Assertion Detection in Clinical Notes: Medical Language Models to the Rescue?

**Betty van Aken[1], Ivana Trajanovska[1], Amy Siu[1],
Manuel Mayrdorfer[2], Klemens Budde[2] and Alexander Löser[1]**

[1]Beuth University of Applied Sciences Berlin, [2]Charité Universitätsmedizin Berlin

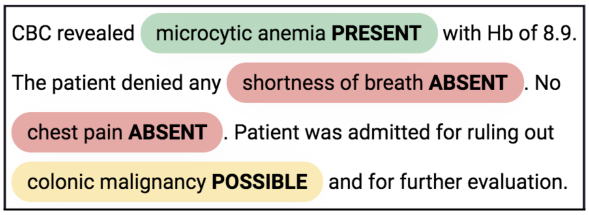
ivtrajanovska@gmail.com
{bvanaken, siu, aloeser}@beuth-hochschule.de
{manuel.mayrdorfer, klemens.budde}@charite.de

Abstract

In order to provide high-quality care, health professionals must efficiently identify the presence, possibility, or absence of symptoms, treatments and other relevant entities in free-text clinical notes. Such is the task of assertion detection – to identify the assertion class (*present*, *possible*, *absent*) of an entity based on textual cues in unstructured text. We evaluate state-of-the-art medical language models on the task and show that they outperform the baselines in all three classes. As transferability is especially important in the medical domain we further study how the best performing model behaves on unseen data from two other medical datasets. For this purpose we introduce a newly annotated set of 5,000 assertions for the publicly available MIMIC-III dataset. We conclude with an error analysis that reveals situations in which the models still go wrong and points towards future research directions.

1 Introduction

The clinical information buried in narrative reports is difficult for humans to access for clinical, teaching, or research purposes (Perera et al., 2013). To provide high-quality patient care, health professionals need to have better and faster access to crucial information in a summarized and interpretable format. In this paper, we focus on English discharge summaries and the task of assertion detection, which is the classification of clinical information as demonstrated in Figure 1.



CBC revealed microcytic anemia **PRESENT** with Hb of 8.9. The patient denied any shortness of breath **ABSENT**. No chest pain **ABSENT**. Patient was admitted for ruling out colonic malignancy **POSSIBLE** and for further evaluation.

Figure 1: Sample output of our demo system. Detected entities are highlighted in red, yellow, and green to indicate *present*, *possible*, and *absent*.

Given a piece of text, we need to identify two pieces of information – a medical entity and textual cues indicating the presence or absence of that entity. Medical entity extraction has been studied extensively (Lewis et al., 2020), we thus focus our work on the task of predicting the *present* / *possible* / *absent* class over a medical entity, addressing an important information need of health professionals. This setting is reflected in the dataset released by the 2010 i2b2 Challenge Assertions Task (de Bruijn et al., 2011a), on which we base our main evaluation.

Clinical assertion detection is known to be a difficult task (Chen, 2019) due to the free-text format of considered clinical notes. Detecting *possible* assertions is particularly challenging, because they are often vaguely expressed, and they occur far less frequently than *present* and *absent* assertions. Language models pre-trained on medical data have shown to create useful representations for a multitude of tasks in the domain (Peng et al., 2019). We apply them to our setup of assertion detection to evaluate whether they can increase performance (especially on the minority class) and where they still need improvement.

We argue that clinical assertion detection models must be transferable to data that differs from the training data, e.g. due to different writing styles of health professionals from other clinics or from other medical fields. As existing datasets do not represent such diversity, we manually annotate 5,000 assertions in clinical notes from several fields in the publicly available MIMIC-III dataset. We then use these annotated notes as an additional evaluation set to test the transferability of the best performing model.

Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, pages 35–40
July 6, 2021.

		present	*possible*	*absent*
2010 i2b2 Challenge Assertion Task	discharge summaries	21,064	1,418	6,144
BioScope	scientific publications	–	3,474	2,161
MIMIC-III Clinical Database (New)	discharge summaries	2,610	250	980
	physician letters	204	34	66
	nurse letters	293	14	59
	radiology reports	249	40	130

Table 1: Distribution of text types and classes in the three employed datasets. Note that *possible* is a minority class across datasets as well as text types. In the i2b2 dataset, for instance, only 5% of all labels are *possible*.

Our **contributions** are summarized as follows:
1) We evaluate medical language models on assertion detection in clinical notes and show that they clearly outperform previous baselines. We further study the transferability of such models to clinical text from other medical areas.
2) We manually annotate 5,000 assertions for the MIMIC-III Clinical Database (Johnson et al., 2016). We release the annotations to the research community[1] to tackle the problem of label sparsity and the lack of diversity in existing assertion data.
3) We conduct an error analysis to understand the capabilities of the best performing model on the task and to reveal directions for improvement. We make our system publicly available as a web application to allow further analyses[2].

2 Related Work

One of the earliest approaches to assertion detection is NegEx (Chapman et al., 2001), where hand-crafted word patterns are used to extract the *absent* category of assertions in discharge summaries. In 2010, the i2b2 Challenge Assertions Task (de Bruijn et al., 2011a) was introduced, and an accompanying corpus was released.

There is a variety of prior work focused on scope resolution for assertions, which differs from our setting in that it does not consider medical concepts but scopes of a certain assertion cue. Representative current approaches for this task setup include a CNN-based (Convolutional Neural Network) one by Qian et al. (2016), reaching an F1 of 0.858 on the more challenging *possible* category. Sergeeva et al. (2019) propose a LSTM-based (Long Short-Term Memory) approach to detect only *absent* scopes. When "gold negation cues" are made available to the model and synthetic features are applied, an F1 of 0.926 is reached. NegBert (Khandelwal and Sawant, 2020) is another approach to detect *absent* scopes. As its name suggests, it is BERT-based and reaches an F1 of 0.957 on BioScope abstracts.

In contrast to these approaches we focus our work on entity-specific assertion detection, the results of which are of more practical help for supporting health professionals. Bhatia et al. (2019) explored extracting entities and negations in a joint setting, whereas the work of Harkema et al. (2009), Chen (2019) and de Bruijn et al. (2011a) is the closest to our task setup, i.e. labelling entities with an assertion class. Harkema et al. (2009) extended the NexEx algorithm with contextual properties. de Bruijn et al. (2011a) use a simple SVM classifier and Chen (2019) apply a bidirectional LSTM model with attention to the task and evaluate it on the i2b2 corpus. While these models reach F1-scores above 0.9 on the majority classes, the challenging *possible* class does not surpass 0.65. We show that medical language models outperform these scores especially regarding the minority class.

Furthermore, Wu et al. (2014) compared then state-of-the-art approaches for negation detection and found a lack of generalisation to arbitrary clinical text. We thus want to examine the transfer capabilities of recent language models to understand whether they can mitigate the phenomenon.

3 Methodology

We want to understand the abilities of medical language models on the task of assertion detection. We hence fine-tune various (medical) language models on the i2b2 corpus described below. We further apply the best performing model to the BioScope dataset and our newly introduced MIMIC-III assertion dataset without further fine-tuning to test their performance on unseen medical data.

[1] Annotated data available at: https://github.com/bvanaken/clinical-assertion-data

[2] Demo application: https://ehr-assertion-detection.demo.datexis.com

Model	F1 for		
	present	*possible*	*absent*
Earlier approaches			
SVM Classifier (de Bruijn et al., 2011b)	0.959	0.643	0.939
Conditional Softmax Shared Decoder (Bhatia et al., 2019)	–	–	0.905
Bi-directional LSTM with Attention (Chen, 2019)	0.950	0.637	0.927
Language models under evaluation			
BERT Base (Devlin et al., 2019)	0.968	0.704	0.943
BioBERT Base (Lee et al., 2020)	0.976	0.759	0.963
Bio+Clinical BERT (Alsentzer et al., 2019)	0.977	0.775	0.966
Bio+Discharge Summary BERT (Alsentzer et al., 2019)	**0.979**	**0.786**	**0.972**
Bio+Clinical Outcome Representations (CORe) (van Aken et al., 2021)	0.975	0.761	0.965
Biomed RoBERTa Base (Gururangan et al., 2020)	0.976	0.723	0.967

Table 2: Results of baseline approaches and (medical) language models on the i2b2 Assertions Task. Pre-trained medical language models outperform all earlier approaches – with a large margin on the *possible* class. Note that Bhatia et al. (2019) only evaluated their model on negation detection.

3.1 Datasets

The **2010 i2b2 Assertion Task** (de Bruijn et al., 2011a) provides a corpus of assertions in clinical discharge summaries. The task is split into six classes, namely *present*, *possible*, *absent*, *hypothetical*, *conditional* and *associated with someone else*. However, the distribution is highly skewed, such that only 6% of the assertions belong to the latter three classes. Hence we only use the *present*, *possible*, and *absent* assertions for our evaluation as they present the most important information for doctors.

BioScope (Vincze et al., 2008) is a corpus of assertions in biomedical publications. It was specifically curated for the study of negation and speculation (or *absent* and *possible* in this paper) scope and does not contain *present* annotations. As mentioned before, the BioScope dataset does not completely match the information need of health professionals and the i2b2 corpus lacks varied medical text types. We thus introduce a new set of labelled assertions to complement existing data.

The **MIMIC-III Clinical Database** (Johnson et al., 2016) provides texts from discharge summaries as well as other clinical notes (physician letters, nurse letters, and radiology reports) representing a promising source of varied medical text. Therefore, two annotators followed the annotation guidelines from the i2b2 challenge, and labelled 5,000 assertions, i.e. word spans of entities and their corresponding *present* / *possible* / *absent* class. The inner-annotator agreement as Cohen's kappa coefficient is **0.847**, which indicates a strong level of agreement. The annotations were further verified by a medical doctor, who provided feedback to correct a small number of labels, and confirmed that the end results were satisfactory.

It is important to note that even though the newly annotated data from MIMIC-III adds variation to the existing corpora, the dataset has its own limitations. The clinical notes are collected from a single institution (with a mostly White patient population) and from Intensive Care Unit patients only. We therefore argue that progress in assertion detection requires further initiatives for releasing more diverse sets of clinical notes.

Table 1 summarizes the assertion distribution in the introduced datasets and shows the unbalanced nature of the data.

3.2 Data Preprocessing

We make predictions about assertions on a per-entity level. However, we want our models to consider the context of an entity. We therefore pass the whole sentence to the models and surround the entity tokens with special *indicator* tokens `[entity]` whose embeddings are randomly initialised. A sample input sequence thus looks as follows: `[CLS] test results were negative for [entity] COVID-19 [entity].`
We apply the same pre-processing to all three datasets.

3.3 Fine-tuning Medical Language Models

There are various pre-trained (bio-)medical and clinical language models available to evaluate on the assertion detection task. We select the most prevalent ones and describe them in short below:

	present	*possible*	*absent*
BioScope			
scientific pub.	–	0.593	0.845
MIMIC-III			
discharge sum.	0.951	0.663	0.939
phys. letters	0.929	0.593	0.892
nurse letters	**0.967**	**0.710**	0.900
radio. reports	0.950	0.691	**0.977**

Table 3: Experimental results (in F1) for the best performing Bio+Discharge Summary BERT model on two further assertion datasets and their different text types. Both datasets were not seen during training. Note that the number of evaluation samples is very low for some text types (i.e. *possible* class in nurse letters), which impairs the expressiveness of these results.

BERT (Devlin et al., 2019) was pre-trained on non-medical data and serves as a baseline for Transformer-base pre-trained language models. **BioBERT** (Lee et al., 2020) is a standard model for medical NLP tasks and is pre-trained on bio-medical publications. **Bio+Clinical BERT** and **Bio+Discharge Summary BERT** (Alsentzer et al., 2019) are built upon BioBERT with additional pre-training on clinical notes / discharge summaries. The **CORe** model (van Aken et al., 2021) uses BioBERT and adds a specialized clinical outcome pre-training. **Biomed RoBERTA** (Gururangan et al., 2020) is based on the RoBERTA model (Liu et al., 2019) and pre-trained on bio-medical publications. After an initial grid search we fix our hyperparameters to a learning rate of 1e-5, batch size of 32, and 2 epochs of training.

4 Evaluation and Discussion

We start by evaluating the mentioned models on the i2b2 corpus. We use training and test data as defined by in the i2b2 challenge and compare our results to previous state-of-the-art approaches in Table 2. Next, we apply the best performing Bio+Discharge Summary BERT to the BioScope and MIMIC-III corpora without additional fine-tuning (Table 3). This way we can see the model's performance on medical text from unseen sources.

4.1 Results

Language models outperform baselines. Table 2 shows that all evaluated medical language models are able to increase F1-scores on all three classes. On the most challenging *possible* class the improvement is the clearest with up to ∼15pp, which shows that the models are better in handling sparse occurrences coupled with vague expressions.

Medical pre-training is important. The vanilla BERT baseline is the weakest of our evaluated models, which shows that models specialized on the medical domain are not only effective for more complex medical tasks but also for assertion detection, which is in line with the claim by Gururangan et al. (2020) that domain-specific pre-training is almost always of use. Bio+Discharge Summary BERT is the best model – probably because it was trained on text very similar to the i2b2 corpus.

Text style matters. Table 3 shows the ability of the Bio+Discharge Summary BERT language model to transfer to other text styles. The assertions in the BioScope corpus are difficult to identify by the model as they clearly differ from the ones used by doctors in clinical notes. The text style in MIMIC-III data is more similar to the originally learned data which is reflected in the results.[3] However, physician letters appear to contain more specialized expressions and therefore evoke more errors. This points towards a lack of generalization possibly caused by the limited variety of assertion cues in the training data.

4.2 Error Analysis

We analyse all errors made by the best model to identify main sources of errors and to point towards future research directions.

Inconsistent data in pre-existing datasets account for roughly 45% of errors. This includes obvious labelling mistakes, but also disagreements among annotators. For example, phrases such as "appeared to be," "concerning for" and "consistent with" are labeled differently, as *present* or as *possible*.

Long range dependencies account for roughly 20% of all errors, in which entities and their cues have dependencies longer than a few tokens apart. While the model's attention mechanism could easily detect distant tokens, the model might have learned to only consider close assertion cues. The following is an example of a distant cue indicating the *absent* class which was missed by the model:

> His rash on the right hand was examined further and is now resolved.

[3] Note that the model's pre-training is based on MIMIC-III and it was thus to an extent exposed to the test data. Due to the difference of the target task and the amount of total pre-training data, this influence should be negligible.

Lists of assertions are found in 8% of error samples. Here the assertion is not directly coupled to an entity but must be inferred by the way it is listed. Such somewhat ambiguous cases are usually easily understood by humans, but difficult for our models.

> No hydrocephalus, subarachnoid hemorrhage, no fracture.

Misspellings account for 5% of all observed errors, but they reveal a critical yet surprising limitation. For instance, the cues "appeas" and "probalbe" that indicate *possible* instances, are missed. While Transformer-based models are generally capable of dealing with misspellings due to subword tokenization, the missing variety of expressions in the data appears to let the models focus on a specific set of textual cues without generalizing to new phrases or even misspellings.

5 Conclusion and Future Work

In this work, we present an evaluation on medical language models to detect assertions in clinical texts and experimental results which show that they outperform baseline approaches. We further provided a new corpus of assertion annotations on the MIMIC-III dataset that will augment existing data collections and shows the model's capability to be transferred to other sources – if the text styles do not strongly differ. We suggest future work to investigate generalization to unseen data and expressions. We further encourage work on multi-task learning of entity extraction and assertions to support health professionals with systems that learn jointly in an end-to-end fashion.

Acknowledgments

Our work is funded by the German Federal Ministry for Economic Affairs and Energy (BMWi) under grant agreement 01MD19003B (PLASS) and 01MK2008MD (Servicemeister).

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Parminder Bhatia, Busra Celikkaya, and Mohammed Khalilia. 2019. Joint Entity Extraction and Assertion Detection for Clinical Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 954–959.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Long Chen. 2019. Attention-based Deep Learning System for Negation and Assertion Detection in Clinical Notes. *International Journal of Artificial Intelligence and Applications*, 10(1).

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 2011a. Machine-learned Solutions for Three Stages of Clinical Information Extraction: The State of the Art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5):557–562.

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel D. Martin, and Xiaodan Zhu. 2011b. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J. Am. Medical Informatics Assoc.*, 18(5):557–562.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy Webber Chapman. 2009. Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *J. Biomed. Informatics*, 42(5):839–851.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, 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(1):1–9.

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A Transfer Learning Approach for Negation Detection and Scope Resolution. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5739–5748.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Patrick S. H. Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop, ClinicalNLP@EMNLP 2020, Online, November 19, 2020*, pages 146–157. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 58–65. Association for Computational Linguistics.

Sujan Perera, Amit Sheth, Krishnaprasad Thirunarayan, Suhas Nair, and Neil Shah. 2013. Challenges in Understanding Clinical Notes: Why NLP Engines Fall Short and Where Background Knowledge Can Help. In *Proceedings of the 2013 International Workshop on Data Management & Analytics for Healthcare*, page 21–26.

Zhong Qian, Peifeng Li, Qiaoming Zhu, Guodong Zhou, Zhunchen Luo, and Wei Luo. 2016. Speculation and Negation Scope Detection via Convolutional Neural Networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 815–825.

Elena Sergeeva, Henghui Zhu, Peter Prinsen, and Amir Tahmasebi. 2019. Negation Scope Detection in Clinical Notes and Scientific Abstracts: A Feature-enriched LSTM-based Approach. *AMIA Summits on Translational Science Proceedings*, 2019:212.

Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A. Gers, and Alexander Löser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2021*. Association for Computational Linguistics.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and Their Scopes. *BMC bioinformatics*, 9(11):1–9.

Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. 2014. Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PLoS One*, 11(9).