Towards Recognition of Textual Entailment in the Biomedical Domain

Noha S. Tawfik[1,2](✉) and Marco R. Spruit[2]

[1] Computer Engineering Department, College of Engineering, Arab Academy for Science, Technology, and Maritime Transport (AAST), Alexandria 1029, Egypt
noha.abdelsalam@aast.edu
[2] Department of Information and Computing Sciences, Utrecht University, 3584 CC Utrecht, The Netherlands
{n.s.tawfik,m.r.spruit}@uu.nl
https://www.uu.nl/en/organisation/department-of-information-and-computing-sciences

Abstract. The medical literature suffers from disagreements among authors discussing the same topic or treatment. With thousands of articles published daily, there is a need to detect inconsistent and often contradictory findings. Natural language inference (NLI) gained a lot of interest in the past years, however, domain-specific NLI systems are yet to be examined in depth. In this paper, we conduct several experiments on sentence pairs extracted from PubMed abstracts, to infer whether they express entailment, contradiction or neutral meanings. The main focus of this research is to recognize textual entailment in published evidence-based medicine findings. We explore popular NLI models and sentence embeddings, adapted to the biomedical domain. We further investigate improving the inference detection abilities of the models by incorporating traditional machine learning (ML) features with deep learning (DL) architecture. The proposed model serves in capturing biomedical language's representations by combining lexical, contextual and compositional semantics.

Keywords: Transfer learning · Textual entailment · Sentence embeddings

1 Introduction

In the last decade, the rate of conducting clinical and medical research has changed dramatically, in terms of both quantity and quality. Subsequently, the number of published results in forms of research papers, clinical trials and textbooks has witnessed a growth spurt. Catillon's synthesis [2] estimates that the number of clinical trials has increased from 10 per day in 1975 to 55 and 95 in 1995 and 2015 respectively. In 2017, the PubMed repository contained around 27 million articles, 2 million medical reviews, 500,000 clinical trials and 70,000 systematic reviews. Contribution of medical research is evaluated according to

E. Métais et al. (Eds.): NLDB 2019, LNCS 11608, pp. 368–375, 2019.
https://doi.org/10.1007/978-3-030-23281-8_32

its applicability in the clinical practice and its ability to aid future research in the same field. It is then critical to assess and resonate with published findings specifically when there is more and more evidence on disagreements and contradiction between outcomes [8,10].

Our work aims to improve the process of evaluating scientific contributions by detecting textual inference between results reported in biomedical abstracts. This paper proposes a model for labeling sentence pairs as entailed, contradictory or neutral. The model relies on linguistic and domain-specific hand-crafted features and recent state-of-the-art sentence encoders. The novelty of our approach is the integration of conventional machine learning features with an encoder-based deep neural network.

2 Related Work

Textual entailment has been widely studied in recent years, with the availability of SNLI, MultiNLI corpora. However, most models fail to generalize across different NLI benchmarks [15], moreover they do not perform accurately on domain-specific datasets. In this section we review textual inference models built specifically for the medical domain. Preclude [11] focuses on extracting conflicts found in health discussions posted in online forums on various health-related topics. The system follows a linguistic rule-based approach to detect inter-advice conflicts. It utilizes MetaMap for semantic clause extraction and tokenization, and then assigns polarity to extracted pairs. More recently, Zadrozny et al. suggested a conceptual framework based on the mathematical sheaf model to highlight conflicting and contradictory criteria in guidelines published by accredited medical institutes. It transforms natural language sentences to formulas with parameters, creates partial order based on common predicates and builds sheaves on these partial orders [17].

There were few scattered attempts on extracting contradictions from scientific articles available online. Sarafraz et al. [12], extracted negated molecular events from biomedical literature using a hybrid of machine learning features and semantic rules. Similarly, De Silve et al. [14], extracted inconsistencies found in miRNA research articles. The system extracts relevant triples and scores them according to an appositeness metric suggested by the authors. Alamri et al. [1], detected contradictory findings through n-grams, negation, sentiment and directionality. Our previous work combined a ranking model to find the most relevant finding per abstract and detected biomedical contradictions through semantic features and biomedical word embeddings [16].

3 Methods

3.1 Dataset

In 2016, Alamri et al. published a dataset of contradictory research claims for medical sentence classification and question answering. It is constructed out of

24 systematic reviews on 4 popular cardiovascular disease topics. Medical experts manually mapped each systematic review to a question and extracted corresponding answers from abstracts of articles referenced in the reviews. Only the most relevant sentence is chosen as answer, it is given a *YES* label if it positively answers the question or *NO* label otherwise. More details on the annotation criteria, process and the corpus statistics can be found in [1]. While the dataset is annotated by experts, its structure is not aligned with the language inference task. For that reason, we reconstruct the corpus by combining claims to build a pairwise-sentence corpus to match conventional NLI datasets. We first fetch the PubMed article ids of all 259 abstracts included in the dataset, and extract the first sentence of each abstract. The first sentence in an abstract often describe the research objective. We enrich the corpus by adding extracted sentences and assigning them with the label *NEUTRAL*. Our choice of objective sentence to fill as neutral is based on the general observation of neutral sentences across different NLI benchmarks where they are usually constructed by adding a purpose clause [7]. Given the unique set of medical questions denoted Q where each question is related to only one systematic review and multiple abstracts. For each q_i that belongs to Q, we assumed the following hypotheses while labeling the sentence pairs as entailed, contradictory or neutral:

- $claim_2$ entails $claim_1$ if $asr_2{=}YES$ AND $asr_1{=}YES$
- $claim_2$ contradicts $claim_1$ if $asr_2{=}YES$ AND $asr_1{=}NO$
- $claim_2$ contradicts $claim_1$ if $asr_2{=}NO$ AND $asr_1{=}YES$
- $claim_2$ is neutral to $claim_1$ if $asr_2{=}YES$ AND $asr_1{=}NEUTRAL$
- $claim_2$ is neutral to $claim_1$ if $asr_2{=}NEUTRAL$ AND $asr_1{=}YES$

Where *asr* denotes the assertion value of each sentence with three possible values *YES, NO, NEUTRAL. Claims* refer to the question answer extracted from abstracts. It is important to note that for formulating the above guidelines, a definition of 'entailment' and 'contradiction' is needed. Therefore, we follow the original corpus interpretation of contradiction as "Two texts, T1 and T2, are said to contradict when, for a given fact F, information inferred about F from T1 is unlikely to be true at the same time as information about F inferred from T2". The final dataset consisted of 2135 sentence pairs with 1080, 608 and 447 entailment, contradiction and neutral class instances respectively.

3.2 Machine Learning

Human Engineered Features. The model has a total of 20 traditional NLP features divided into 3 main categories. The main selection criteria of features was to capture context, lexical and semantic representations of text with a limited and optimized feature set.

String-Based Features. This sub category includes *editDist, LevSim, CosSim, JacSim* to represent shortest/longest edit distance, Levenshtein similarity, Cosine similarity and jaccard similarity respectively. In addition, we calculate

4 variations of length measures between the two sentences: *LenMax, LenMin, LenAbs, LenAvg*

Contradiction-Based Features. Negation is still a robust measure of appositeness, we define 4 features to detect negation in sentences. *NegationBin* as a binary feature, *NegOverlap* as the jaccard similarity of negated words only, *AntScore* as a score between the count of antonyms found between sentences. To expand the antonyms coverage we use both WordNet and VerbOcean lexicons, and also *ModOverlap* as the similarity between modal words found in both input. In addition to the above set we also try to detect the outcome polarity through Subjectivity and sentiment (*SubjScore, SentLabel*) using the NLTK sentiment analyzer. Moreover, the results sentence of scientific articles are often accompanied by a "change clause" that affects the final output [9]. The key is to detect whether changes occurring in both sentences are bad, good or neutral. To measure the final pairwise polarity, we include more features such as *PolarityBin* as a binary feature set to 1 when both sentences share the same polarity and 0 otherwise, and *ChangePolarity* that scores each sentence according to a predefined list of change keywords labelled good (+ve score values) or bad (-ve score values).

Context-Based Features. To include domain knowledge we add *EntityOverlap* that computes the similarity between medical UMLS concepts identified by MetaMap[1]. We also rely on word embeddings to capture context. Our hypothesis is that models trained on domain knowledge would generate vector representation capable of learning conceptual meaning of the domain. We compute *EmbedSim* as the cosine similarity between the two embedding vectors and the *EmbedAvg* as the similarity between embedding average for each sentence pooling of all word embeddings. The word embeddings are extracted using FastText model pre-trained on the PubMed Central open access subset[2]. We add the Word Mover's Distance *WMDSim* as measure of similarity between both sentences.

Classification. We experiment with different classification algorithms available in the Scikit-learn toolkit. The experiments include Support Vector machine, Linear regression model, Random Tree, Gradient boost and Naive Bayes.

3.3 Deep Learning

Sentence Embeddings. Text embedding are considered a key element in various NLP tasks. Popular word embeddings such as Word2Vec and GloVe outperform existing models that rely on co-occurrence counts because of their ability to better represent distributional semantics. To encode sentences with one of the prior models, a simple average of their corresponding word embeddings would yield strong results. Nonetheless, during the last two years we witnessed a rise of different supervised and unsupervised approaches towards learning representations of sequences of words, such as sentences or paragraphs. They are able

[1] https://metamap.nlm.nih.gov/.

[2] https://github.com/lucylw/pubmed_central_fasttext_pretrained.

to identify the order of words within a sentence and hence capture more context. The developed sentence representations extend the success of earlier word vectors with interesting results and increasing potential in different tasks. We focus our research on the two of the most popular sentence encoding schemes InferSent and Universal Sentence Encoder. We argue that fine tuning these models and leveraging transfer learning could possibly lead to a good performance in a domain-specific settings. Both chosen encoders were trained partially or fully on textual inference data which fits perfectly with our task.

InferSent is a sentence encoder proposed by Facebook [6]. Its main advantage over other models is its supervised training over SNLI, a large text inference dataset manually annotated. The original model[3] is trained on 570k human-generated English sentence-pairs with a bi-directional Long Short Term Memory (BiLSTM) encoder.

Universal Sentence Encoder (USE) was developed by Google [3]. It has two variations, the first is a transformer-based encoder which yields high-accuracy at the cost of high complexity and extra computational resources. The second model uses a deep averaging network that averages word embeddings and serve as input to a deep neural network. In our model, we deploy the transformer architecture as it was proven to yield better results in several NLP tasks. The universal sentence encoder[4] training data contains supervised and unsupervised sources such as Wikipedia articles, news, discussion forms, dialogues and question/answers pairs. It is also partially augmented with instances from the SNLI corpus.

Deep Learning Network. Our DL model follows a siamese-like architecture where the first set of layers are parallel duplicates and share same weights. For merging the two inputs, we concatenate the element-wise difference and then multiply both vectors. Following that, there are multiple intermediate dense layers. The nodes are directly connected to the nodes in the next layer and use rectified linear activation (ReLU) function. Given the small dataset size, we introduce a dropout layer with a dropout rate of 0.3. Finally, the prediction layer with 3 nodes predicts the probability of each of the inference classes, and a softmax activation function. We adopt an exponentially decaying learning rate, and an l2 regularization weight of 0.01.

3.4 A Feature-Assisted Neural Network Architecture Model

With the small size of the dataset, traditional features demonstrate good performance in comparison with the neural network models. This, along with more evidence on the usefulness of combining traditional features in deep learning architecture [5,13], encouraged us to build a hybrid model. An essential dilemma for building the feature-assisted model is how to incorporate engineered features

[3] Pre-trained model for InferSent available at https://github.com/facebookresearch/InferSent.

[4] Pre-trained model for USE available at https://tfhub.dev/google/universal-sentence-encoder-large/3.

to sentence embeddings inputs. Directly appending the traditional ML features to the encoded representations generated from InferSent or USE would not influence the performance. In that scenario, the features' effect on the classification decision would almost be nonexistent given the large size of sentence encoding vector versus the feature vector size of 21 values. Figure 1 gives an overview of the final feature-assisted framework we propose.

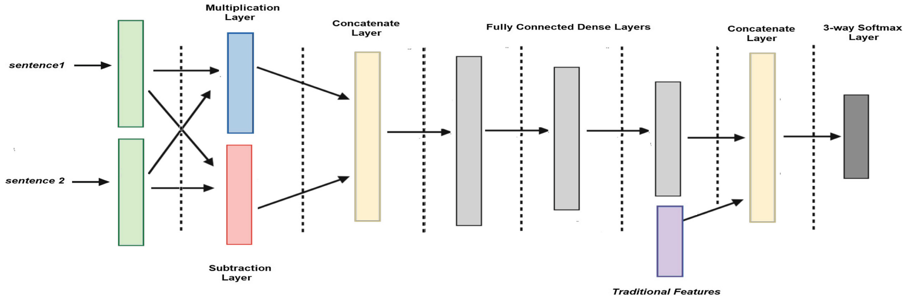

Fig. 1. The feature-assisted neural network architecture.

4 Results and Evaluation

All the following results are calculated as the average results of standard cross-validation with 10 folds. The results reported for the machine learning approach are the output of the best two classifiers: Random Forest (RF) and extreme gradient boosting (XGBoost). It is generally observed that XGBoost almost always achieves higher accuracy than RF. Table 1 shows the results details of the model, The baseline performance is 50.56% based on the majority classifier output. We note, that the ML experiments were not meant for direct comparison with the DL model. The conducted evaluations serve at choosing the best feature combination that could further boost the DL network.

Table 1. Machine learning features with Random Forest and XGBoost classifiers based on 10-fold cross validation. Reported numbers correspond to average accuracy and standard deviation

Feature set	Random Forest	XGBoost
context-based	53.26% (+/- 1.80%)	49.16 % (+/- 2.67%)
contradiction-based	67.81% (+/- 1.28%)	69.49% (+/- 1.77%)
context + string	61.01% (+/- 1.97%)	64.91% (+/- 2.98%)
all features	72.30% (+/- 2.32%)	**76.94% (+/- 1.24%)**

As for the deep learning algorithms, we ran multiple experiments while varying the number of hidden layers and the corresponding number of nodes. Adding more layers test our model capacity, in other terms, with small number of layers the model may struggle to fit the data. On the other hand, over-scaling the network size leads to great results on training data and performs poorly on the test data. Our experiments show that there was a minimal overfitting effect with increasing the number of layers, however, there was no added accuracy. Deep Learning experiments' results are shown in Table 2. In all cases, InferSent encoder outperforms USE encoder with approximately 8%. This finding is consistent with previous published findings [4]. Both encoders are considered universal and should represent sentences efficiently given the amount of data they are trained on. The performance difference between the two encoders could be attributed to the difference in the embedding vector dimension (512 vs 4096) and the nature of inference data *InferSent* is trained on. We added the traditional features to the best performing model with 3 layers and a number of nodes decreasing by 50% with each hidden layer that is deeper in the neural network. No remarkable achievement were noticed in the *USE* encoder case(only 0.6% difference). However, the hybrid model achieves the best result with an average accuracy of 96.21% and a minimum of 94.32% when combined with the *InferSent* encoder. Even with a limited dataset, the results suggest that the machine learning features and deep learning models are complementary. Their combination in an end-to-end model can enhance the learning process and improve the predictions on unseen data.

Table 2. Deep Learning performance results on 10-fold cross validation with respect to the number of hidden layers in the DNN architecture. Reported numbers corresponds to average accuracy and standard deviation

Hidden layers	Hidden units	USE *(Dim.:512)*	InferSent *(Dim:4096)*
1 layers	512	72.56% (+/- 1.14%)	89.88% (+/- 3.91%)
3 layers	512,256,128	82.27% (+/- 1.63%)	**93.95% (+/- 1.39%)**
3 layers	512,256,64	83.17% (+/- 2.20%)	93.86% (+/- 1.48%)
5 layers	512,256,256,128,128	83.68% (+/- 1.50%)	92.24% (+/- 0.79%)
3 layers	512,256,128,64,64	83.68% (+/- 1.50%)	93.18% (+/- 1.73%)

5 Conclusion

We attempt to detect medical text inference from published scientific articles. Various experiments have been applied in different scenarios including ML features and DL network built on top of sentence encoders. Our proposed hybrid architecture is the optimal configuration in terms of size and number of hidden layers. The final results are promising, however, the model must be re-evaluated on a larger corpus to generalize its effect. We could enhance the sentence encoder power by re-training them on domain-specific sources such as research articles and clinical notes. We also believe that a feature ablation analysis over a bigger range of features could potentially select a better boosting vector for assisting the neural network.

References

1. Alamri, A.: The detection of contradictory claims in biomedical abstracts. Ph.D. thesis, University of Sheffield (2016)
2. Catillon, M.: Medical Knowledge Synthesis: A Brief Overview (2017). https://www.hbs.edu/faculty/Pages/item.aspx?num=54337
3. Cer, D., et al.: Universal Sentence Encoder. arXiv preprint, March 2018
4. Chen, Q., Kim, S., Wilbur, W.J., Du, J., Lu, Z.: Combining rich features and deep learning for finding similar sentences in electronic medical records. In: Proceedings of the BioCreative/OHNLP Challenge 2018 (2018)
5. Chen, R.C., Yulianti, E., Sanderson, M., Bruce Croo, W.: On the benefit of incorporating external features in a neural architecture for answer sentence selection. ACM Ref. Format (2017). https://doi.org/10.1145/3077136.3080705
6. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. arXiv e-prints, May 2017. http://arxiv.org/abs/1705.02364
7. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., Smith, N.A.: Annotation artifacts in natural language inference data. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2018)
8. Ioannidis, J.P.A.: Why most published research findings are false. PLoS Med. **2**(8), e124 (2005). https://doi.org/10.1371/journal.pmed.0020124
9. Niu, Y., Zhu, X., Li, J., Hirst, G.: Analysis of polarity information in medical text. In: AMIA ... Annual Symposium proceedings. AMIA Symposium 2005, pp. 570–574 (2005)
10. Prasad, V., Cifu, A., Ioannidis, J.P.A.: Reversals of established medical practices: evidence to abandon ship. Jama **307**(1), 37–38 (2012)
11. Preum, S.M., Mondol, A.S., Ma, M., Wang, H., Stankovic, J.A.: Preclude2: personalized conflict detection in heterogeneous health applications. Pervasive Mob. Comput. **42**, 226–247 (2017). https://doi.org/10.1016/J.PMCJ.2017.09.008
12. Sarafraz, F.: Finding conflicting statements in the biomedical literature. Ph.D. thesis, University of Manchester (2012)
13. Sequiera, R., et al.: Exploring the Effectiveness of Convolutional Neural Networks for Answer Selection in End-to-End destion Answering. arXiv e-prints (2017)
14. de Silva, N., Dou, D., Huang, J.: Discovering inconsistencies in PubMed abstracts through ontology-based information extraction. In: ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB) (2017)
15. Talman, A., Chatzikyriakidis, S.: Testing the generalization power of neural network models across NLI benchmarks. Technical report (2018)
16. Tawfik, N.S., Spruit, M.R.: Automated contradiction detection in biomedical literature. In: Perner, P. (ed.) MLDM 2018. LNCS (LNAI), vol. 10934, pp. 138–148. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-96136-1_12
17. Zadrozny, W., Garbayo, L.: A sheaf model of contradictions and disagreements. Preliminary report and discussion. In: International Symposium on Artificial Intelligence and Mathematics, Florida