Supplementary Material

Effective Crowd-Annotation of Participants, Interventions, and Outcomes in the Text of Clinical Trial Reports

A Annotation Interfaces

We present the annotation interface for SENBASE in Figure 1 and for SENSUPPORT in Figure 2. Notice that in each interface the annotator has the optional choice to read the full abstract text in that the annotated sentence appears. In this abstract, the currently annotated sentence is highlighted by a blue border, as shown in Figure 3.

We make the interface available for download at `https://github.com/Markus-Zlabinger/pico-annotation`.

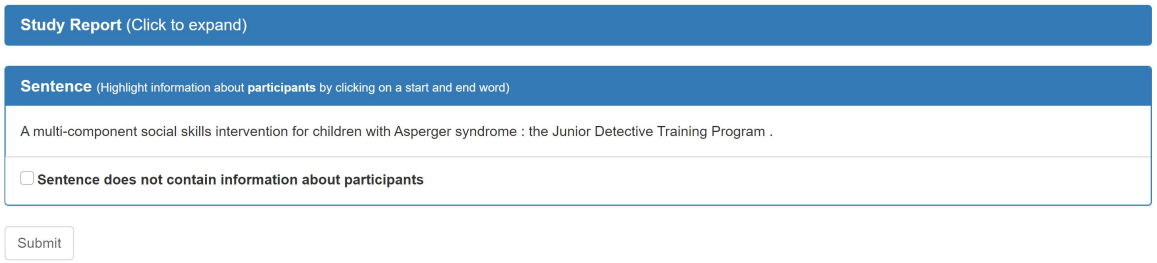

Figure 1: Annotation interface of SENBASE

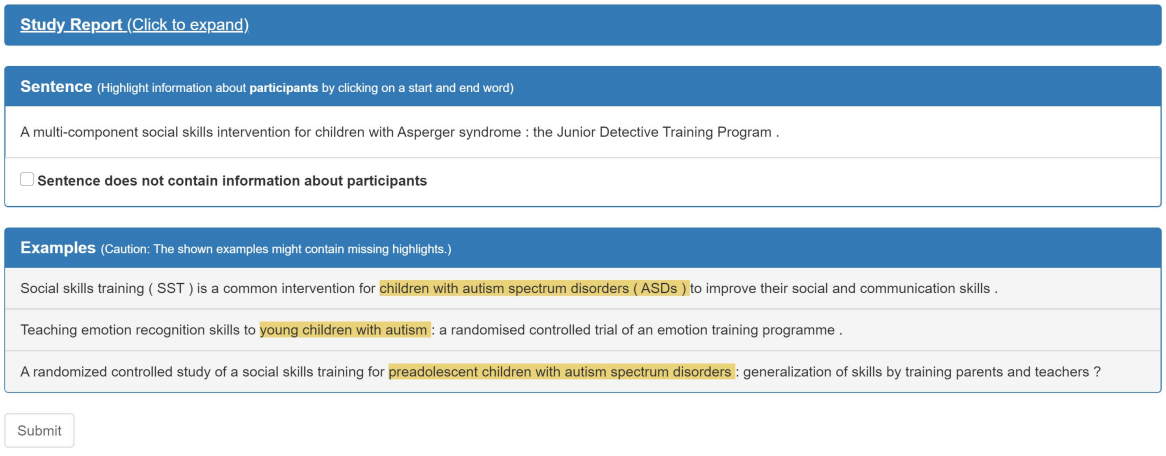

Figure 2: Annotation interface of SENSUPPORT

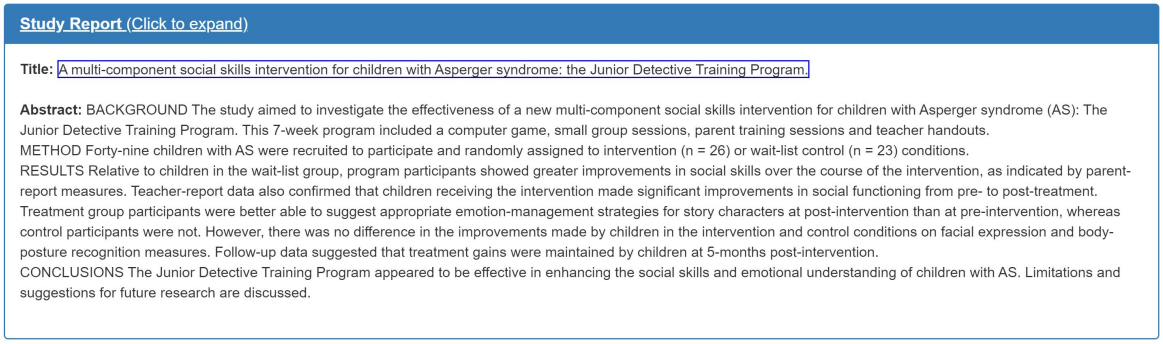

Figure 3: Optionally, annotators can in both interfaces examine the full abstract text.

B Pretrained Models for Similarity Methods

An overview of all pretrained models is given in Table 1. The evaluation results of the unsupervised semantic similarity methods for all pretrained models is presented in Table 2.

Embedding	Model	Training Data	Used by Method
Word	BioWord2Vec [12]	PubMed abstracts, MIMIC III corpus [8]	AVG, WAVG, SIF
	FastText [10]	English Wikipedia	AVG, WAVG, SIF
	PubMedW2VSmall (window 2) [6]	PubMed abstracts	AVG, WAVG, SIF
	PubMedW2VLarge (window 30) [6]	PubMed abstracts	AVG, WAVG, SIF
Text	SciBERT [2]	Semanticscholar full-text papers	SenBERT
	BioBERT [9]	PubMed abstracts	SenBERT
	ClinicalBERT [1]	MIMIC III corpus [8]	SenBERT
	BioSent2Vec [5]	PubMed abstracts, MIMIC III corpus [8]	Sent2Vec
	USE 4.0 [4]	Wikipedia, web news, online forums, SNLI corpus [3]	USE
	InferSent 2.0 [7]	SNLI corpus [3]	InferSent
	WikiUnigram [11]	English Wikipedia	Sent2Vec

Table 1: Overview of the pretrained models.

C Computation Infrastructure

We ran the experiments for the similarity methods and the PIO task designs on a single machine with following specs:

- CPU: Intel Core i7-8750H
- GPU: GeForce GTX 1060
- RAM: 16 GB DDR4

Method	Model	Preprocessing	MedSTS	BIOSSES	Avg.
TFIDF	-	Lower	0.74	0.70	0.72
TFIDF	-	LowerStop	0.74	0.73	0.74
Levensthein	-	Lower	0.55	0.64	0.60
Levensthein	-	LowerStop	0.64	0.69	0.66
AVG	BioWord2Vec	Lower	0.61	0.72	0.66
AVG	BioWord2Vec	LowerStop	0.72	0.77	0.75
AVG	PubMedW2VSmall	Lower	0.45	0.65	0.55
AVG	PubMedW2VSmall	LowerStop	0.64	0.76	0.70
AVG	PubMedW2VLarge	Lower	0.48	0.63	0.55
AVG	PubMedW2VLarge	LowerStop	0.66	0.76	0.71
AVG	FastText	Lower	0.19	0.45	0.32
AVG	FastText	LowerStop	0.57	0.71	0.64
WAVG	BioWord2Vec	Lower	0.73	0.75	0.74
WAVG	BioWord2Vec	LowerStop	0.76	0.77	0.76
WAVG	PubMedW2VSmall	Lower	0.64	0.74	0.69
WAVG	PubMedW2VSmall	LowerStop	0.68	0.76	0.72
WAVG	PubMedW2VLarge	Lower	0.67	0.74	0.70
WAVG	PubMedW2VLarge	LowerStop	0.71	0.77	0.74
WAVG	FastText	Lower	0.50	0.65	0.57
WAVG	FastText	LowerStop	0.62	0.72	0.67
SIF	BioWord2Vec	Lower	0.79	0.75	0.77
SIF	BioWord2Vec	LowerStop	0.78	0.76	0.77
SIF	PubMedW2VSmall	Lower	0.71	0.75	0.73
SIF	PubMedW2VSmall	LowerStop	0.70	0.76	0.73
SIF	PubMedW2VLarge	Lower	0.72	0.75	0.74
SIF	PubMedW2VLarge	LowerStop	0.71	0.76	0.73
SIF	FastText	Lower	0.59	0.69	0.64
SIF	FastText	LowerStop	0.65	0.73	0.69
SenBERT	ClinicalBERT	Identity	0.65	0.69	0.67
SenBERT	BioBERT	Identity	0.78	0.58	0.68
SenBERT	SciBERT	Identity	0.60	0.68	0.64
USE	USE 4.0	Identity	0.66	0.72	0.69
InferSent	InferSent 2.0	Identity	0.49	0.65	0.57
Sent2Vec	BioSent2Vec	Lower	0.81	0.74	0.78
Sent2Vec	BioSent2Vec	LowerStop	0.81	0.77	0.79
Sent2Vec	WikiUnigram	Lower	0.65	0.74	0.70
Sent2Vec	WikiUnigram	LowerStop	0.64	0.77	0.70

Table 2: Pearson correlation between the ground truth labels and the unsupervised semantic similarity methods. This table includes the results of *all* evaluated pretrained models.

References

[1] ALSENTZER, E., MURPHY, J. R., BOAG, W., WENG, W.-H., JIN, D., NAUMANN, T., AND MCDERMOTT, M. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (2019), Association for Computational Linguistics, pp. 72–78.

[2] BELTAGY, I., LO, K., AND COHAN, A. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019), pp. 3606–3611.

[3] BOWMAN, S. R., ANGELI, G., POTTS, C., AND MANNING, C. D. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).

[4] CER, D., YANG, Y., KONG, S.-Y., HUA, N., LIMTIACO, N., JOHN, R. S., CONSTANT, N., GUAJARDO-CESPEDES, M., YUAN, S., TAR, C., SUNG, Y.-H., STROPE, B., AND KURZWEIL, R. Universal Sentence Encoder. *arXiv:1803.11175 [cs]* (Apr. 2018).

[5] CHEN, Q., PENG, Y., AND LU, Z. BioSentVec: Creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)* (2019), IEEE, pp. 1–5.

[6] CHIU, B., CRICHTON, G., KORHONEN, A., AND PYYSALO, S. How to Train good Word Embeddings for Biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing* (Berlin, Germany, Aug. 2016), Association for Computational Linguistics, pp. 166–174.

[7] CONNEAU, A., KIELA, D., SCHWENK, H., BARRAULT, L., AND BORDES, A. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen, Denmark, Sept. 2017), Association for Computational Linguistics, pp. 670–680.

[8] JOHNSON, A. E., POLLARD, T. J., SHEN, L., LI-WEI, H. L., FENG, M., GHASSEMI, M., MOODY, B., SZOLOVITS, P., CELI, L. A., AND MARK, R. G. MIMIC-III, a freely accessible critical care database. *Scientific data 3* (2016), 160035.

[9] LEE, J., YOON, W., KIM, S., KIM, D., KIM, S., SO, C. H., AND KANG, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (Sept. 2019), btz682.

[10] MIKOLOV, T., GRAVE, E., BOJANOWSKI, P., PUHRSCH, C., AND JOULIN, A. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)* (2018).

[11] PAGLIARDINI, M., GUPTA, P., AND JAGGI, M. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics* (2018).

[12] ZHANG, Y., CHEN, Q., YANG, Z., LIN, H., AND LU, Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data 6*, 1 (May 2019), 1–9.