Comparing Annotated Datasets for Named Entity Recognition in English Literature

Rositsa V. Ivanova[1], Sabrina Kirrane[1], and Marieke van Erp[2]

[1] Vienna University of Economics and Business, [2] KNAW Humanities Cluster

Objective

Over the years, researchers and engineers have investigated various approaches that could potentially improve the recognition and linking of named entities. The growing interest in named entity recognition (NER) in various domains has led to the creation of different benchmark datasets. We take a closer look at existing annotated NER datasets in the domain of English literature and compare the performance of NER tools using such annotated datasets as a means to detect the differences between the datasets.

Datasets

	Litbank (Bamman et al., 2020)	OWTO (Dekker et al. 2019)	New CoNLL	New Extended
Novels	100	40	12	12
Annotators	95 by one ann. 5 by two ann.	two ann. 20 novels each	two ann. per novel	two ann. per novel
Innter-ann. F_1	86.00	-	95.25	78.17
Guidelines	ACE 2005 OntoNotes	-	CoNLL-2003	CoNLL-2003 LitBank
Ann. layers	multiple	one	one	one

Table 1:Characteristics of datasets

Tools

- BookNLP (Bamman et al., 2014) - targeting the domain of English novels
- Flair (Akbik et al., 2019; Schweter and Akbik, 2020) - one of the best performing NER tools

Results

Novel	LitBank			OWTO			New (CoNLL)			New (Ext)		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
Alice in Wonderland	80.00	54.05	64.52	92.00	74.19	82.14	100.00	80.65	**89.29**	100.00	12.89	22.83
David Copperfield	100.00	18.06	30.59	44.44	85.71	**58.54**	0.00	0.00	0.00	0.00	0.00	0.00
Dracula	45.45	10.87	17.54	14.29	33.33	20.00	45.45	35.71	**40.00**	45.45	2.16	4.12
Emma	86.67	38.24	53.06	83.10	98.33	**90.08**	40.00	36.73	38.30	37.78	6.59	11.22
Frankenstein	77.78	9.33	16.67	41.67	100.00	58.82	77.78	70.00	**73.68**	77.78	2.47	4.79
Huckleberry Finn	82.61	33.33	47.50	73.53	78.12	**75.76**	60.87	50.00	54.90	56.52	4.91	9.03
Moby Dick	71.43	6.58	12.05	37.50	100.00	54.55	71.43	62.50	**66.67**	42.86	1.42	2.75
Oliver Twist	73.33	11.96	20.56	70.00	100.00	82.35	93.33	87.50	**90.32**	86.67	6.81	12.62
Pride and Prejudice	95.74	42.06	58.44	73.08	98.28	**83.82**	31.91	31.25	31.58	29.79	4.58	7.93
The Call of the Wild	84.21	14.81	25.20	94.74	41.86	**58.06**	84.21	37.21	51.61	78.95	6.52	12.05
Ulysses	92.98	50.48	65.43	81.58	98.41	89.21	96.49	94.83	**95.65**	92.98	18.21	30.46
Vanity Fair	70.15	31.33	43.32	74.59	88.35	**80.89**	22.39	18.99	20.55	14.93	4.50	6.92
Mean	80.03	26.76	37.91	65.04	83.05	**69.52**	60.32	50.45	54.38	55.31	5.92	10.39
Standard deviation	14.46	16.89	19.75	24.8	23.1	20.34	32.32	29.02	29.9	32.14	5.11	8.65
Median	81.31	24.7	36.96	73.31	93.32	**78.33**	66.15	43.61	53.26	50.99	4.75	8.48

Table 2:Evaluation of BookNLP

Novel	LitBank			OWTO			New (CoNLL)			New (Ext)		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
Alice in Wonderland	76.92	54.05	63.49	92.31	82.76	82.27	100.00	83.87	**91.23**	100.00	13.40	23.64
David Copperfield	87.50	19.44	**31.82**	6.25	7.69	6.90	6.25	6.67	6.45	6.25	0.49	0.90
Dracula	57.14	8.70	15.09	0.00	0.00	0.00	57.14	28.57	**38.10**	57.14	1.72	3.35
Emma	60.00	26.47	36.73	57.78	68.42	62.65	71.11	65.31	**68.09**	68.89	12.02	20.46
Frankenstein	46.15	8.00	13.64	15.38	50.00	23.53	69.23	90.00	**78.26**	61.54	2.83	5.41
Huckleberry Finn	72.41	36.84	48.84	62.07	81.82	70.59	72.41	75.00	**73.68**	68.97	7.55	13.61
Moby Dick	88.89	10.53	18.82	33.33	100.00	50.00	88.89	100.00	**94.12**	66.67	2.84	5.45
Oliver Twist	6.67	10.87	18.69	66.67	90.91	76.92	86.67	81.25	**83.87**	80.00	6.28	11.65
Pride and Prejudice	29.41	14.02	18.99	21.57	31.43	25.58	94.12	100.00	**96.97**	88.24	14.71	25.21
The Call of the Wild	88.24	27.78	42.25	88.24	75.00	**81.08**	88.24	69.77	77.92	85.29	12.61	21.97
Ulysses	90.48	54.29	67.86	71.43	100.00	83.33	88.89	96.55	**92.56**	87.30	18.90	31.07
Vanity Fair	37.23	23.33	28.69	36.17	49.28	41.72	55.32	65.82	**60.12**	41.49	17.57	24.68
Mean	61.75	24.53	33.74	45.93	61.44	50.38	73.19	71.9	**71.78**	67.65	9.24	15.62
Standard deviation	27.33	16.42	18.6	31.46	34.1	30.37	25.47	28.56	26.5	25.09	6.44	10.16
Median	66.21	21.39	30.26	46.98	71.71	56.33	79.54	78.13	**78.09**	68.93	9.79	17.04

Table 3:Evaluation of Flair

Contact Information

- rivanova@wu.ac.at
- https://github.com/therosko/annotated_datasets_en_comparisson

Analysis

- The main difference between the performance comes down to honorifics.

	BookNLP	Flair	Litbank	OWTO	New CoNLL	New Extended
Mr.	I-PER	O	B-PER	I-PERSON	O	O
Bennet	I-PER	B-PER	I-PER	I-PERSON	B-PER	B-PER

- In some novels (e.g. Ulysses) personal pronouns make up to 10% of all tokens, leading to a drop in the recall whenever tools do not tag them.

	BookNLP	Flair	Litbank	OWTO	New CoNLL	New Extended
He	O	O	O	O	O	B-PER

	BookNLP	Flair	Litbank	OWTO	New CoNLL	New Extended
his	O	O	B-PER	O	O	B-PER
strong	O	O	I-PER	O	O	O
wellknit	O	O	I-PER	O	O	O
trunk	O	O	I-PER	O	O	O

- Inconsistencies in the gold standards (e.g. same entity tagged differently throughout the novel) lead to drops in the precision score of tools.
- Handling of common phrases (e.g. a boy) is done differently in the various gold standards, however neither of the tools tags them. This leads to further decrease in the recall values of the tools.

	BookNLP	Flair	Litbank	OWTO	New CoNLL	New Extended
their	O	O	B-PER	O	O	B-PER
friends	O	O	I-PER	O	O	I-PER

Discussion

Using existing gold standards:

+ datasets already exist
+ adaptation of existing annotations is possible
– literary texts differ from other types of texts (e.g. news)
– the annotation has been made following specific guidelines

Dataset maintenance and using old datasets:

+ evaluation is comparable over the years and easy to execute
– very limited view on the problem of NER
– tools are adapted to perform best with the datasets

Evaluation Metrics

+ evaluation is comparable over the years and easy to execute
– only accepts full matches as correct
– missing differentiation between “ambiguous” and “incorrect” tags
– small differences in the definition of an entity is amplified by the frequency of its occurrence

Annotation and Training Challenges

+ work could be invested into recognising shortcomings of existing datasets
– insufficient annotation guidelines for inexperienced annotators
– longer texts (needed for the literary domain) are difficult to annotate
– annotators may lack knowledge about the respective period of writing of the text

Selected References

[1] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

[2] D. Bamman, T. Underwood, and N. A. Smith. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, 2014.

[3] D. Bamman, S. Popat, and S. Shen. An annotated dataset of literary entities. 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)*, pages 2138–2144. Association for Computational Linguistics, June 2019. doi: 10.18653/v1/N19-1220.

[4] D. Bamman, O. Lewke, and A. Mansoor. An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54. European Language Resources Association, May 2020. ISBN 979-10-95546-34-4.

[5] N. Dekker, T. Kuhn, and M. van Erp. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science*, 5:e189, 2019.

[6] S. Schweter and A. Akbik. Flert: Document-level features for named entity recognition. *arXiv preprint arXiv:2011.06993*, 2020.