Cross-institution extraction of social determinants of health from clinical notes: an evaluation of methods

Madhumita Sushil[1], PhD, Atul J. Butte[1], MD, PhD, Ewoud Schuit[2], PhD, Artuur M. Leeuwenberg[2], PhD
[1] Bakar Computational Health Sciences Institute, University of California, San Francisco, USA
[2] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, The Netherlands

Introduction

Social determinants of health (SDOH), like employment status, living condition, alcohol and tobacco use can provide important information about a patient's health, and are in part stored in free text clinical notes in the electronic health record (EHR).[1] This study compared several natural language processing (NLP) techniques to extract SDOH from English clinical notes, as part of the National NLP Clinical Challenge on extraction of SDOH.[2]

Methods

Data used for training and evaluation consisted of annotated clinical notes from MIMIC-III,[3] and the University of Washington (UW) and 75 million unannotated notes from the University of California, San Francisco (UCSF). The task was set up in three phases (detailed in Table 1): an in-domain setting (phase A: train and test data from the same institution), a cross-domain setting (phase B: train and test data come from different institutions), and a domain-transfer setting (phase C: a small additional dataset from the target institution is available for training). Our basic system architecture first extracts event triggers with a sequence tagging model, and then for each detected event extracts typed arguments, again using a sequence tagging model per argument type. For sequence tagging, a BIO (Begin, In, Out) scheme is used, together with Long Short Term Memory network (LSTM) models. This basic architecture was extended with simple techniques known in the literature to improve the generalizability: fixed pre-trained fastText-based[4] embeddings (concatenation of embeddings trained on MIMIC and UCSF notes) and fine-tuned contextual BERT[5] embeddings (trained on UCSF notes), self-training (in 186 unlabeled MIMIC or 250 UCSF notes), loss reweighting of the target domain UW data, and final layer fine tuning on the target domain UW data. Per phase, only three submissions were permitted, so we submitted only a subset of all techniques per phase. After release of the test data, we performed additional experiments to better assess the impact of each technique.

Results[1]

Regarding our shared task submissions (see Table 1): In task A, the in domain evaluation, all our submissions (indicated by *) obtained above-average F1 scores. In task B, the cross-domain evaluation, interestingly, the mean F1 was higher than for task A, and our submitted systems performed below average. In task C, the domain transfer setting, two of our submissions perform above average, and one below (due to a later identified bug in the code). Regarding the impact of the explored techniques: across phases, pre-trained fastText embeddings increased recall and F1 score (comparing rows 1 vs. 3, 8 vs. 9, and 14 vs. 15). BERT embeddings further improve recall, and F1 (comparing rows 1 vs. 5, 2 vs. 6, 8 vs. 11, 14 vs. 19). Self-training, explored for phase A and B, gave mixed results (comparing rows 1 vs. 2, 3 vs. 4, 5 vs. 6, 9 vs. 10, 11 vs. 12). In phase C, increasing the weight of target domain data in the loss function only sometimes improved the performance (comparing rows 15 vs. 18, and 19 vs. 21). Final layer fine tuning in the target domain slightly increased F1 (comparing rows 14 vs. 16, 15 vs. 17, and 19 vs. 20).

Discussion

Overall BERT-based models performed best in all evaluation settings. Pretrained embeddings on large unlabeled notes are critical to both cross-domain generalization as well as improved domain transfer. Optimizing only the final layer of the model further helps domain transfer, but is relatively less impactful. We conjecture that further analysis of training (fixing pretrained embeddings; larger models; larger corpora), decoupling data-specific features during model optimization, as well as exploring alternative training schemes (instead of the BIO scheme) may further improve cross-domain generalization. However, extensive research is needed to make concrete conclusions.

Acknowledgements

We thank the organizers of the 2022 N2C2 Shared Task on Challenges in NLP for Clinical Data for their provision of the data and independent system evaluation. We further thank the Information Commons team at UCSF for curating the deidentified clinical data warehouse (IRB #18-25163), which can be used for research without further IRB.

[1] Code corresponding to this work can be found at: https://github.com/tuur/sdoh_n2c2track2_ucsf_umcu

Table 1. Results per phase and setting. P=Precision, R=Recall, F1=F1 score, ST=Self training, fT=pre-trained fastText embeddings, FLF=Final layer fine tuning, $LW_{\times 10}$=10 times higher weight for the target domain loss. Official submissions are indicated with an upper star (*). Bold is best performance, underline is best submitted performance.

	Setting	Training	Testing	Model	P	R	F1
1	Phase A (same domain)	$MIMIC_{train+dev}$ (1504 docs.)	$MIMIC_{test}$ (373 docs.)	*LSTM	0.84	0.65	<u>0.74</u>
2				LSTM + ST_{MIMIC}	0.87	0.64	0.73
3†				*LSTM + $fT_{MIMIC@UCSF}$	~~0.88~~0.83	~~0.56~~0.69	~~0.68~~0.76
4†				*LSTM + $fT_{MIMIC@UCSF}$ + ST_{MIMIC}	~~0.85~~0.85	~~0.60~~0.71	~~0.70~~0.77
5				$BERT_{UCSF}$ + LSTM	0.81	0.76	**0.78**
6				$BERT_{UCSF}$ + LSTM + ST_{MIMIC}	0.78	0.75	0.76
7				Shared task mean/median (14 teams)	0.69	0.56	0.60/0.69
8	Phase B (cross domain)	$MIMIC_{train+dev}$ (1504 docs.)	$UW_{train+dev}$ (2010 docs.)	LSTM	0.72	0.42	0.53
9†				*LSTM + $fT_{MIMIC@UCSF}$	~~0.72~~0.72	~~0.26~~0.49	~~0.39~~0.58
10†				*LSTM + $fT_{MIMIC@UCSF}$ + ST_{UCSF}	~~0.70~~0.76	~~0.30~~0.46	~~0.42~~0.57
11				*$BERT_{UCSF}$ + LSTM	0.68	0.54	**<u>0.61</u>**
12				$BERT_{UCSF}$ + LSTM + ST_{UCSF}	0.65	0.56	0.60
13				Shared task mean/median (10 teams)	0.68	0.58	0.62/0.70
14	Phase C (domain transfer)	$MIMIC_{train+dev}$ (1504 docs.) + $UW_{train+dev}$ (2010 docs.)	UW_{test} (518 docs.)	LSTM	0.86	0.66	0.75
15†				*LSTM + $fT_{MIMIC@UCSF}$	~~0.90~~0.85	~~0.55~~0.70	~~<u>0.68</u>~~0.77
16				LSTM + FLF_{UW}	0.85	0.68	0.76
17†				*LSTM + $fT_{MIMIC@UCSF}$ + FLF_{UW}	~~0.81~~0.88	~~0.42~~0.68	~~0.55~~0.77
18†				*LSTM + $fT_{MIMIC@UCSF}$ + $LW_{\times 10}$	~~0.89~~0.87	~~0.55~~0.71	~~0.68~~0.78
19				$BERT_{UCSF}$ + LSTM	0.83	0.75	0.78
20				$BERT_{UCSF}$ + LSTM + FLF_{UW}	0.84	0.74	**0.79**
21				$BERT_{UCSF}$ + LSTM + $LW_{\times 10}$	0.82	0.75	0.78
22				Shared task mean/median (10 teams)	0.70	0.61	0.64/0.73

† A code bug affecting loading of pre-trained embeddings was found in these models after submission. Corrected results for these settings are presented on the same line behind the original crossed out results.

References

1. Lybarger, K., Ostendorf, M. & Yetisgen, M. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *J. Biomed. Inform.* **113**, 103631 (2021).
2. National NLP Clinical Challenges - Track 2 - Extracting Social Determinants of Health. https://n2c2.dbmi.hms.harvard.edu/2022-track-2.
3. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
4. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **5**, 135–146 (2017).
5. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in *Proceedings of NAACL, Volume 1,* 4171–4186 (ACL, 2019).