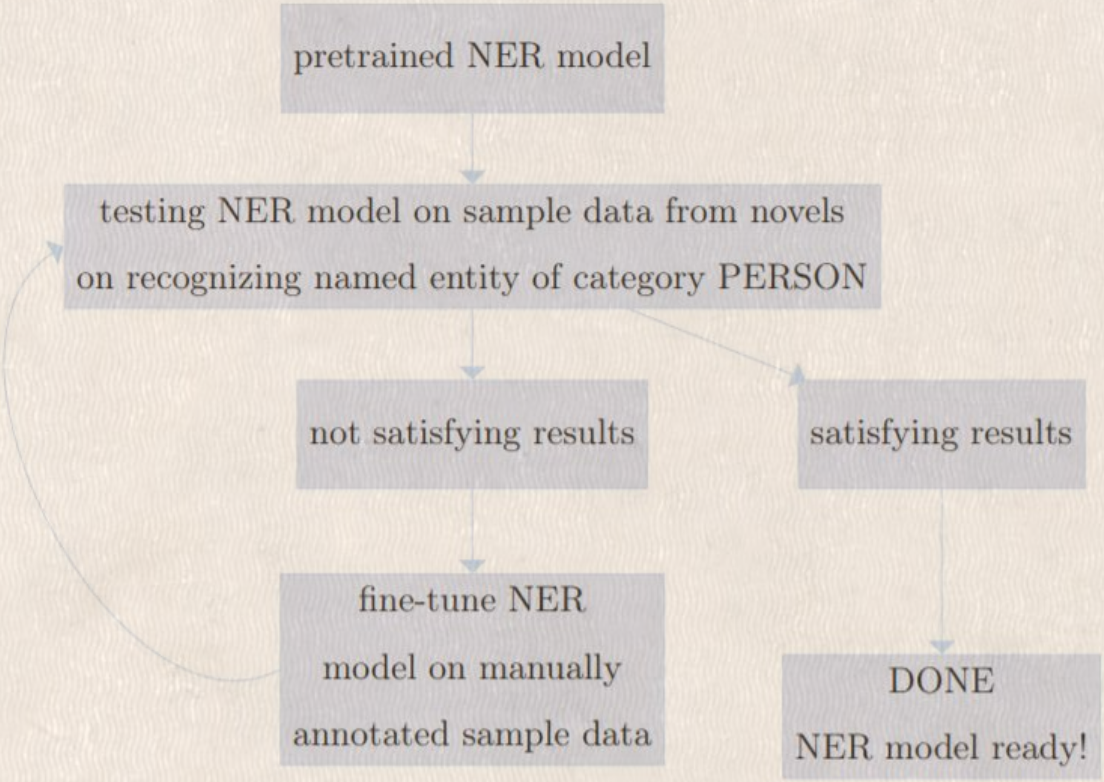
Novel title	precision	recall	F-measure	support
The Picture of Dorian Gray	0.69	0.41	0.51	90
Frankenstein	0.91	0.62	0.74	93
Treasure Island	0.75	0.66	0.7	97
Emma	0.84	0.77	0.81	115
Jane Eyre	0.86	0.78	0.82	97
Wuthering Heights	0.95	0.87	0.91	108
Pride and Prejudice	0.85	0.87	0.86	124
Dracula	0.86	0.94	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

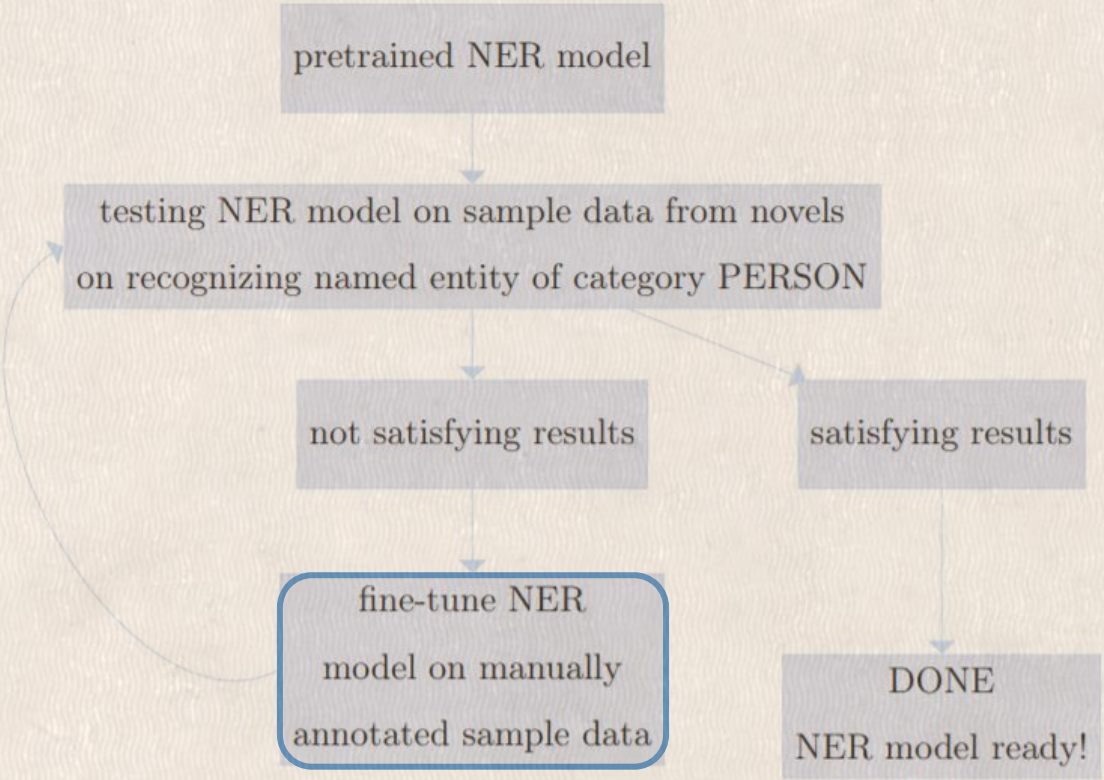
Not recognized named entities

Novel title	Named entities of category *PERSON* 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

ORG



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 *PERSON* 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."*

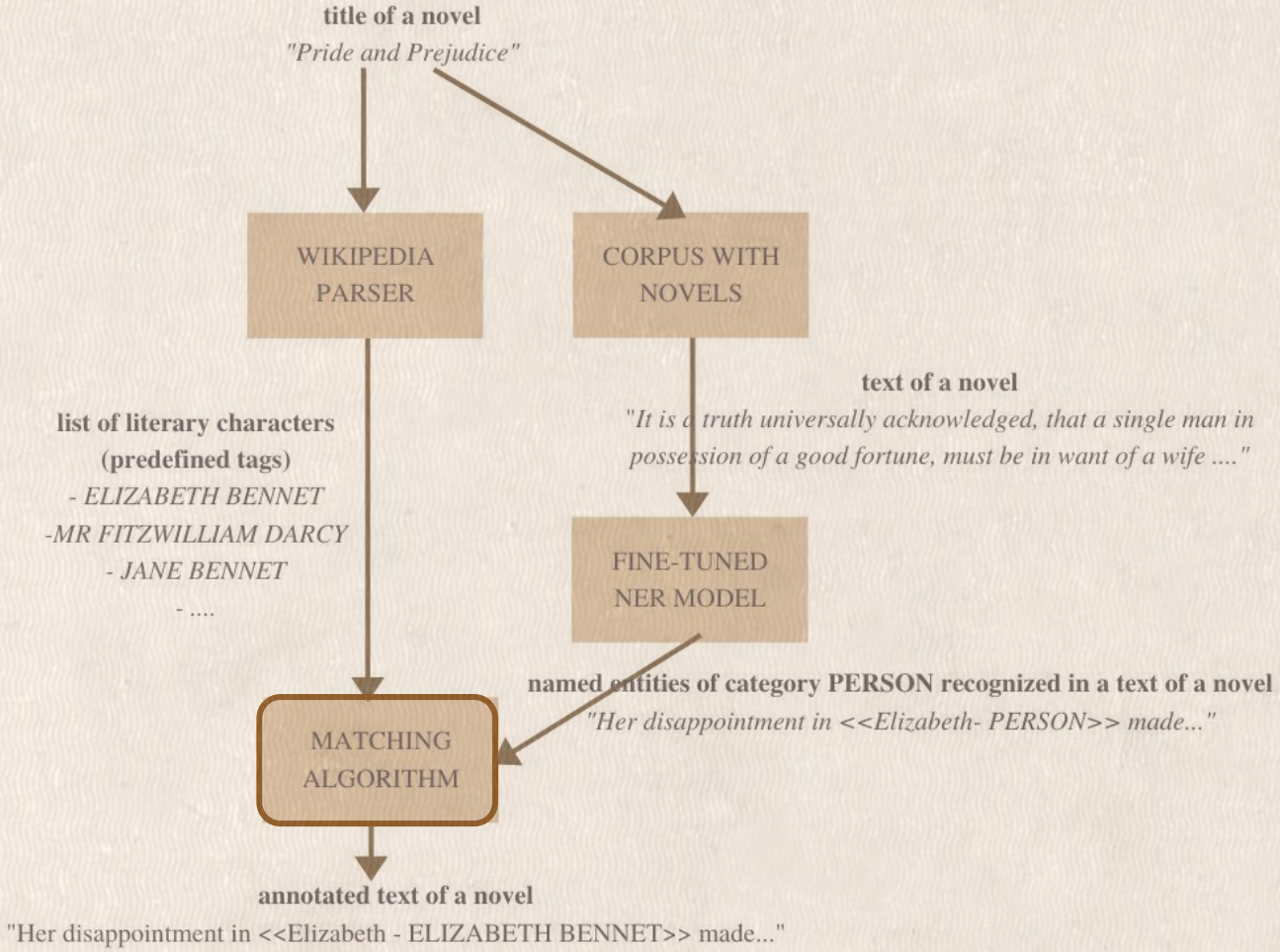
*"**Deborah**'s delicate sense of honour would not allow her to speak to **Harvey** privately of what **Lydia** had let fall; **Harvey** was glad of it; till it appeared whether her inquiries would receive any satisfaction, she had rather be without a confidante."*

Fine-tuned NER models performance

Large testing set →

Small testing set (totally new novels) →

NER model	precision	recall	F-measure
standard	0.84	0.8	0.82
fine-tuned	0.77	0.99	0.87
standard	0.78	0.79	0.78
fine-tuned	0.69	0.95	0.8



Assigning found named entities to specific literary character

***

matching algorithm

Approximate text matching

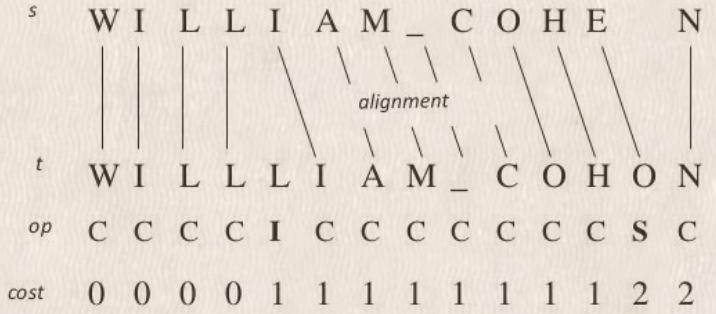
Given a long text of length n and a comparatively short pattern of length m, both sequences over an alphabet find the text positions that match the pattern with at most k "errors".

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”)



Known by many names

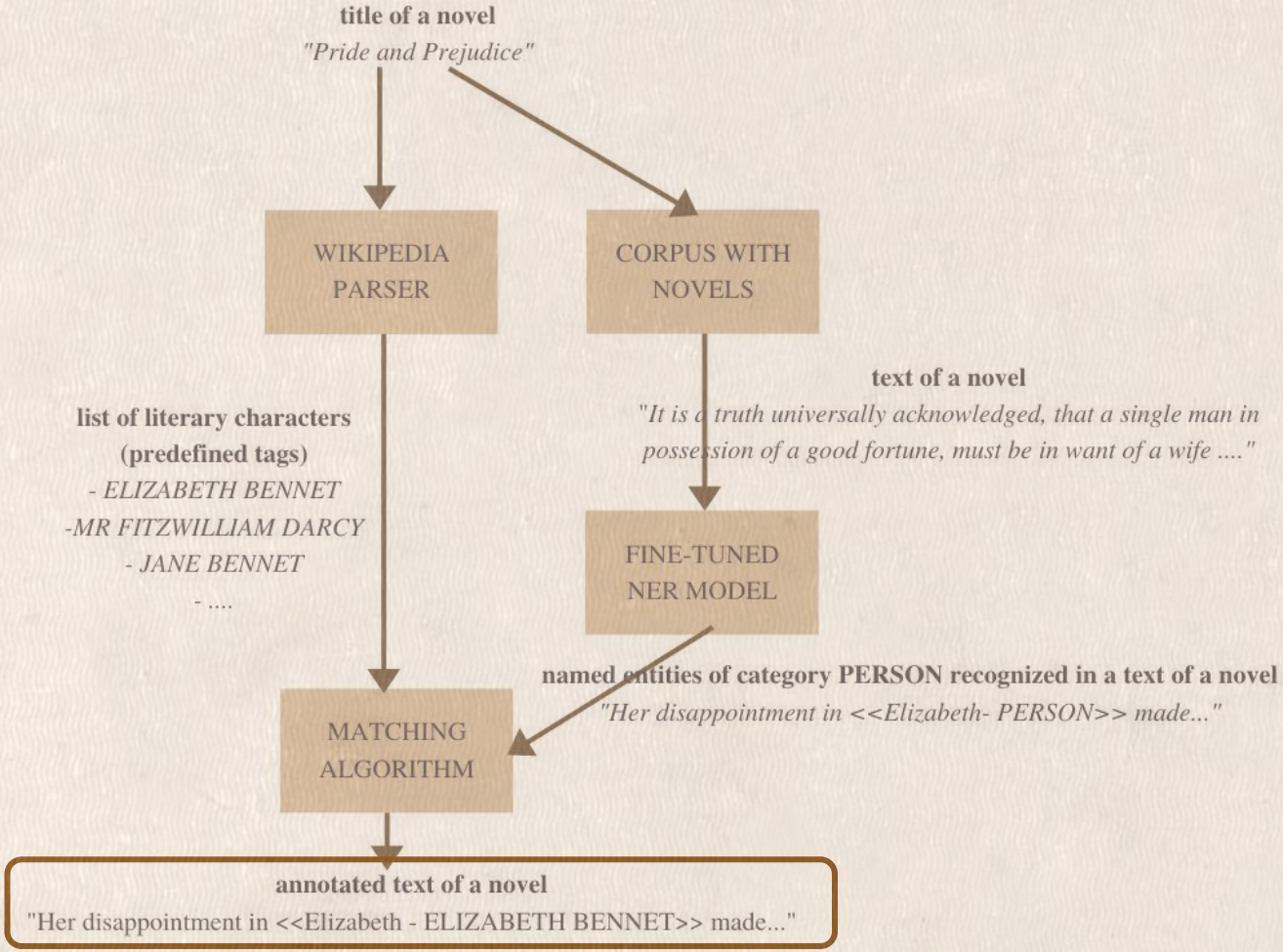
Entity	Appearances
Elizabeth	635
Lizzy	96
Miss Bennet	72
Miss Elizabeth	12
Elizabeth Bennet	8

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

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



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

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

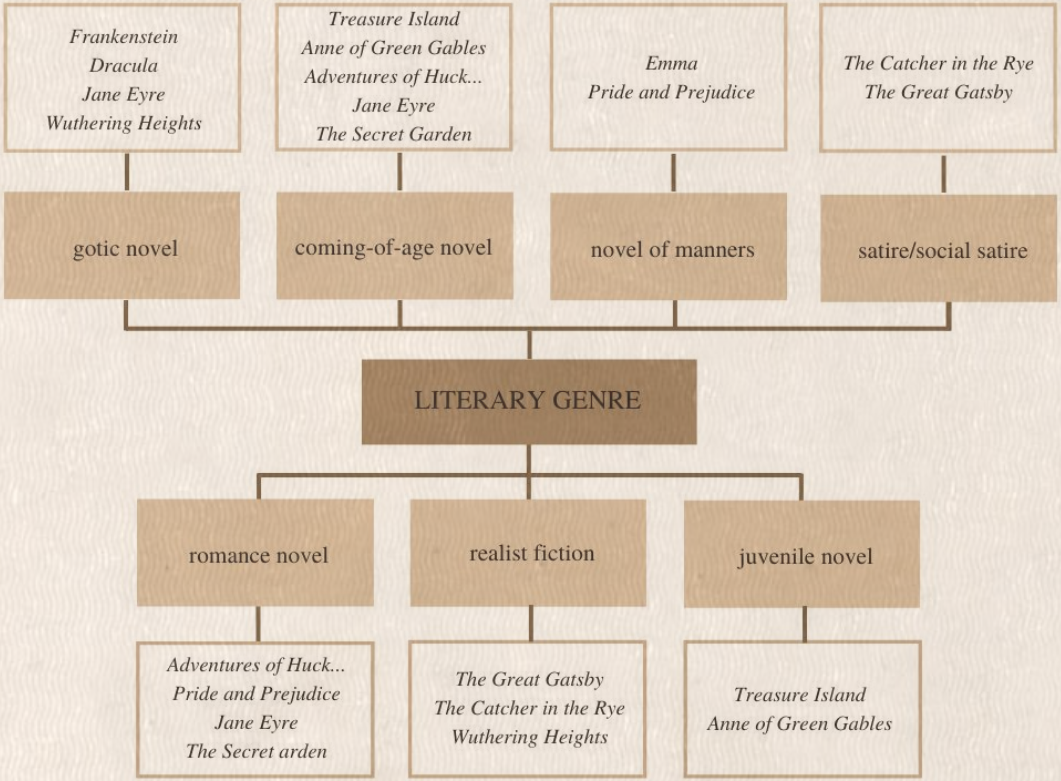
Future work

- 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
  - investigating human opinions
  - sentiment analysis

Thank you for your attention

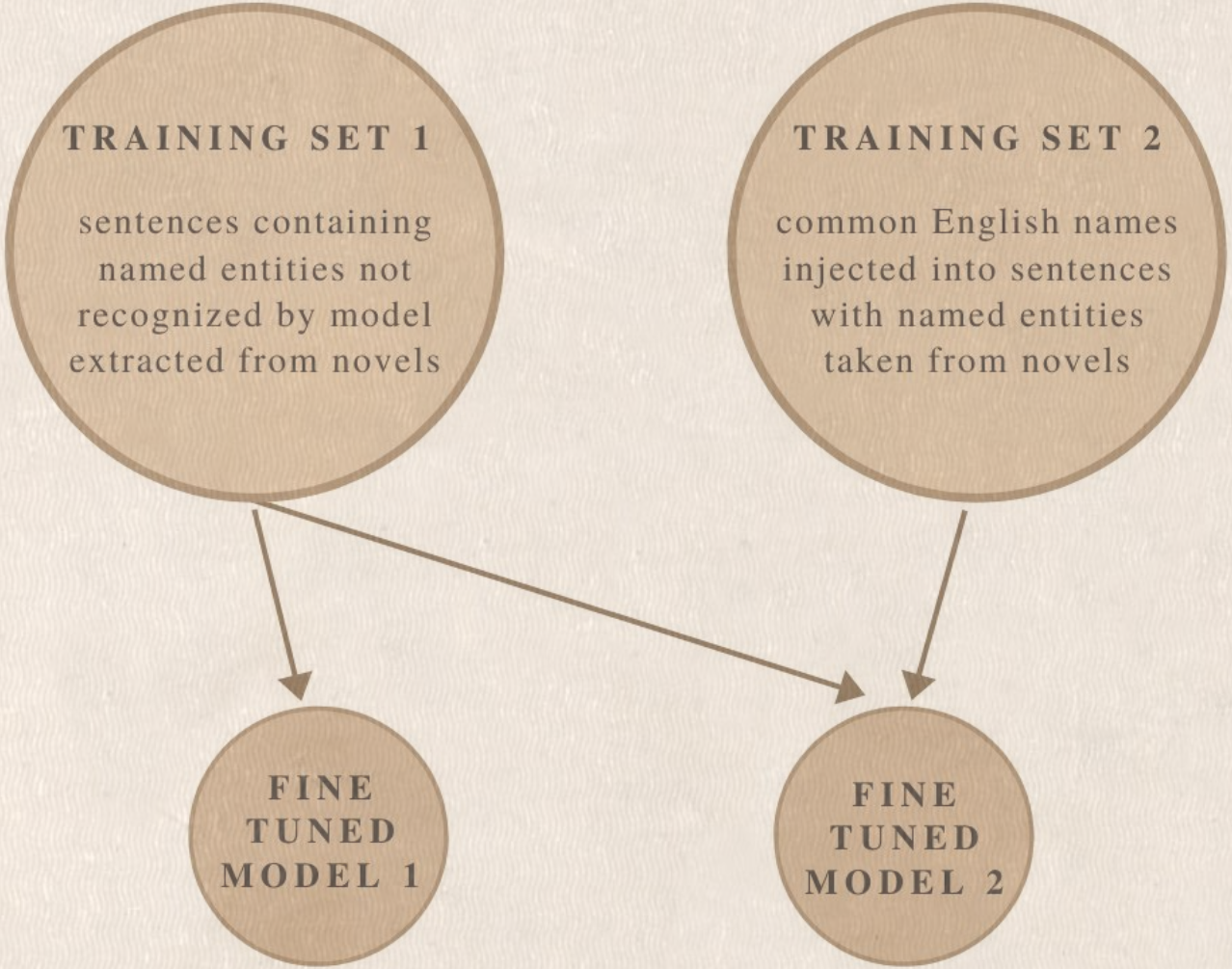
***



Literary genre vs *protagonistTagger* performance

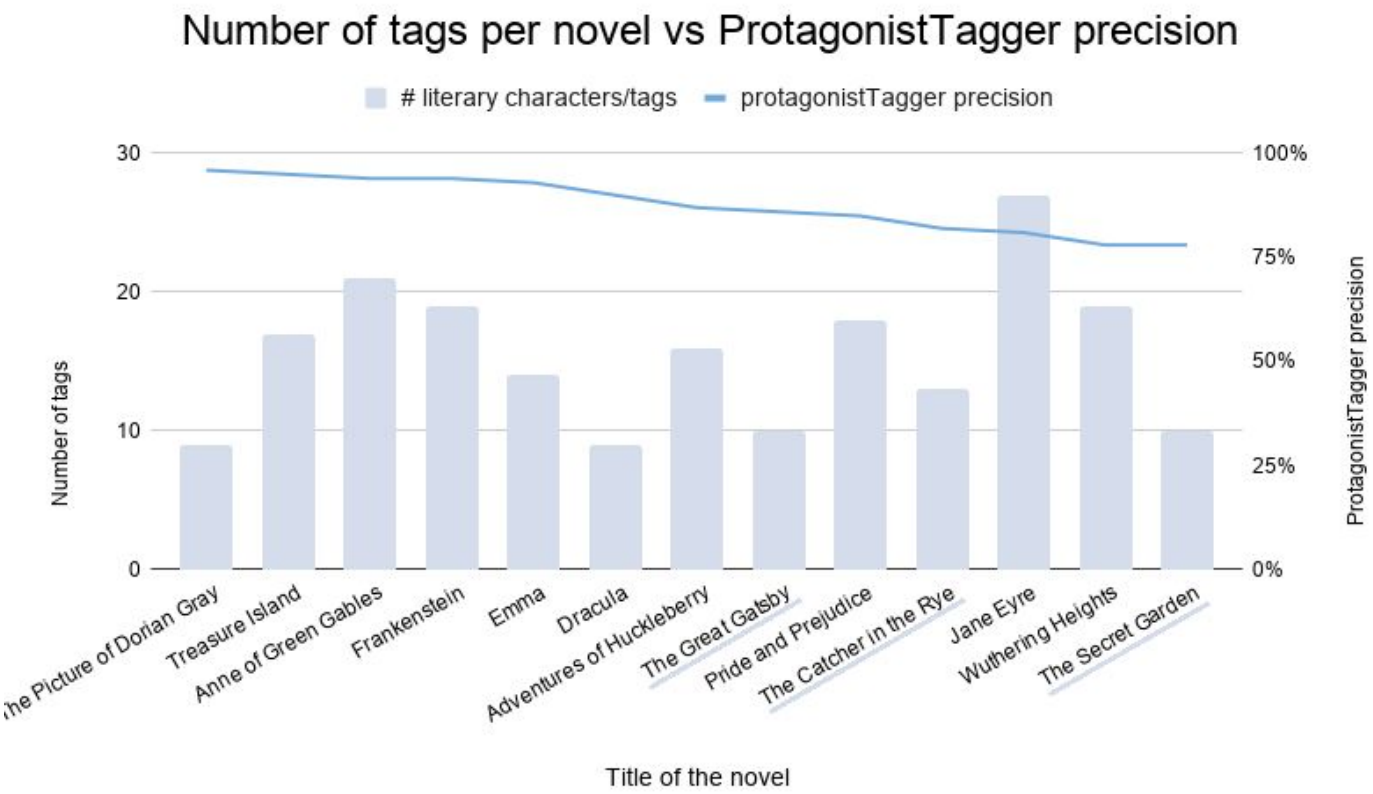
Novel title	precision	recall	F-measure
Pride and Prejudice	0.85	0.87	0.86
The Picture of Dorian Gray	0.96	0.97	0.96
Anne of Green Gables	0.94	0.96	0.95
Wuthering Heights	0.78	0.76	0.77
Jane Eyre	0.81	0.74	0.75
Frankenstein	0.94	0.92	0.93
Treasure Island	0.95	0.95	0.95
Adventures of Huckleberry Finn	0.87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



Tested NER models

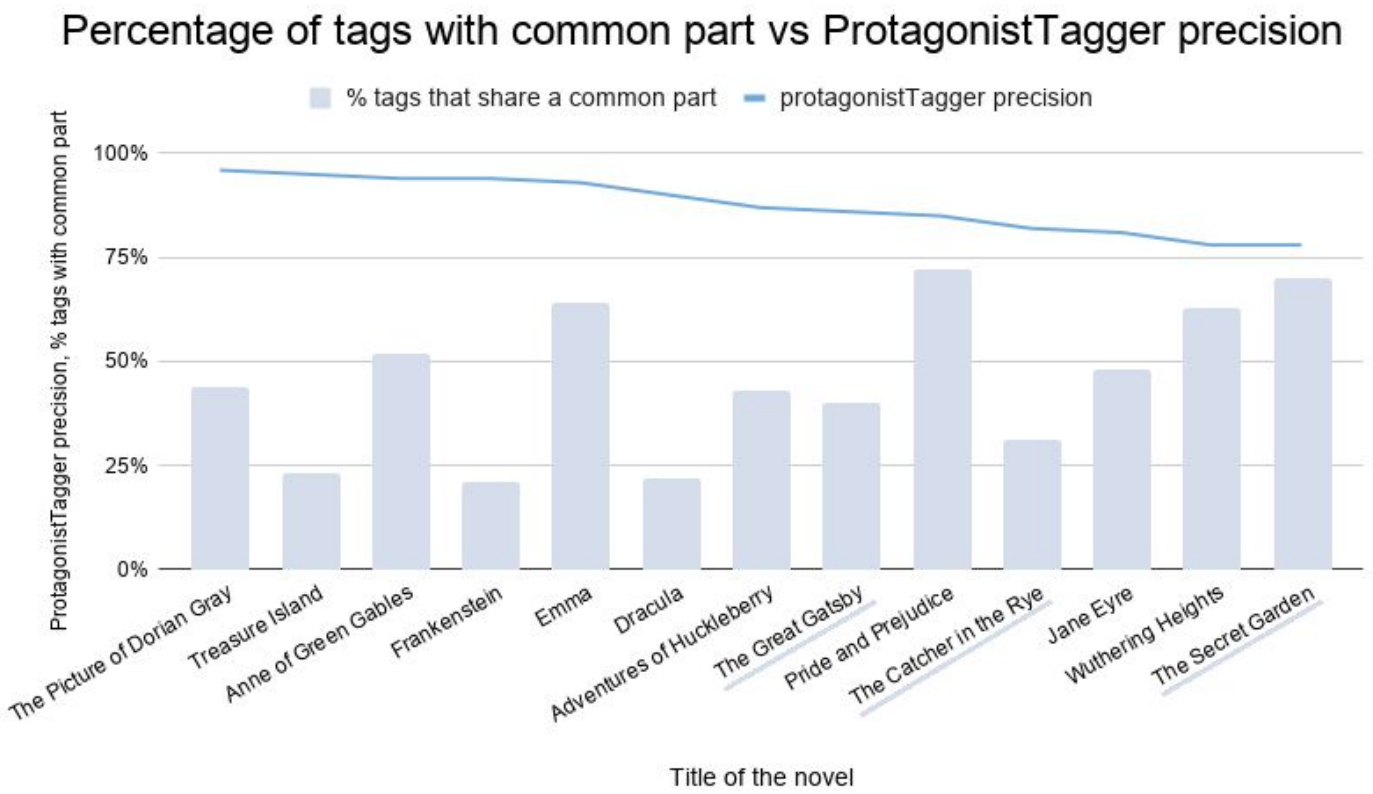
ProtagonistTagger performance vs number of tags



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



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%

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,addy
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

Analysis of *ProtagonistTagger* results

***

ProtagonistTagger performance vs NER performance