Overview of BioCreative/OHNLP Challenge 2018 Task 2: Clinical Semantic Textual Similarity

Yanshan Wang, Naveed Afzal, Sijia Liu, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Sunyang Fu, Hongfang Liu
Department of Health Sciences Research, Mayo Clinic, Rochester, MN

***Abstract*—Clinical Semantic Textual Similarity (ClinicalSTS) measures the degree of semantic equivalence between two snippets of clinical text. It removes redundancy in clinical texts to facilitate the secondary use of electronic health records (EHRs). Applications of ClinicalSTS include clinical text summarization, clinical question answering, clinical semantic information retrieval, and clinical decision support. While prior STS shared tasks constrained themselves to texts in the general domain, the first ClinicalSTS shared task focuses on clinical texts generated from clinical notes in the real world, aiming to motivate the natural language processing and biomedical informatics communities to study STS in the clinical domain by curating clinical text dataset and assessing the state-of-the-art computational methods. The ClinicalSTS shared task attracted 4 participating teams producing a total of 12 system submissions.**

Keywords—clinical semantic textual similarity; natural language processing; electronic health records

I. Introduction

The wide adoption of electronic health records (EHRs) has led to an improvement in healthcare quality by electronically documenting a patient's medical conditions, thoughts and actions among the care providers [1]. Those EHR data, with the vast majority being free-texts (e.g., clinical notes, discharge summaries, radiology reports, and pathology reports), have been utilized for primary and secondary purposes, such as documentation need in care process, clinical decision support, outcome improvement, biomedical research and epidemiologic monitoring of the nation's health [2]. Due to the ease of use of EHRs, the frequency use of copy-and-paste, templates, and smart phrases have resulted in redundant texts in clinical notes, which reduces the quality of EHR data and adds cognitive burden of tracking complex medical records in clinical practice [3]. An analysis of 23, 630 progress notes written by 460 clinicians shows that 18% of the text was manually entered; 46%, copied; and 36%, imported [4]. Therefore, there is a growing need for tools that can aggregate data from diverse sources and minimize data redundancy, and organize and present the EHR data in a user-friendly way to reduce physicians' cognitive burden.

One technique for automatically reducing redundancy in free text EHRs is to compute semantic similarity between clinical text snippets and remove highly similar snippets. Semantic textual similarity (STS) is a common task in the general English domain to assess the degree to which the underlying semantics of two segments of text are equivalent to each other. The assessment is usually performed using ordinal scaled output ranging from complete semantic equivalence to complete semantic dissimilarity. STS shared task has been held annually since 2012 to encourage and support research in this area [5-10]. However, these series of STS tasks used texts in the general English domain and no STS shared task focuses on the text data in the clinical domain.

To motivate the natural language processing (NLP) and biomedical informatics communities to study STS in the clinical domain, we initiated this ClinicalSTS shared task at the 2018 BioCreative/OHNLP Challenge to provide a venue for evaluation of the state-of-the-art algorithms and models.

II. Task Overivew

ClinicalSTS provides paired clinical text snippets for each participant. The clinical text snippets are mostly sentences from clinical notes. The participating systems are asked to return a numerical score indicating the degree of semantic similarity between the pair of two sentences. Performance is measured by the Pearson correlation coefficient between the predicted similarity scores and human judgments. The ClinicalSTS scores fall on an ordinal scale, ranging from 0 to 5 where 0 means that the two clinical text snippets are completely dissimilar (i.e., no overlap in their meanings) and 5 means that the two snippets have complete semantic equivalence. Table 1 illustrates clinical text examples of the ordinal similarity scale. Participating systems can use real valued scores to indicate their semantic similarity prediction.

III. Data Preparation

The data for the ClinicalSTS shared task is an annotated subset of the MedSTS dataset [11] collected from EHRs at the Mayo Clinic's clinical data warehouse. Here we briefly describe how MedSTS was curated. From the data warehouse, we selected unique sentences from 3 million de-identified clinical notes of patients receiving their primary care at Mayo Clinic. Then we removed protected health information (PHI) by employing a frequency filtering approach [12] based on the assumption that sentences appearing in multiple patients' records tend to contain no PHI information, which resulted in 14.9 million unique sentences with 361.9 million tokens. We used the averaged value ($\geq$ 0.45) of three surface lexical similarities, namely Ratcliff/Obershelp pattern matching algorithm [13], cosine similarity [14], and Levenshtein distance [15], as a cutoff value to obtain candidate sentence pairs with

TABLE I. SIMILARITY SCORES WITH EXPLANATIONS AND EXAMPLES FOR THE CLINICALSTS TASK FROM [11].

Score	Examples
5	*The two sentences are completely equivalent, as they mean the same thing.* S1 → Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA Aerosol 2 puffs by inhalation every 4 hours as needed. S2 → Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA Aerosol 1-2 puffs by inhalation every 4 hours as needed #1 each.
4	*The two sentences are mostly equivalent, but some unimportant details differ.* S1 → Discussed goals, risks, alternatives, advanced directives, and the necessity of other members of the surgical team participating in the procedure with the patient. S2 → Discussed risks, goals, alternatives, advance directives, and the necessity of other members of the healthcare team participating in the procedure with the patient and his mother.
3	*The two sentences are roughly equivalent, but some important information differs/missing.* S1 → Cardiovascular assessment findings include heart rate normal, Heart rhythm, atrial fibrillation with controlled ventricular response. S2 → Cardiovascular assessment findings include heart rate, bradycardic, Heart rhythm, first degree AV Block.
2	*The two sentences are not equivalent, but share some details.* S1 → Discussed risks, goals, alternatives, advance directives, and the necessity of other members of the healthcare team participating in the procedure with (patient) (legal representative and others present during the discussion). S2 → We discussed the low likelihood that a blood transfusion would be required during the postoperative period and the necessity of other members of the surgical team participating in the procedure.
1	*The two sentences are not equivalent, but are on the same topic.* S1 → No: typical 'cold' symptoms; fever present (greater than or equal to 100.4 F or 38 C) or suspected fever; rash; white patches on lips, tongue or mouth (other than throat); blisters in the mouth; swollen or 'bull' neck; hoarseness or lost voice or ear pain. S2 → New wheezing or chest tightness, runny or blocked nose, or discharge down the back of the throat, hoarseness or lost voice.
0	*The two sentences are completely dissimilar.* S1 → The risks and benefits of the procedure were discussed, and the patient consented to this procedure. S2 → The content of this note has been reproduced, signed by an authorized physician in the space above, and mailed to the patient's parents, the patient's home care company.

some level of prima facie similarity. Please refer to [11] for more details of how these methods were employed.

The MedSTS dataset consists of a total of 174,629 sentence pairs. A randomly selected dataset of 1,068 sentence pairs from the MedSTS dataset was finally annotated by human experts and used for the ClinicalSTS shared task. In order to ensure no PHI exists in this dataset, we manually removed PHI from each sentence.

In the annotation phase, two clinical experts were asked to independently annotate each sentence pair in the ClinicalSTS dataset on the basis of their semantic equivalence. Both annotators were vastly experienced with many years of experience of clinical domain. The agreement between the two annotators was high, with a weighted Cohen's Kappa of 0.67. We utilized the average of their scores as the gold standard for evaluating submitted systems. 70% of the ClinicalSTS dataset (750 sentence pairs with gold standard) was released as training data to each team to develop and tune their systems. The participating systems are asked to return a numerical score indicating the degree of semantic similarity for the remaining 30% testing data (318 sentence pairs).

IV. SYSTEM EVALUATION

This section reports participant evaluation results for the ClinicalSTS shared task.

A. Participation

Participating teams were required to sign a Mayo Data Use Agreement to get access to the dataset. Each team can submit up to 3 runs for the testing data where each run should have one line for each sentence pair that provides the similarity score assigned by the system as a floating point number. We note that the

ncbi_sennlp team was granted permission to submit 4 runs as the team has 3 runs using machine learning models and they asked to submit an additional run using merely deep learning models. The task attracted 4 participating teams with a total of 12 system submissions.

B. Evaluation Metric

Participant systems were evaluated based on the Pearson correlation coefficient between the predicted scores and the gold standard on the testing set.

C. Rankings

The rankings for the ClinicalSTS task are given in Table 2. The best performance is obtained by ncbi_sennlp's Regression-Run, which achieves a Pearson correlation of 0.8328. Overall the median correlation score for the testing dataset is 0.8016.

TABLE II. CLINICALSTS RANKINGS.

Team	Run	Pearson correlation	Run Rank	Team Rank
ncbi_sennlp	Regression-Run	0.8328	1	1
ncbi_sennlp	DLML-Run	0.8258	2	
ncbi_sennlp	RFDNN-Run	0.8246	3	
HITSZ	Ensemble	0.8143	4	2
HITSZ	CNN-LSTM	0.8121	5	
ncbi_sennlp	RF-Run	0.8106	6	
HITSZ	CNN	0.7925	7	
PennTeam	Conceptor-Run_PSL	0.7789	8	3
PennTeam	Conceptor-Run_PSL_combined	0.7779	9	
NTTU	Vote-Run1	0.7090	10	4
NTTU	Libsvm(epsllon-SVR)-Run3	0.7046	11	
NTTU	Libsvm(nu-SVR)-Run2	0.7005	12	

D. Methods

Participating teams explored techniques ranging from conventional machine learning models to the state-of-the-art deep learning models.

The winner, ncbi_sennlp, submitted four systems. The first system is the Random Forest model using 63 features including string similarity features, entity similarity features, number similarity features, and deep learning features. The second system uses the average score of the first system and Dense Neural Networks. The third system applies a regression model on 8 trained models including the Random Forest model, the Bayesian Ridge regression model, the Lasso regression model, the linear regression model, the Extra Tree model, the DNN using the Universal Sentence Encoder, the DNN using the inferSent encoder, and the Encoder-MLP using the inferSent encoder. The fourth system uses the average score of the first system and the Encoder-MLP using the inferSent encoder. The third system achieved the best performance among all submitted systems with a Pearson correlation of 0.8328.

Second place was achieved by HITSZ using Attention-Based Convolutional Neural Network (ABCNN) and Bi-direction Long Short Term Memory networks (Bi-LSTM). One submission uses ABCNN with traditional NLP features. Another is a hybrid model of ABCNN and Bi-LSTM, with the traditional NLP features. The third run ensembles previous two systems by calculating the average scores. The ensemble model performs best among their submitted systems.

The third place team, PennTeam, proposed a sentence embedding method that represents a sentence as a weighted average of word vectors followed by a soft projection. They used a self-regularized identity map named Conceptors to correct the common component bias in linear sentence embedding.

The forth place team, NTTU, utilized majority voting and two different support vector regression models with only word embedding representation features for their submissions. The best performance was achieved by the majority voting method.

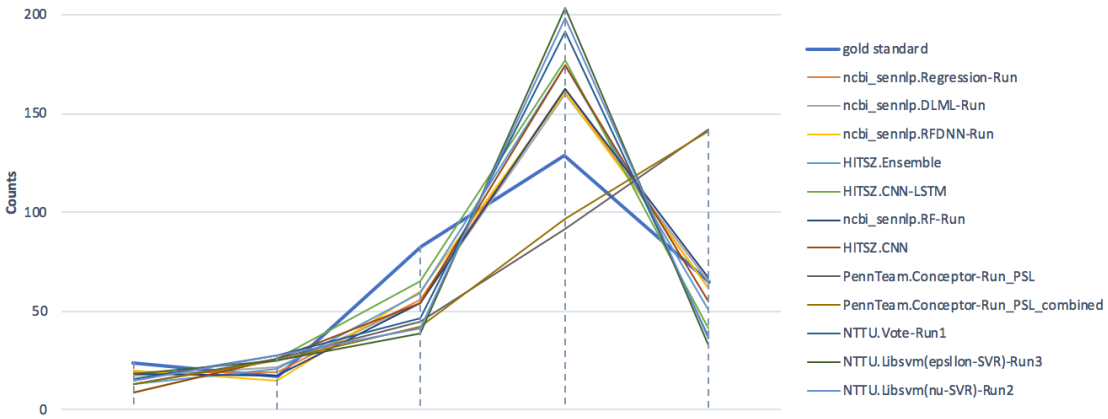

Fig. 1. Counts of the similarity scores assigned by the participating systems.

V. Analysis

Fig. 1 plots the counts of the similarity scores assigned by the participating systems and human experts. Most systems returned similarity scores in the ranges of (2,3] and (3,4] as the gold standard except two systems from the PennTeam. However, most systems predicted more sentences in the score range of (3,4], and less in the range of (2,3] than gold standard. The predicted scores are well aligned with the gold standard for extreme ranges, such as [0,1] and [4,5]. It indicates that machines are good at distinguishing complete similar or complete dissimilar sentence pairs but it is difficult for both human and machines to distinguish sentences with somewhat semantic similarity.

VI. Conclusion

We have presented the results of the ClinicalSTS shared task at the 2018 BioCreative/OHNLP Challenge. The shared task focuses on computing semantic similarity for clinical text sentences generated from clinical notes in the real world. There are 12 submissions from 4 teams. The best system has a high Pearson correlation of 0.8328. The ClinicalSTS shared task is the first attempt to encourage the NLP and biomedical informatics communities to tackle STS in clinical domain. Since we need to manually check and remove PHI from the data, we only released a subset of the collection in this shared task. In the future, we plan to release larger training data to facilitate more sophisticated deep learning models.

Acknowledgment

We would like to thank Donna Ihrke and Liwei Wang for annotating the corpus. This work was made possible by NIGMS R01GM102282, NLM R01LM11934, NIBIB R01EB19403, and NCATS U01TR02062. The student travel grant and awards are sponsored by NIGMS 5R01GM080646-12.

References

1. Blumenthal, D., *Implementation of the federal health information technology initiative*. New England Journal of Medicine, 2011. **365**(25): p. 2426-2431.
2. Wang, Y., et al., *Clinical information extraction applications: a literature review*. Journal of biomedical informatics, 2017.
3. Zhang, R., et al. *Evaluating measures of redundancy in clinical texts*. in *AMIA Annual Symposium Proceedings*. 2011. American Medical Informatics Association.
4. Wang, M.D., R. Khanna, and N. Najafi, *Characterizing the source of text in electronic health record progress notes*. JAMA internal medicine, 2017. **177**(8): p. 1212-1213.
5. Agirre, E., et al. * *SEM 2013 shared task: Semantic textual similarity*. in *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. 2013.
6. Agirre, E., et al. *Semeval-2014 task 10: Multilingual semantic textual similarity*. in *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*. 2014.
7. Agirre, E., et al. *Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability*. in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*. 2015.
8. Agirre, E., et al. *Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation*. in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 2016.
9. Cer, D., et al., *SemEval-2017 Task 1: Semantic Textual Similarity-Multilingual and Cross-lingual Focused Evaluation*. arXiv preprint arXiv:1708.00055, 2017.
10. Afzal, N., Y. Wang, and H. Liu. *MayoNLP at SemEval-2016 Task 1: Semantic textual similarity based on lexical semantic net and deep learning semantic model*. in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 2016.
11. Wang, Y., et al., *MedSTS: A Resource for Clinical Semantic Textual Similarity*. 2018.
12. Li, D., et al. *A frequency-filtering strategy of obtaining PHI-free sentences from clinical data repository*. in *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*. 2015. ACM.
13. Black, P.E., *Ratcliff/Obershelp pattern recognition*. Dictionary of Algorithms and Data Structures, 2004. **17**.
14. Sohn, S., et al., *Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions*. Journal of the American Medical Informatics Association, 2017. **25**(3): p. 353-359.
15. Soukoreff, R.W. and I.S. MacKenzie. *Measuring errors in text entry tasks: an application of the Levenshtein string distance statistic*. in *CHI'01 extended abstracts on Human factors in computing systems*. 2001. ACM.