Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220606

Clustering Nursing Sentences - Comparing Three Sentence Embedding Methods

Hans MOEN[a,1], Henry SUHONEN[b], Sanna SALANTERÄ[b,c], Tapio SALAKOSKI[d], and Laura-Maria PELTONEN[b]
[a]*Department of Computer Science, Aalto University, Espoo, Finland*
[b]*Department of Nursing Science, University of Turku, Turku, Finland*
[c]*Turku University Hospital, Turku, Finland*
[d]*Department of Mathematics and Statistics, University of Turku, Turku, Finland*

Abstract. In health sciences, high-quality text embeddings may augment qualitative data analysis of large amounts of text by enabling, e.g., searching and clustering of health information. This study aimed to evaluate three different sentence-level embedding methods in clustering sentences in nursing narratives from individual patients' hospital care episodes. Two of these embeddings are generated from language models based on the BERT framework, and the third on the Sent2Vec method. These embedding methods were used to cluster sentences from 20 patient care episodes and the results were manually evaluated. Findings suggest that the best clusters were produced by the embeddings from a BERT model fine-tuned for the proxy task of predicting subject headings for nursing text.

Keywords. Text clustering, natural language processing, sentence embeddings, nursing documentation, electronic health records

1. Introduction

Vectorized representations (embeddings) of text that captures meaning in a semantic space are important for many tasks related to natural language processing (NLP). This includes searching, clustering, summarization, and classification. The interest in textual embeddings has rapidly been growing since the introduction of the Word2Vec method [1]. The current direction focuses on contextualized embeddings with pre-trained language models like ELMo [2], and more recently transformer-based models like BERT [3]. Studies show that BERT-based language models without fine-tuning for a downstream task perform quite poorly, often underperforming compared to averaging the word embeddings from traditional (global) word embedding methods [4]. Thus, fine-tuning of BERT models on relevant tasks and domains seems important for generating embeddings that capture the semantics of the targeted domain.

In health sciences, text clustering may augment qualitative data analysis of large amounts of health-related text to support both clinical work as well as research. Our aim is to group sentences from nursing notes from individual patients' hospital stays (care episodes) into clusters that each focus on one and the same topic, or possibly a coherent set of topics for sentences that cover more than one topic.

[1]Corresponding author, Hans Moen, Tietotekniikan laitos, P.O. Box. 15400, FI-00076 AALTO, Finland; E-mail: hans.moen@aalto.fi.

We explore performance of three sentence-level embedding methods. We limited the experiment to using the same clustering approach for all three. Given that our focus is clinical text, which differs from the type of text used when pre-training the used BERT model [5], we hypothesize that the fine-tuning of the model on a proxy classification task with nursing text will yield better embeddings. We also test the Sent2Vec [6] method, which has shown strong performance in unsupervised word and sentence embedding tasks.

2. Methods

2.1. Data

The clinical data set used in this study consists of nursing documents from electronic health records (2005-2020) of almost 94,000 cardiac patients from a Finnish hospital district. Ethical (hospital ethics committee 17.2.2009 §67; UTU ethics committee 9/2020) and administrative approvals were obtained (2/2009; J14/20). In this hospital district, nurses structure their text according to the Finnish Care Classification standard (FinCC) [7], which is a taxonomy (with >600 headings) of nursing diagnoses, interventions and outcomes. We used a subset of about 1 million nursing documents (3.4 million sentences) for model training. Another subset of 20 care episodes was used in the manual evaluation (135 documents, 1032 paragraphs, 2301 sentences).

2.2. Automatic Clustering

We used k-means clustering [8] which is a centroid-based algorithm because it indicated best performance in an initial pilot evaluation. For implementation we used the PyClustering library [9]. Other clustering algorithms tested in the pilot study were OPTICS [10], DBSCAN [11] and Agglomerative Hierarchical clustering (see, e.g., [12]). The Euclidean distance gave better or, at worst, similar results as Cosine similarity based on the pilot test. Since an optimal number of clusters was unknown, we used an automatic approach for this. Both OPTICS and DBSCAN are designed to solve this problem. However, none of these two outperformed k-means clustering together with a technique for determining optimal cluster number. We calculated the Silhouette coefficient for the different number of clusters (k) and picked the one that had the highest coefficient (see e.g. [13]). We used the implementation in Scikit-learn [14]. Another technique considered for determining the optimal cluster number was the elbow technique [15, 16].

2.3. Sentence Embeddings

We explored three embedding methods for generating sentence embeddings:

- `BERT-BASIC` is the BERT model pre-trained for Finnish text on news, online discussions, and internet crawls [5]. When inputting a sentence, the embedding is extracted from the representation of the '[CLS]' token in its last layer.
- `BERT-FINE_TUNED` is the Finnish BERT model further fine-tuned on the nursing text dataset for a proxy task focusing on classifying nursing sentences based on subject headings in the FinCC standard (consists of > 600 headings)

(see [17] for more information about this task). Machine learning libraries used are Huggingface [19], PyTorch [20] and Keras/Tensorflow [21].

- SENT2VEC is an embedding model trained using the Sent2Vec method [6]. It learns static word n-gram embeddings, using an approach similar to C-BOW in Word2vec [1] where the context window equals sentence length. These are combined into sentence embeddings. We trained this on the mentioned nursing text dataset using default parameters except dim=200 and loss=hs. We did not incorporate the FinCC headings here.

2.4. Manual Evaluation

A common way to evaluate clustering results is to manually create gold clusters, and then use a cluster similarity score like the Rand index [18] to compare the generated clusters against. We found manually forming gold clusters very difficult due to the complexity and size of the data. Instead, it was easier for domain experts to evaluate clustering results retrospectively. Thus, we formulated an evaluation scheme where evaluators were instructed to score each cluster according to two criteria - topic coherency (Evaluation A), and uniqueness of topic(s) (Evaluation B). See Table 1. For evaluation B, cases with less than 4 clusters in a care episode, evaluation scores of 2 or 3 were changed to 4 retrospectively to not favor very large and few clusters. A sample of 20 care episodes was used in the manual evaluation (see Data section). Evaluators were two specialists in nursing. Interrater agreement was calculated with Cohen's Kappa.

Table 1. Evaluation scheme used for scoring the clustering results by the different methods

Evaluation A Topic coherency of each cluster		**Evaluation B** Uniqueness of the topic(s) to each cluster	
Class	**Description**	**Class**	**Description**
1-ideal	All sentences cover the same topic(s).	1-ideal	Topic(s) are unique to this cluster.
2-semi-optimal	One topic found here is not in all sentences.	2-semi-optimal	Same topic(s) also occurs in one other cluster.
3-poor	Two of the topics here are not found in all sentences.	3-poor	Same topic(s) occur in two other clusters.
4-very bad	Three or more of the topics are not found in all sentences.	4-very bad	Same topic(s) occur in three or more clusters.
5-unable to assess	-	5-unable to assess	-

3. Results

Evaluation scores are shown in Table 2 (Evaluation A) and Table 3 (Evaluation B). We report the scores on sentence level to compensate for differences in cluster sizes. The BERT-FINE_TUNED method outperformed both BERT-BASIC and SENT2VEC with a larger number of clusters (more sentences) rated class 1 and 2 (ideal and semi-optimal): For the topic cluster coherency evaluation criteria (A), 77.53-78.45% of clusters formed by BERT-FINE_TUNED belong to classes 1 or 2. For BERT-BASIC this number is 59.24-62.53%, while for SENT2VEC it is 34.72-35.59%. When it comes to evaluating the uniqueness of the topic(s) to each cluster, criteria (B), 16.82-26.90% of clusters formed by BERT-FINE_TUNED belong to classes 1 or 2, while this number is 4.19-7.48% for BERT-BASIC, and 11.12-16.04% for SENT2VEC.

Table 2. Scores from manual evaluation A - topic coherency of each cluster.

	SENT2VEC				BERT-BASIC				BERT-FINE_TUNED			
Evaluator:	**1**		**2**		**1**		**2**		**1**		**2**	
Class	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
1	689	30	643	28	1036	45	1060	46	1037	45	1092	47
2	130	6	156	7	403	17	303	13	747	32	713	31
3	135	6	61	3	242	10	132	6	231	10	178	8
4	1345	58	1397	61	612	27	749	33	279	12	259	11
5	2	.1	44	2	8	.4	57	2	7	.3	59	3

Table 3. Scores from manual evaluation B - uniqueness of the topic(s) to each cluster.

	SENT2VEC				BERT-BASIC				BERT-FINE_TUNED			
Evaluator:	**1**		**2**		**1**		**2**		**1**		**2**	
Class	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
1	18	.8	173	7	17	.7	64	3	75	3	323	14
2	238	10	196	8	156	7	49	2	381	17	298	13
3	176	8	223	10	223	10	26	1	537	23	74	3
4	1830	79	1546	67	1872	81	2075	90	1278	55	1546	67
5	39	2	163	7	33	1	87	4	30	1	60	3

The interrater agreement varied between methods and evaluation criteria. For SENT2VEC the overall agreement was *moderate* (A: 0.55, n=489, $p<0.05$; B: 0.43, n=489, $p<0.05$). For BERT-BASIC the agreement was only *fair* to *moderate* (A: 0.47, n=761, $p<0.05$; B: 0.30, n=761, $p<0.05$). Also, for BERT-FINE_TUNED the agreement was *fair* to *moderate* (A: 0.43, n=533, $p<0.05$; B: 0.21, n=553, $p<0.05$).

4. Discussion

As BERT-FINE_TUNED gives the most suited sentence embeddings for this task, this highlights the importance of domain and task specificity in the fine-tuning of these models. For specialized domains in healthcare there are usually very few task-specific labeled datasets available. This study shows that, by formulating a proxy classification task on the data and labels that are available, we can still fine-tune generic language models to better represent the semantics of such specialized text. The evaluation showed that it is an easier task to generate coherent clusters compared to generating clusters with no topical overlap. This mirrors well the holistic nature of nursing documentation, where different aspects of care are interconnected. The interrater evaluation scores confirm the difficulty of determining what constitutes a coherent cluster and how to discern between intercluster overlap. Our scores could also indicate that the interpretation of the evaluation scheme differs somewhat between the evaluators. Further research is needed to build a theoretical framework for clustering and evaluating nursing text as well as for validating the findings of this study with a larger sample.

Study limitations include limited evaluation sample and modest evaluation scheme. The focus of this study was not to find the optimal clustering algorithm or distance metric. Still, better results can likely be achieved through a more thorough evaluation of different clustering algorithms, embedding methods, distance metrics and techniques for determining optimal cluster counts. However, this requires a different evaluation approach than what we have used. Additionally, future research should consider explainability aspects and assess confidence of methods used.

5. Conclusions

Contextual sentence embeddings generated by a BERT model fine-tuned on a proxy classification task shows promising results when used for clustering nursing text from cardiac patients' narratives. The findings can be used to develop tools to augment health science researchers in qualitative analysis of large data sets. As future work we plan to test these embeddings for the purpose of extractive summarization of nursing text.

Acknowledgements: This work was supported by the Academy of Finland (grants 315376, 336033, 315896), Business Finland (grant 884/31/2018), and EU H2020 (grant 101016775). We thank Jari Björne for helping with fine-tuning the BERT model.

References

[1] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs.CL]. 2013 Jan 16.
[2] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. arXiv:1802.05365 [cs.CL]. 2018 Feb 15.
[3] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL 2019. 2019 June; Minneapolis, Minnesota. 4171-86.
[4] Jiang T, Huang S, Zhang Z, Wang D, Zhuang F, Wei F, et al. PromptBERT: Improving BERT Sentence Embeddings with Prompts. arXiv:2201.04337 [cs.CL]. 2022 Jan 12.
[5] Virtanen A, Kanerva J, Ilo R, Luoma J, Luotolahti J, Salakoski T, et al. Multilingual is not enough: BERT for Finnish. arXiv:1912.07076 [cs.CL]. 2019 Dec 15.
[6] Pagliardini M, Gupta P, Jaggi M. Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In: Proceedings of NAACL 2018. 2018, June. New Orleans: Louisiana. 528-40.
[7] Hoffrén P, Leivonen K, Miettinen M. Nursing standardized documentation in Kuopio University Hospital. Stud Health Technol Inform 2009;146:776-7.
[8] Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory 1982;28(2):129-37.
[9] Novikov AV. PyClustering: Data mining library. J Open Source Softw 2019;4(36),1230.
[10] Ankerst M, Breunig MM, Kriegel HP, Sander J. OPTICS: Ordering points to identify the clustering structure. ACM Sigmod record 1999;28(2),49-60.
[11] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD-96 Proceedings. 1996, Aug. 226-231.
[12] Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv 1999 Sept;31(3),264-323.
[13] Frahling G, Sohler C. A fast k-means implementation using coresets. Int J Comput Geom Appl 2006 Jan;18(06),605-25.
[14] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011 Oct;12,2825-30.
[15] Thorndike RL. Who belongs in the family? Psychometrika 1953 Dec;18(4),267-76.
[16] Syakur MA, Khotimah BK, Rochman EMS, Satoto BD. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. IOP Conference Series: Materials Science and Engineering 2018 Apr;336(1),012017.
[17] Moen H, Hakala K, Peltonen LM, Suhonen H, Ginter F, Salakoski T, et al. Supporting the use of standardized nursing terminologies with automatic subject heading prediction: a comparison of sentence-level text classification methods. JAMIA 2020;27(1),81-8.
[18] Hubert L, Arabie P. Comparing partitions. J Classif 1985 Dec;2(1)193-218.
[19] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Huggingface's transformers: State-of-the-art natural language processing. arXiv:1910.03771 [cs.CL]. 2019 Oct 9.
[20] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems. 2019 Dec 8-14. Vancouver: Canada. 8026-37.
[21] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467 [cs.DC]. 2016 Mar 14.