Evaluating Semantic Textual Similarity in Clinical Sentences Using Deep Learning and Sentence Embeddings

Rui Antunes
DETI/IEETA
University of Aveiro
Aveiro, Portugal
0000-0003-3533-8872

João Figueira Silva*
DETI/IEETA
University of Aveiro
Aveiro, Portugal
0000-0001-5535-754X
joaofsilva@ua.pt

Sérgio Matos
DETI/IEETA
University of Aveiro
Aveiro, Portugal
0000-0003-1941-3983
aleixomatos@ua.pt

ABSTRACT

The wide adoption of electronic health records (EHRs) has fostered an improvement in healthcare quality, with EHRs currently representing a major source of medical information. Nevertheless, this process has also brought new challenges to the medical environment since the facilitated replication of information (e.g. using copy-paste) has resulted in less concise and sometimes incorrect information, which hinders the understandability of this data and can compromise the quality of medical decisions drawn from it. Due to the high volume and redundancy in medical data, it is imperative to develop solutions that can condense information whilst retaining its value, with a possible methodology involving the assessment of the semantic similarity between clinical text excerpts.

In this paper we present an approach that explores neural networks and different types of text preprocessing pipelines, and that evaluates the impact of using word embeddings or sentence embeddings. We present the results following our participation in the n2c2 shared-task on clinical semantic textual similarity, perform an error analysis and discuss obtained results along with possible future improvements.

CCS CONCEPTS

• **Applied computing** → **Health care information systems**; **Health informatics**; • **Computing methodologies** → *Natural language processing*; *Neural networks*;

KEYWORDS

Natural language processing, clinical information extraction, semantic textual similarity, deep learning, sentence embeddings

ACM Reference Format:

Rui Antunes, João Figueira Silva, and Sérgio Matos. 2020. Evaluating Semantic Textual Similarity in Clinical Sentences Using Deep Learning and Sentence Embeddings. In *The 35th ACM/SIGAPP Symposium on Applied*

Computing (SAC '20), March 30–April 3, 2020, Brno, Czech Republic. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3341105.3373987>

1 INTRODUCTION

For years, technology advancements have been applied in health care with the goal of preventing, diagnosing and treating diseases, as well as improving the quality of life of the general population. These technological breakthroughs have brought new tools and information sources to physicians, aiding them in clinical workflows like patient follow-up and clinical decision making, whilst also playing an important role in the transition into the personalized medicine paradigm.

Electronic health records (EHRs) are an example of such widely adopted technologies in medicine, providing an electronic infrastructure that can aggregate a multitude of medical information and support the medical act. EHRs provide a longitudinal view of patient trajectory, comprehending the past and present of a patient's health condition¹, and despite containing rich contextual information in structured data that could for instance be explored for prediction modelling purposes [14, 15, 32], large amounts of valuable patient information are stored in unstructured notes, commonly referred to as free text, which are often underexplored due to difficulties in processing this type of text [24].

Due to the large extent of information contained in an EHR, physicians are provided with the key aspect of *context* when reasoning on their medical decisions, making EHR a key component of the patient-centered notion of health care. However, even though its wide adoption has enabled an improvement on healthcare quality, certain challenges also arose with it. An example of such issue is the facilitated process of replicating information in medical text reports through copy-paste actions or by using pseudo-templates, which has impacted on EHR data quality since it can lead to less concise documentation, and an increased chance of introducing erroneous information, that can consequently compromise the quality of the medical act [13, 27].

Owing to the utmost importance of reducing the dimension and redundancy in EHR data, solutions for annotating relevant data and summarizing clinical text [26] have been the focus of much research, mostly targeting clinical natural language processing (NLP). A possible recently explored approach to reduce clinical text redundancy is by assessing the semantic textual similarity (STS) between different text excerpts from an EHR.

¹The EHR can also comprehend a future component of the patient trajectory if future treatments and appointments are considered.

*Contributed equally with the first author to this work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SAC '20, March 30–April 3, 2020, Brno, Czech Republic

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6866-7/20/03...\$15.00

<https://doi.org/10.1145/3341105.3373987>

STS has attracted particular attention in past years. SemEval, an ongoing series of evaluations of computational semantic analysis systems, started a pilot STS challenge track in 2012 which attracted the attention of the research community [5] and, due to its success, progressively organized a series of STS challenge tracks from 2012 through 2017 [1–4, 8]. However, these shared tasks had the major drawback of being centered in general-domain text, whereas clinical text is inherently different in its characteristics. Therefore, an additional effort for the clinical domain was required.

The increasing interest in exploring and pushing forward existing research with clinical data, to further improve healthcare quality, led to the creation of dedicated resources and challenges. To that extent, in recent years, text corpora specifically focused on clinical text were created along with shared tasks on clinical STS that leverage from those corpora. A more detailed coverage on clinical resources and STS challenges is further provided in Section 2.

In the present paper we describe an approach for evaluating the STS between clinical sentence pairs. For that, we explored the use of neural networks and evaluated the usage of different clinical text preprocessing pipelines. Moreover, we compared the impact of using word embeddings *versus* sentence embeddings. The proposed method was used in a participation in the 2019 n2c2 shared task on clinical semantic textual similarity².

2 RELATED WORK

Over the years, healthcare quality has benefited from the *symbiosis* between the medicine and technology fields. Despite the significant research focus given to certain data modalities, such as medical imaging or -omics data, text data from the clinical narratives stored in EHRs has been regaining increased interest since natural language provides a flexible convoy for physicians to track and report each medical situation, enabling complete descriptions and overcoming the limitations from ambiguous and unspecific terms that physicians can find when using coding standards such as RxNorm [23] or SNOMED-CT [29]. Therefore, the potential in free text is great as it can frequently contain information that is otherwise not obtainable from other data sources [16].

However, free text is very challenging to process and, on top of that, can be cluttered with redundant information. Data redundancy is in fact an increasing problem since the digital processes used by physicians to document medical data have the major issue of facilitating information replication (*e.g.* through copy-paste actions), which can reduce the quality of EHR data and increase the probability of introducing incorrect information that can lead to clinical error [13, 27]. This problem has been acknowledged by physicians, who state the need to improve the clinical documentation process to cope with these growing issues [18]. A possible approach to tackle the abovementioned issue is by exploring clinical STS to detect text excerpts containing repeated and/or similar content within EHR documents.

2.1 Biomedical and Clinical STS Resources

Recent years have shown efforts to bring STS to the clinical data domain. In 2017, motivated by the rapidly increasing availability

of textual data in the biomedical domain, by the need to facilitate the retrieval and analysis of this data, and also by the lack of suitable datasets for the development of appropriate systems, Soğancıoğlu *et al.* created BIOSSES: a benchmark dataset containing 100 sentence pairs from biomedical literature where each sentence pair was scored in a scale from 0 to 4 regarding its semantic similarity, and used it for the development of STS methods [28]. Despite being an interesting initial effort, as the type of text found in biomedical literature significantly differs from that of medical narratives, an additional effort was still required.

In 2018, with the goal of exploring clinical STS to reduce clinical text redundancy, Wang *et al.* assembled MedSTS, a dataset containing 174 629 pairs of clinical sentences extracted from a clinical corpus at Mayo Clinic. From this pool of clinical sentence pairs, 1 068 pairs were annotated by two medical experts regarding their semantic similarity, who classified each pair with a value within 0 (dissimilar) and 5 (equivalent), resulting in the creation of the MedSTS_ann dataset. The authors used MedSTS_ann to compare the performance of existing STS approaches on general and clinical domain STS datasets, and as expected observed that performances obtained on MedSTS_ann were in general lower, demonstrating the higher complexity of clinical text [30].

2.2 Clinical STS Shared Tasks

Driven by the interest of motivating the research community to solve real world clinical problems, Wang *et al.* organized a shared task on clinical STS where they released MedSTS_ann, hence making it the first available resource for the study of clinical STS. The pioneering shared task was titled BioCreative/OHNLN Challenge 2018 Task 2: Clinical Semantic Textual Similarity³ and attracted the participation of 4 teams who developed automatic systems to measure the semantic relation between sentence pairs from MedSTS_ann [31].

Submitted solutions explored various techniques ranging from conventional machine learning to deep learning models. The winning system consisted of a regression model applied on 8 trained models: Random Forest, Bayesian Ridge regression, Lasso regression, linear regression, Extra Tree model, Dense Neural Network (DNN) using the Universal Sentence Encoder, DNN using the inferSent encoder, and Encoder-MLP using the inferSent encoder. This system attained the highest Pearson correlation of 0.8328 [9].

The second placing team obtained its best result (0.8143) with an ensemble that averaged the predictions from two models: (i) an Attention-Based Convolutional Neural Network (ABCNN) with NLP features, and (ii) a Bidirectional Long Short-Term Memory network (Bi-LSTM) [33].

The third team used sentence embeddings by performing a weighted average of word vectors followed by a soft projection. To address common component bias in linear sentence embedding, a self-regularized identity map (Conceptors) was used. This approach obtained a Pearson correlation of 0.7789 [20]. The final team attained a best Pearson correlation of 0.7090 and did not disclose a detailed approach.

On a more recent note, building upon the experience from the BioCreative/OHNLN shared task on clinical STS, a collaboration

²<https://portal.dbmi.hms.harvard.edu/projects/n2c2-2019-t1/>

³<https://sites.google.com/view/ohnlp2018/home>

between n2c2 and OHNLP resulted in the 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity with the objective of expanding previous work by providing novel annotated data, thus allowing further development and evaluation of systems on previously unseen data.

2.3 Sentence Representation: BioWordVec versus BioSentVec

Posterior to the BioCreative/OHNLP shared task, members from the winning team proceeded their work by exploring the field of sentence representation, focusing on word and sentence embeddings. Due to the limitations of traditional word embeddings, which are computed at the word level and trained for general domain text, Zhang *et al.* decided to develop a set of biomedical word embeddings that combined subword information from biomedical text with the biomedical vocabulary MeSH (Medical Subject Headings). The resulting biomedical word embeddings, named BioWordVec⁴, performed better than word2vec embeddings in medical word pair similarity benchmarks, and were made publicly available to the community [34]. The authors further improved these embeddings by training them on a new corpus containing over 30 million documents from PubMed articles and clinical notes from the MIMIC-III Clinical Database [17].

Word embeddings capture representations at word level, thus to represent a sentence it is necessary to decompose it in words and combine the representations from each word. However, it is also possible to represent sentence text in a more straightforward approach, using instead sentence embeddings. Similarly to the word embedding scenario, despite the existence of pre-trained sentence encoders for the general domain, biomedical and clinical text remained an unexplored field, thus Chen *et al.* trained sentence embeddings on the same PubMed + MIMIC-III dataset used for BioWordVec. The result was BioSentVec⁴, a set of publicly available sentence embeddings suitable for biomedical and clinical text applications [11].

To assess the adequacy and effectiveness of these sentence embeddings, Chen *et al.* tested BioSentVec with BIOSSES and MedSTS_ann. Whilst using a much simpler 5-layer deep learning model that only received two sentence vectors generated by BioSentVec as input, the authors managed to obtain a Pearson correlation of 0.836 in MedSTS_ann, slightly improving the previous state-of-the-art. This demonstrated that BioSentVec embeddings can effectively capture sentence semantics in clinical text [11]. The authors performed a further comparative evaluation in MedSTS_ann using various sentence similarity models, including the latest bidirectional transformers in the clinical domain, such as BioBERT [19], and observed that (i) the simpler approach proposed in [11] still obtained the best performance, and (ii) embeddings trained on large corpora are the best solution to capture sentence semantics in small datasets [10].

3 MATERIALS AND METHODS

In this work we evaluate a machine learning system in a supervised setting for predicting the semantic relatedness between clinical sentences. For this task, a real value in the interval $[0, 5]$ is attributed to each pair of sentences: if two sentences are completely unrelated

Table 1: Dataset statistics detailing the number of pairs for different levels of similarity. N: number of sentence pairs.

Set	N	Similarity scores				
		[0, 1]	]1, 2]	]2, 3]	]3, 4]	]4, 5]
Training	1642	312	154	394	509	273
Test	412	238	46	32	62	34
Total	2054	550	200	426	571	307

they have a similarity score of 0 (minimum); otherwise if they share the same semantic meaning their similarity value is 5 (maximum). Alike previous challenges on STS [8, 31] the evaluation metric is the Pearson correlation coefficient, where two lists of similarity values (system predictions and ground truth) are compared.

In the following subsections we describe the dataset used in this work, the different approaches used to preprocessing clinical text, the process of feature representation, and finally the deep learning models employed in our simulations.

3.1 Dataset

The dataset contained a total of 2 054 sentence pairs, and it was beforehand split by the n2c2 organizers into training and test subsets. The training data allowed for model development, whereas the test data was used solely for official evaluation in the challenge. Only afterwards, the ground truth scores of the test set were made available to the participating teams. Table 1 presents some statistics about the dataset. Surprisingly, the distribution of the similarity scores between the two subsets (training and test) are somewhat discrepant since the test data contains a higher number of pairs with scores in the interval $[0, 1]$.

A list of example clinical sentence pairs is provided in Table 2, containing the sentence pairs, their respective similarity score, and an explanation of the criteria used to assign the score.

3.2 Clinical Text Preprocessing

Before converting sentences into embedding representations, it was necessary to apply a preprocessing step to the clinical sentence pairs. In this work, three different pipelines were tested.

3.2.1 Base Preprocessing. The baseline preprocessing pipeline was simple, consisting only of two steps: (i) lowercase conversion of the text, and (ii) tokenization using the NLTK tokenizer⁵.

3.2.2 Advanced Preprocessing With Full Stopword Removal. This pipeline was inspired on the preprocessing used by Chen *et al.* in the BioCreative/OHNLP 2018 Task on Clinical Semantic Textual [9], and had the objective of retaining as much semantic information as possible in preprocessed sentences.

The pipeline started with the separation of number ranges (e.g. $0.3-1.8 \rightarrow 0.3$ to 1.8) which are frequent in lab analysis data. The second step was to extend numbers into their textual counterpart

⁴<https://github.com/ncbi-nlp/BioSentVec>

⁵<https://www.nltk.org/>

Table 2: Examples of clinical sentence pairs and respective similary scores with explanations from Wang *et al.* [30].

Score	Examples
5	<p><i>The two sentences are completely equivalent, as they mean the same thing</i></p> <p>S1: Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA Aerosol 2 puffs by inhalation every 4 h as needed</p> <p>S2: Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA Aerosol 1-2 puffs by inhalation every 4 h as needed #1 each</p>
4	<p><i>The two sentences are mostly equivalent, but some unimportant details differ</i></p> <p>S1: Discussed goals, risks, alternatives, advanced directives, and the necessity of other members of the surgical team participating in the procedure with the patient</p> <p>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</p>
3	<p><i>The two sentences are roughly equivalent, but some important information differs/missing</i></p> <p>S1: Cardiovascular assessment findings include heart rate normal, Heart rhythm, atrial fibrillation with controlled ventricular response</p> <p>S2: Cardiovascular assessment findings include heart rate, bradycardic, Heart rhythm, first degree AV Block</p>
2	<p><i>The two sentences are not equivalent, but share some details</i></p> <p>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)</p> <p>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</p>
1	<p><i>The two sentences are not equivalent, but are on the same topic</i></p> <p>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</p> <p>S2: New wheezing or chest tightness, runny or blocked nose, or discharge down the back of the throat, hoarseness or lost voice</p>
0	<p><i>The two sentences are completely dissimilar</i></p> <p>S1: The risks and benefits of the procedure were discussed, and the patient consented to this procedure</p> <p>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</p>

(e.g. 78 → seventy-eight; 0.9 → zero point nine). Next, words connected with slashes, dashes and dots where separated with spaces (e.g. yes/no → yes / no; point-of-care → point - of - care). Then starting white spaces are removed, and double spaces are converted to single spaces.

The resulting text is converted to lowercase and tokenized with the NLTK tokenizer. Finally, a stopword removal is performed using a complete stopword list for biomedical literature⁶, punctuation is removed from the tokens, and tokens composed of a single character are discarded.

3.2.3 Advanced Preprocessing With Partial Stopword Removal. This approach is similar to the previous processing pipeline, differing only in the stopword removal part. Since the list of stopwords

for biomedical text contains terms that can be important for retaining the semantics in clinical text, a smaller stopword list from Luo *et al.* [21] was used instead, being composed of the following terms: 's, a, an, any, her, his, patient, that, the, these, this, those, your.

3.3 Feature Representation

A crucial step in machine learning is feature representation, which aims to transform any kind of data (text, image, and others) into a numeric representation. Regarding textual representation, the bag-of-words technique (one-hot encoding) is commonly employed in traditional machine learning classifiers for tasks such as document classification or word sense disambiguation. However, in the last years, more efficient methods for textual representation have been proposed. Distributed representation of words, or word embeddings,

⁶<https://www.ncbi.nlm.nih.gov/IRET/DATASET>

allow the computation of word vector representations, of relatively small dimension, from large unlabeled textual data [6, 7, 22]. Also, vector representations of sentences, or sentence embeddings, have been investigated [25]. Word and sentence embeddings have been successfully combined with neural network models achieving state-of-the-art results in many NLP tasks.

For this clinical STS task we evaluated, separately, the use of word embeddings and sentence embeddings. We employed the publicly available BioWordVec and BioSentVec models created by Chen *et al.* [11]. To encode the sentences using word embeddings we normalized the sum of the embedding vectors of their constituent words.

3.4 Deep Learning Model

The Keras library [12] was used to implement and test different neural network models. Our proposed model derived from other state-of-the-art works [9, 11] that use word and sentence embeddings. We tested various types and configurations of deep learning models, yet simpler models yielded better results, similarly to what was observed in [11].

We present a neural network whose inputs are the embedding representations of the respective two sentences encoded by (i) word embeddings or (ii) sentence embeddings. In both cases we also included its multiplication (element-wise) and dot product. However, in the latter case we additionally included the cosine similarity since the sentence embedding vectors were not normalized, and the cosine similarity provided valuable information (preliminary results were higher).

The neural network contained a first layer with 512 units and ReLU activation. Also, we used Xavier normal initialization, a bias constant of 0.01 and L2 regularization of 0.001. Afterwards, a dropout rate of 0.4 was set, and a final unit with sigmoid activation performed the predictions. The stochastic gradient descent optimizer was employed with a learning rate of 1.0, and the mean squared error as loss function. A simplified scheme of the neural network with sentence embeddings is presented in Figure 1.

Figure 1: Neural network architecture using sentence embeddings.

For fine-tuning the hyperparameters and adapting the model architecture we used repeated K-fold cross-validation: for each “repetition” we applied cross-validation where we split the training data into three subsets: training, validation, and test. These allowed for consistent model development without biasing in regard to the test set. We used the training subset to update the network weights, the validation subset to evaluate the model performance in each epoch, and finally the test subset was used for unbiased evaluation. The model that obtained the highest result in the validation subset was chosen to evaluate model performance in the test subset. Model training was halted when the performance in the validation subset did not improve for a period of 20 epochs.

After this intensive model refinement with thorough evaluation on training data, the configuration model was left unmodified to be applied on unseen test data. In the next section we evaluate the impact of using different preprocessing approaches, as well as using word embeddings or sentence embeddings as input vectors in the proposed neural network model.

4 RESULTS AND DISCUSSION

For gathering results we applied different text preprocessing pipelines and sentence representations: word embeddings *versus* sentence embeddings. We evaluated performance on (i) training data using repeated cross-validation and on (ii) test data from predictions of a single evaluation. Results presented in the training set were obtained by averaging 30 separate scores, whereas evaluation on test data was slightly different: we averaged the predictions of 30 individual models. The idea behind this was to increase model robustness against the test data, because the final model is able to ‘see’, and learn from, the whole training data by using diverse folds for training and validation.

During model development on training data, a compromise between the sizes of training, validation, and test subsets sizes was necessary. Because of that, we explored the use of different K-fold split values to assess the impact of using less data for training and more for validation, and *vice versa*. We hypothesized an ‘optimal’ threshold considering enough training data and a solid evaluation on validation data could reflect an improvement on unseen data.

Table 3 presents detailed results from all these experiments. In general, the use of sentence embeddings provided superior results especially when using the base text preprocessing. Surprisingly, the full text preprocessing provided better results when using word embeddings, proving the benefit of stopwords removal. To assess the effect of using different K-fold splits, we highlight the top scores according to the evaluation in the training set: for 10, 5, and 3-folds these were, respectively, 0.811, 0.812, and 0.792 in training data, and 0.837, 0.819, and 0.836 in test data. Therefore, we conclude the use of different splits for training and validation did not affect significantly the results, and thus any of the number of folds we used would be acceptable.

The highest scoring model in training data (0.812), which consisted in using sentence embeddings with partial text preprocessing (5-fold setting), produced a correlation of 0.819 in test data. However, better results on test data were achieved when using sentence embeddings with the base text preprocessing (0.837, 0.831, and 0.836 for 10, 5, and 3-folds respectively). Furthermore, it is interesting to

Table 3: Results (Pearson correlation) obtained in the training and test sets by averaging, respectively, the results and the predictions of 30 individual models (number of folds $\times$ number of repetitions). For each distinct K-fold evaluation, the results of the configuration that obtained the highest performance in the training set, together with the corresponding results in the test set, are highlighted in bold. WE: word embeddings normalized sum. SE: sentence embeddings.

Set	Features	Text preprocessing	10-fold (3 repetitions)	5-fold (6 repetitions)	3-fold (10 repetitions)
Training	WE	Base	0.716	0.735	0.672
		Full	0.795	0.794	0.770
		Partial	0.771	0.773	0.728
	SE	Base	0.811	0.810	0.792
		Full	0.750	0.771	0.692
		Partial	0.809	0.812	0.791
Test	WE	Base	0.804	0.786	0.802
		Full	0.823	0.826	0.824
		Partial	0.819	0.818	0.819
	SE	Base	0.837	0.831	0.836
		Full	0.808	0.780	0.802
		Partial	0.818	0.819	0.826

Table 4: Error analysis: performance evaluation for different similarity levels using the following model from Table 3: sentence embeddings, base text preprocessing, and 10-fold evaluation. The three highest F1-scores in each set are highlighted in bold.

Metric	Training					Test				
	[0, 1]	]1, 2]	]2, 3]	]3, 4]	]4, 5]	[0, 1]	]1, 2]	]2, 3]	]3, 4]	]4, 5]
Precision	0.826	0.206	0.361	0.488	0.491	0.853	0.115	0.097	0.365	0.318
Recall	0.416	0.362	0.244	0.625	0.532	0.269	0.391	0.188	0.307	0.618
F1-score	0.553	0.263	0.291	0.548	0.511	0.409	0.177	0.128	0.333	0.420

notice that word embeddings with full text preprocessing also attained good test results (0.823, 0.826, and 0.824 for 10, 5, and 3-folds respectively) similarly to those with sentence embeddings. Based on this, we believe that the combination of word and sentence embeddings may provide further improvement.

Overall, when analysing training and test performances it is noteworthy that systems obtained higher scores on test data by a margin of approximately 0.02. We suspect that one of the reasons for this is the fact that evaluation on test data was performed by averaging several models (where each model was trained and validated on distinct data subsets) whilst the results reported on training data are the average of several simulations made by an individual model.

Finally, we compare our highest test result (0.837) with those achieved during the n2c2/OHNLN 2019 Clinical STS task where a total of 90 system predictions were submitted: final aggregated results (Pearson correlation) presented a mean correlation of 0.712, a median of 0.829, and a maximum of 0.901. We consider that our model attained a positive performance given its simplicity, being slightly above the median score, but also that there still exists large margin for progress given the maximum result achieved.

4.1 Error Analysis

A detailed error analysis was performed to better understand model behaviour — *i.e.* which similarity levels the model predicted more correctly — and to perceive if this could be another reason why test results are higher than those on training (due to the dataset imbalance). As such, we used the same similarity intervals as expressed in the dataset statistics (Table 1): [0, 1], ..., [4, 5].

To perform this evaluation, we started by computing the number of true positives, false positives, and false negatives to calculate precision, recall, and F1-score. A prediction was considered a true positive if the ground truth was in the same similarity interval. Otherwise, the prediction was assigned as false positive and false negative in the corresponding intervals. To demonstrate this procedure, assume the model prediction is 3.7 whereas the ground truth is 4.1. In such case, a false positive and a false negative would be added in the similarity intervals]3, 4] and]4, 5], respectively.

Table 4 presents this detailed error analysis in training and test data, for a model with the following configuration: sentence embeddings, base text preprocessing pipeline, and 10-fold evaluation. Overall, it is noticeable that the system had more difficulty in correctly predicting sentence pairs with scores in the interval]1, 3].

It is interesting to note that the highest precision (around 0.8) was observed in the interval $[0, 1]$ showing the model's ability to correctly detect completely dissimilar sentences. Since the majority of samples (around 58%) in test data were in this interval, this also supports our assumption that test results were higher because of their scores imbalance. Additionally, one can observe that the F1-scores were higher in extreme similarity intervals, corroborating the assumption from Wang *et al.* that machines can succeed at distinguishing completely similar or dissimilar sentence pairs but, unlike humans, they struggle in distinguishing less clear relations of semantic similarity [31].

Finally, despite the fact that F1-scores on test data are somewhat smaller to those on training, we emphasize this type of error analysis does not elucidate directly the model performance, since (i) the test set is highly imbalanced and (ii) near-correct scores can be misinterpreted as complete failures as they fall in a different interval by a slight margin (e.g. the model predicts a score of 0.9 whereas the ground truth is 1.1).

5 CONCLUSIONS AND FUTURE WORK

The increasing involvement of technology in the medical field has resulted in an improvement of the quality of healthcare. However, the wide use of digital processes in medicine does also have its shortcomings, namely in the clinical documentation process where the increased easiness of replicating information results in more redundant free text and in an enhanced chance of introducing erroneous information. This is a serious concern as it can potentially lead to clinical error. A possible solution to reduce redundancy in free text is the use of STS methods to detect text excerpts with equivalent semantic content.

This paper presents the approach used in the 2019 n2c2/OHNLTP Track on Clinical Semantic Textual Similarity. Our system used a simple neural network to tackle the problem of ranking the degree of similarity between snippets of clinical text, and explored the impact of using different preprocessing methods, as well as using different feature representation methods (word embeddings *versus* sentence embeddings).

Our study demonstrated that sentence embeddings provided superior text representation than word embeddings, better capturing sentence semantics. However, word embeddings were not far behind in terms of system performance (Pearson correlation of 0.823 vs 0.837). It was also noticeable that word embeddings benefited from using a more thorough text preprocessing pipeline, whereas sentence embeddings obtained better test results with a basic preprocessing approach.

As future work we intend to explore the fusion of features from word embeddings and sentence embeddings in the same model. Moreover, other architectures could be explored such as convolutional neural networks or bidirectional long short-term memory networks with attention mechanisms, along with other types of embedding representations, such as contextual embeddings. We also aim to investigate data augmentation or distant supervision techniques.

ACKNOWLEDGMENTS

This work is financed by national funds through FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) and by ERDF – European Regional Development Fund, through COMPETE 2020 – Operational Programme for Competitiveness and Internationalisation, within project PTDC/EEI-ESS/6815/2014; POCI-01-0145-FEDER-016694. Rui Antunes and João Figueira Silva are funded by national funds through FCT under grants SFRH/BD/137000/2018 and PD/BD/142878/2018 respectively.

Copyright © 2020 The authors. This is the authors' version of the work. It is posted here for your personal use. Not for redistribution. The definitive version was published in the Proceedings of the 35th ACM/SIGAPP Symposium on Applied Computing and is available online at <https://doi.org/10.1145/3341105.3373987>.

REFERENCES

- [1] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. 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)*. Association for Computational Linguistics, Denver, Colorado, 252–263. <https://doi.org/10.18653/v1/S15-2045>
- [2] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, Dublin, Ireland, 81–91. <https://doi.org/10.3115/v1/S14-2010>
- [3] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. Association for Computational Linguistics, San Diego, California, 497–511. <https://doi.org/10.18653/v1/S16-1081>
- [4] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *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*. Association for Computational Linguistics, Atlanta, Georgia, USA, 32–43. <https://www.aclweb.org/anthology/S13-1004>
- [5] Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: a pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Montréal, Canada, 385–393. <http://dl.acm.org/citation.cfm?id=2387636.2387697>
- [6] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* 3 (Feb. 2003), 1137–1155. <http://jmlr.org/papers/v3/bengio03a.html>
- [7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5, 1 (2017), 135–146. https://doi.org/10.1162/tacl_a_00051
- [8] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: semantic textual similarity - multilingual and cross-lingual focused evaluation. *arXiv:1708.00055* (July 2017). <https://arxiv.org/abs/1708.00055>
- [9] Qingyu Chen, Jingcheng Du, Sun Kim, W. John Wilbur, and Zhiyong Lu. 2018. Combining rich features and deep learning for finding similar sentences in electronic medical records. In *Proceedings of the BioCreative/OHNLTP Challenge 2018*. Washington, DC, USA. <https://sites.google.com/view/ohnltp2018>
- [10] Qingyu Chen, Jingcheng Du, Sun Kim, W. John Wilbur, and Zhiyong Lu. 2019. Evaluation of five sentence similarity models on electronic medical records. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM, Niagara Falls, NY, USA, 533. <https://doi.org/10.1145/3307339.3343239>
- [11] Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2018. BioSentVec: creating sentence embeddings for biomedical texts. *arXiv:1810.09302* (Oct. 2018). <https://arxiv.org/abs/1810.09302>
- [12] François Chollet and Others. 2015. Keras. <https://keras.io/>

- [13] Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2013. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics* 14, 10 (Jan. 2013). <https://doi.org/10.1186/1471-2105-14-10>
- [14] J. C. Ferrão, F. Janela, M. D. Oliveira, and H. M. G. Martins. 2013. Using structured EHR data and SVM to support ICD-9-CM coding. In *2013 IEEE International Conference on Healthcare Informatics*. IEEE, Philadelphia, PA, USA, 511–516. <https://doi.org/10.1109/ICHI.2013.79>
- [15] José Carlos Ferrão, Mónica Duarte Oliveira, Filipe Janela, and Henrique M. G. Martins. 2016. Preprocessing structured clinical data for predictive modeling and decision support. *Applied Clinical Informatics* 07, 04 (Dec. 2016), 1135–1153. <https://doi.org/10.4338/ACI-2016-03-SOA-0035>
- [16] Kasper Jensen, Cristina Soguero-Ruiz, Karl Oyvind Mikalsen, Rolv-Ole Lindsetmo, Irene Kouskoumvekaki, Mark Girolami, Stein Olav Skrovseth, and Knut Magne Auestad. 2017. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific Reports* 7, 46226 (April 2017). <https://doi.org/10.1038/srep46226>
- [17] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3, 160035 (May 2016). <https://doi.org/10.1038/sdata.2016.35>
- [18] Thomson Kuhn, Peter Basch, Michael Barr, Thomas Yackel, and for the Medical Informatics Committee of the American College of Physicians. 2015. Clinical documentation in the 21st century: executive summary of a policy position paper from the American College of Physicians. *Annals of Internal Medicine* 162, 4 (Feb. 2015), 301–303. <https://doi.org/10.7326/M14-2128>
- [19] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* btz682 (Sept. 2019). <https://doi.org/10.1093/bioinformatics/btz682>
- [20] Tianlin Liu, João Sedoc, and Lyle Ungar. 2018. Correcting the common discourse bias in linear representation of sentences using conceptors. In *Proceedings of the BioCreative/OHNL Challenge 2018*. Washington, DC, USA. <https://sites.google.com/view/ohnlp2018>
- [21] Yen-Fu Luo, Weiyei Sun, and Anna Rumshisky. 2019. MCN: a comprehensive corpus for medical concept normalization. *Journal of Biomedical Informatics* 92 (April 2019), 103132. <https://doi.org/10.1016/j.jbi.2019.103132>
- [22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781* (Jan. 2013). <https://arxiv.org/abs/1301.3781>
- [23] Stuart J. Nelson, Kelly Zeng, John Kilbourne, Tammy Powell, and Robin Moore. 2011. Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association* 18, 4 (July 2011), 441. <https://doi.org/10.1136/amiajnl-2011-000116>
- [24] Amy Neustein, S. Sagar Imambi, Mário Rodrigues, António Teixeira, and Liliana Ferreira. 2014. Application of text mining to biomedical knowledge extraction: analyzing clinical narratives and medical literature. In *Text Mining of Web-based Medical Content*. De Gruyter, 3–32. <https://doi.org/10.1515/9781614513902.3>
- [25] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2017. Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv:1703.02507* (March 2017). <https://arxiv.org/abs/1703.02507>
- [26] Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association* 22, 5 (Sept. 2015), 938–947. <https://doi.org/10.1093/jamia/ocv032>
- [27] Hardeep Singh, Traber Davis Giardina, Ashley N. D. Meyer, Samuel N. Forjuoh, Michael D. Reis, and Eric J. Thomas. 2013. Types and origins of diagnostic errors in primary care settings. *JAMA Internal Medicine* 173, 6 (March 2013), 418–425. <https://doi.org/10.1001/jamainternmed.2013.2777>
- [28] Gizem Sogancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics* 33, 14 (July 2017), i49–i58. <https://doi.org/10.1093/bioinformatics/btx238>
- [29] Michael Q. Stearns, Colin Price, Kent A. Spackman, and Amy Y. Wang. 2001. SNOMED clinical terms: overview of the development process and project status. In *Proceedings of the AMLA Symposium*. American Medical Informatics Association, Washington, DC, USA, 662–666. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243297/>
- [30] Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2018. MedSTS: a resource for clinical semantic textual similarity. *Language Resources and Evaluation* (Oct. 2018). <https://doi.org/10.1007/s10579-018-9431-1>
- [31] Yanshan Wang, Naveed Afzal, Sijia Liu, Majid Rastegar-Mojarad, Liwei Wang Feichen Shen, Sunyang Fu, and Hongfang Liu. 2018. Overview of BioCreative/OHNL Challenge 2018 Task 2: clinical semantic textual similarity. In *Proceedings of the BioCreative/OHNL Challenge 2018*. Washington, DC, USA. <https://sites.google.com/view/ohnlp2018>
- [32] Jionglin Wu, Jason Roy, and Walter F. Stewart. 2010. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical Care* 48, 6 (June 2010), S106–S113. <https://doi.org/10.1097/MLR.0b013e3181de9e17>
- [33] Ying Xiong, Shuai Chen, Yedan Shen, Xiaolong Wang, Qingcai Chen, Jun Yan, and Buzhou Tang. 2018. A hybrid system for clinical semantic textual similarity. In *Proceedings of the BioCreative/OHNL Challenge 2018*. Washington, DC, USA. <https://sites.google.com/view/ohnlp2018>
- [34] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data* 6, 1 (May 2019), 52. <https://doi.org/10.1038/s41597-019-0055-0>