Literature review

Semantic Textual Similarity (STS) is one of the core tasks in Natural Language Understanding field. Its objective is to determine whether two pieces of texts (here sentences) are similar or not, sometimes with the use of adequate scores indicating the similarity level. Text similarity assessment is also used as an intrinsic method to evaluate text embedding quality. It is believed that good text vectors should reflect the similarity dependences in testing.

What is more, we recognize two types of text similarity systems: supervised and unsupervised. The first one needs to be trained on an annotated dataset with similarity scores assigned to each sentence pair. The scores reflect the similarity level between two sentences. It can be a regression (i.e. score ranges from 0 to 5) or classification (i.e. labels: similar, dissimilar) model. On the other hand, the unsupervised approach takes two texts as input and try to estimate a similarity score for this pair. It can be done based on some vector similarity metrics. One of the most popular measures is cosine similarity. The key part of the unsupervised task is to develop a good text representation model and use a reliable similarity metric.

The task of sentence similarity indication becomes even harder when it applies to domain-specific corpora. It requires a specialized model which can express the sentence meaning in a vector of features and a specific dataset to evaluate the model on. In this research we will focus on the sentences coming from medical domain, describing patient eligibility for clinical trials. Since there is no annotated corpus for this task, we will develop an unsupervised approach for semantic similarity recognition.

The following literature review consists of two types of articles, describing semantic similarity algorithms for general texts and for biomedical domain. They show different approaches to the problem.

Chen at al. demonstrate a solution which won the Semantic Textual Similarity challenge - an ensemble of two models, Random Forest and Encoder Network. They built the algorithm on a dataset of 1068 annotated sentence pairs from clinical notes. The labels were scores from 0 (completely dissimilar) to 5 (semantically equivalent) provided by two medical experts. The average of their labels was taken as a final score in the dataset. Since there was an annotated dataset available in the challenge, the scientists could create a supervised model for semantic similarity estimation. However, they also report the correlation between the similarity scores and the features. The features were built with unsupervised methods. The preprocessing of the sentences included: converting into lowercase, splitting words joined by punctuation, tokenization with TreeBank, removing stopwords and

punctuation. The traditional machine learning model, Random Forest, used human engineered set of features, which were classified into 5 categories: token-, character-, sequence-, semantic- and entity-based. They either were related the syntax of the sentence or to the semantics. Token-based features were calculated from an unordered list of tokens and included: Jaccard similarity and its generalized version, Dice similarity, Ochiai similarity and TF-IDF similarity. Character-based features were created from substrings of length n (n-grams) where n was set to 3 and 4. The Q-gram similarity was measured. Sequence-based features indicated the number of transformations (insertions, deletions, substitutions) that one needs to apply to transform one sentence into another. The following methods were used to calculate it: Bag similarity, Levenshtein similarity, Needleman-Wunsch similarity and Smith Waterman similarity. This category takes the order of tokens into account. The semantic-based features were created with the use of BioSentVec model and the cosine similarity between them. Entity-based features were built on the clinical concepts extracted with CLAMP and mapped to Concept Unique Identifiers (CUI) with UMLS. Then, the entity similarity was calculated which is the number of entities which share the same CUI divided by the maximum number of CUIs in one of the two sentences. The authors also used an interesting approach for number which often appear in clinical notes. They replaced digits with text and applied Word Mover's Distance to calculate similarity between numerical values. Beside Random Forest, the researchers evaluated also 3 deep learning architectures: Convolutional Neural Network, Recurrent Neural Network and Encoder-Decoder model. The encoder-decoder network with BioSentVec embeddings as input achieved the best performance. The two models, Random Forest and Encoder-Decoder, were finally ensembled with linear regression model and used to predict the similarity score. From all the features Q-gram similarity had the highest correlation with the similarity score given by the annotators. BioSentVec also seemed to be a good approximation. However, the authors proved that the feature significance depends on the text type which is processed, as the features focus on different perspectives i.e. sentences with similar words and different meaning or sentences with different words but the same meaning. Therefore, they recommend to use many features catching different dependences.

On the other hand, Blagec et al. focused on different neural sentence embedding models for semantic similarity assessment of biomedical texts. They evaluated unsupervised and supervised approaches. They used the BIOSSES benchmark dataset. They assessed the unsupervised models by correlation calculation between the estimated cosine similarity score and the score assigned by annotators. Unsupervised algorithms included: fastText (skip-gram and CBOW), sent2vec, skip-thoughts and Paragraph Vector (PV-DM and PV-DBOW). The best results were achieved by the Paragraph Vector (0.819 Pearson correlation) but sent2vec, which is significantly less complex algorithm, also attained high correlation score – 0.798.

Moreover, the authors proved that the quality of the embeddings can be increased by removing long lines from corpus and splitting words joined by hyphens. They also reported a big difference in the skip-gram and CBOW performance (0.766 vs. 0.253 Pearson correlation) which can be caused by the rare words which are hard to be modelled by CBOW. All the unsupervised models were ensembled by averaging their similarity scores and a hybrid model was created. The scientists also evaluated a supervised method, linear regression which predicted the average similarity score based on features derived from unsupervised estimation and string-based similarity indexes – Jaccard and Q-gram. Furthermore, the scientists experimented with the contradiction detection problem and they prepared a small contradiction dataset for that purpose. They found out that the models may achieve good results in semantic similarity estimation bur are not able to indicate contradiction between similar texts. This is caused by a similar context of the words (in contradicting and similar sentences) which results in similar embeddings. The authors suggested that it can be overcome with the use of sentiment polarity labels.

Ranasinghe et al. explored the use of contextualized word representations in STS task. They used four embedding models – ELMo, BERT, Flair and a stacked model, ELMo+BERT. The models were compared to word2vec representations. The scientists focused on three unsupervised methods of similarity estimation. The first one was the cosine similarity of average word vectors in the sentences with two additional options – ignoring stopwords and calculating a weighted average based on TF-IDF scores. Another method was Word Mover's Distance which takes into account the word embeddings from one sentence and measures the minimum distance that they need to move to reach the words from the second sentence. The authors also evaluated the result of removing stopwords. The last investigated method was cosine similarity with Smooth Inverse Frequency which calculates weighted average of word embeddings based on their frequency in a corpus. After that it projects the embeddings on the first principal component and subtract the projections from the embeddings. This solution reduces impact of the most frequent words in the corpus. The authors wanted to explore all three algorithms on general English corpora and other domains. Therefore, there were three annotated datasets, used for evaluation of the created algorithms – SICK dataset (9927 samples), Spanish STS dataset (1250 samples) and biomedical English STS dataset (100 samples). All the algorithms were assessed based on three metrics – Pearson correlation, Spearman correlation and Mean Squared Error, which are very often use for STS task evaluation. All the unsupervised method except for cosine similarity with Smooth Inverse Frequency did not show any improvement when using contextualized word embeddings versus word2vec representations. However, Smooth Inverse Frequency with the embeddings created by ELMO+BERT model achieved the best results. The authors also proved that their solution did not depend on the language features and could be applied to other domains – biomedical and Spanish corpora. They used Bio-BERT, biomedical ELMo, multilingual BERT

and Spanish ELMo for that purpose. They also created stacked models of the relevant ELMo and BERT models. Similar to general English, Smooth Inverse Frequency with the stacked model performed the best on Spanish and biomedical datasets.

There was also a research on clinical trial outcomes. Koroleva et al. investigated differences between planned and reported outcomes by measuring similarity between them. They applied a binary approach for that. It means that there was no scale to assess the similarity level but only two labels – similar or different. If two sentences were related to the same concept (e.g. blood pressure, survival), they were annotated as similar. All the details, e.g. numbers, time constraints, were not taken into account in judgment. Since there was no relevant dataset for that, the authors needed to create a new annotated set of sentence pairs. The researchers investigated deep learning contextualized models – BERT, BioBERT, SciBERT and their different variations – cased and uncased. The models were fine-tuned on the annotated dataset. Similarity labels were assigned with the use of sentence pair classification architecture. The models were compared to different single similarity measures (unsupervised algorithms) from four categories: string measures (Levenhstein distance, proportion of characters occurring in two sentences), lexical measures (proportion of lemmas/stems occurring in two sentences), vector-based measures (cosine similarity between LSA genism vectors, cosine similarity between average of spaCy vectors) and ontology-based measures (path similarity score based on WordNet, Leacock-Chodorow similarity score, Wu-Palmer similarity score). Furthermore, in case of the lexical measures the sentences were preprocessed so that the digits, stopwords and some words with general semantics were removed. All the listed methods returned a score (usually between 0 and 1) which had to be mapped to two categories based on a threshold chosen individually for each metric. The scientists also evaluated nine supervised algorithms – SVM, K-neighbor Classifier, MLP Classifier, Decision Tree Classifier, Random Forest Classifier, Ada Boost Classifier, Extra Trees Classifier, Gradient Boosting Classifier and Gaussian Process Classifier. All the scores calculated with the unsupervised methods were used as features for those algorithms. Among all the models, the BioBERT model achieved the best F1-score. Random Forest Regressor outperformed other basic classifiers and the stem-based algorithm was the winner among the single similarity measures. In addition, it was mentioned that a big advantage of BERT-based algorithms is that they do not need any additional text processing or resources (e.g. UMLS or WordNet). In the error analysis it was shown that some of the sentence pairs required domain knowledge to correctly judge on the similarity. There were also some sentences with different level of detail, which were also hard for the model to process.

Another interesting model achieving high results in semantic textual similarity is Sentence-BERT (Reimers and Gurevych, 2019). The researchers created this model as a response to the poor performance of embeddings extracted from BERT model.

Despite the success that BERT model achieved in many NLP tasks, using it as a model for sentence encoding is not a good choice. The common practice to generate sentence embeddings from BERT is to average its last hidden layer or use the CLS token representation. The authors showed that this approach often achieved worse results in textual similarity that an average of GloVe embeddings. Therefore, they created an SBERT model which used Siamese and triplet networks. Firstly, two sentences were processed by separate BERT models and then there was mean pooling of last hidden layers. The scientists also evaluated max pooling and the CLS token representation performance but the mean pooling achieved the best result. This architecture was tested in three applications – classification, regression and triples loss. In case of classification, there was an additional step implemented before softmax layer – a concatenation of two sentence vectors. In the regression architecture a cosine similarity was used as the last layer. The model was assessed on many different datasets for textual similarity and entailment. It was evaluated in unsupervised and supervised applications. In the unsupervised STS it was shown that the SBERT model outperformed all the state-of-the-art models on 6 out of 7 datasets. However, it needs to be emphasized that the assessed datasets included samples from general English sources. There was no evaluation for the biomedical domain. Furthermore, the authors reported that the fine-tuned SBERT achieved even better results. In all the experiments the embeddings generated by BERT model performed poorly so it was concluded that they are not useful in the cosine similarity measurement.

The reviewed articles concerned different aspects of semantic textual similarity task. They presented both the unsupervised and the supervised approaches. The decision on which one to use depends on the availability of an annotated dataset. Three out of five articles were related directly to the biomedical domain. Instead of using a general English model, their authors applied models trained on the relevant corpora – e.g. BioBERT or BioSentVec and evaluated them on different biomedical datasets. Chen at al. assessed various syntax- and semantic-based unsupervised algorithms, as well as classification models. Blagnec et. al tested different standard text representations (e.g. fastText, sent2vec) and used cosine similarity measure. The research of Koroleva et. al was very strongly related to the clinical trial domain. It was about trial outcome similarity. In contrast to all the other articles, the scientists did not focus on the similarity level estimation but they formulated the task as a binary classification. They used thresholds with similarity metrics to judge if sentences are similar or different. That research assessed contextualized biomedical models – BioBERT and SciBERT. Ranasinghe et. al compared standard word representations to contextualized embeddings and proved that the context has impact on the similarity task. The best model was a stacked ELMo+BERT model. The scientists also showed that cosine similarity with Smooth Inverse Frequency outperformed other unsupervised methods. One of the newest models is Sentence-BERT which is created directly to perform a similarity task. The authors of it proved that the contextualized

representations from BERT do not achieve good results with cosine similarity measure. However, SBERT is not trained on biomedical corpus.

Up to our knowledge there was no research related to semantic similarity of clinical trial eligibility criteria. Moreover, there is no annotated dataset available for that purpose. Evaluation of the described methods on the trial criteria would be very interesting. What is more, there were many new biomedical BERT-based models developed recently, like PubMedBERT, Coder, ClinicalBERT etc. They could be investigated on the criteria similarity task. Another possibility is to apply zero shot learning to recognize the main concepts (topics) of sentences. In this approach the texts with the same concept could be judged as similar. Furthermore, there are many different text preprocessing methods, e.g. UMLS entity linking or abbreviation resolution from ScispaCy which might be evaluated.

References

Blagec, K., Xu, H., Agibetov, A. et al. Neural sentence embedding models for semantic similarity estimation in the biomedical domain. *BMC Bioinformatics* 20, 178 (2019).

Chen, Q., Du, J., Kim, S. et al. Deep learning with sentence embeddings pre-trained on biomedical corpora improves the performance of finding similar sentences in electronic medical records. *BMC Med Inform Decis Mak* **20**, 73 (2020).

Koroleva, A., Kamath, S. & Paroubek, P. Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. *Journal of Biomedical Informatics*. (2019).

Ranasinghe, T., Orasan, C. & Mitkov, R. Enhancing Unsupervised Sentence Similarity Methods with Deep Contextualised Word Representations. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. (2019).

Reimers, N., & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *ArXiv, abs/1908.10084*. (2019).