Assertion Detection in Electronic Health Records

Ivana Trajanovska
01.09.2020

“Society is aging and healthcare costs keep rising. By digitizing the system, health services can be provided at lower cost and higher quality”

Introduction

Related Work

Methodology

Evaluation

Conclusion

1

Introduction

Motivation, Task Definition

Introduction

Motivation

- Increasing number of Electronic Health Records (EHRs)
- Clinicians need to have a better access to important information
- Extracting knowledge from EHRs
- High quality patient care

Introduction

What is Assertion Detection?

“Assertions are an attribute of the medical problem concepts that are marked in the concept extraction task”

Types of assertions:

- **Present**
- **Absent**
- **Possible**
- Conditional
- Hypothetical
- Not associated with the patient

Introduction

Task Definition

Given an entity in a medical text, identify its asserted class from the context

Two steps approach

1. Given a paragraph, detect the entities it contains.

   She showed signs of pneumonia, but has no pain ⟶ pneumonia, pain

2. Having the entities marked in a paragraph, identify their asserted class

   She showed signs of pneumonia, but no pain ⟶ABSENT

Introduction

Scope and Limitations

Limitations

- Not enough labeled data
- BioScope dataset - Radiology reports not available
- NegPar dataset - Not free of charge

Scope

- 2010 i2b2/va challenge on assertions
- MIMIC - III data → requires manual labeling

  Discharge Summaries, Radiology Reports, Nurse Reports, Physician Letters
- Focus on Present, Absent and Possible

Introduction
Hypotheses

Research Question 1

- The chosen fine-tuned model BioBERT + Discharge Summaries should surpass the current state-of-the-art models

Research Question 2

- The model can be transferred to the same task on datasets coming from different distributions

2

Related work

Current solutions and Language Models

Related Work
Current Solutions

Rule Based model focused on negation - Negex (Chapman et al., 2001)

2010 i2b2 challenge on Assertion Detection

- Best model - SVM with F1 of 93.62 (de Bruijn et al., 2011)

Conditional Softmax Shared Decoder on negation (Bhatia et al., 2019)

- F1 score of 90.5

Assertion Detection - Bidirectional LSTM with Attention (Chen, 2019)

- F1 of 95 on the Present class, 93 on the Absent class and 64 on the Possible class

Related Work
Language Models

Word embeddings - low-dimensional, continuous, dense vectors

- word2vec

Language models - Contextualized word representation

- ELMo - Embeddings from Language Models
- **BERT - Bidirectional Encoder Representations from Transformers**

3

Methodology

Data and Architecture

Methodology

What are discharge summaries?

"Clinical reports prepared by a health professional at the conclusion of a hospital stay or series of treatments"

CHIEF COMPLAINT AND HISTORY OF PRESENT ILLNESS: Pt. 111 is a 45-year-old female with squamous cell carcinoma of the top of mouth (stage T2 N0) that was biopsied by her dentist. Pathology was reviewed revealing invasive cell carcinoma. The possibility of metastatic carcinoma could not be excluded. She presented on 2018-01-23 for resection.

Sections with most relevant information:

- History of Present Illness
- Past Medical History
- Impression
- Chief Complaint

Methodology

2010 i2b2/VA challenge on assertions - Discharge Summaries

Subset of Classes

- Present – includes all problems that are present in a patient.
- Absent – indicates that a specific medical problem doesn't exist in a patient.
- Possible – asserts that there is some possibility that a patient has a specific medical problem.

Present	Absent	Possible
21064	**6144**	**1418**

Methodology
MIMIC - III Data Annotation

Annotation Guideline from 2010 i2b2/VA challenge on assertions

Two annotators

Doccano as annotation tool

Cohen’s kappa coefficient as our evaluation metric

- Score of 0.847, strong level of agreement.
- 64 - 81% of the data are reliable

Methodology

MIMIC - III Data distribution on annotated samples

A freely accessible critical care database containing anonymized data associated with patients who stayed in critical care units of the Beth Israel Deaconess Medical Center

Dataset	Present	Absent	Possible
Discharge Summaries	**2613**	**980**	**250**
Physician Letters	**204**	**66**	**34**
Nurse Letters	**293**	**59**	**14**
Radiology Reports	**249**	**130**	**40**

Methodology
BioScope

Data on Negation and Speculation (Absent and Possible)
- Abstracts and Papers only available

	Clinical	Papers	Abstracts
Negation samples	**871**	**404**	**1757**
Speculation samples	**1137**	**783**	**2691**

Methodology

Architecture - Step 1 - Given a paragraph, detect the entities it contains

Named Entity Recognition Task

Comparison of two models on subset of i2b2 data

NER-style F1

- TeXoo, F1 of 46
- ScispaCy, F1 of 69

Methodology

Architecture - Step 2 - Having the entities marked in a paragraph, identify their asserted class

- Classification task
- BioBERT + Discharge summaries and a classification layer on top

Methodology
Pipeline

Methodology
Final Product

An endpoint to test the model

History of present illness : A 36-year-old male with history of myocardial infarction in 2019-09-30 with stent to the LAD and 50% to the mid LAD , had no signs of restenosis in 2018-04-02 and then underwent brachytherapy to the RCA , there his vitals were initially stable with a hct of 36.7, though he was felt to be hypovolemic.

Present Absent Possible

4

Evaluation

Results, Comparisons and Error Analysis

Evaluation

Hyperparameter Optimization

Parameter	Values
Learning Rate	**1e-5**, 2e-5, 3e-5
Batch Size	16, **32**
Weighted Cross Entropy	True, **False**
Epochs	**2**, 3

Evaluation
Results

	Precision	Recall	F1
Present	**0.9877**	**0.9795**	**0.9836**
Absent	**0.9832**	**0.9927**	**0.9879**
Possible	**0 .8091**	**0.8641**	**0.8357**

F1 Macro - 0.9357

Evaluation

Adding MIMIC - III (250) samples to the Possible class

	Precision	Recall	F1
Present	**0.9855**	**0.9825**	**0.9840**
Absent	**0.9926**	**0.9806**	**0.9866**
Possible	**0 .7946**	**0.8641**	**0.8279**

F1 Macro - 0.9328

Evaluation
Comparison with and without MIMIC-III

	F1 i2b2 only	F1 with MIMIC
Present	0.9836	**0.9840**
Absent	**0.9879**	0.9866
Possible	**0.8357**	0.8279

The initial model is used further

Evaluation

Human Baseline - 20% of test data

Cohen's kappa score - 0.7382 → 35–63% of the data are reliable

88% overlap on samples from the Present class, 75% overlap on Absent samples, and 72% overlap on samples labeled as Possible

	Annotator 1	Annotator 2	Our model
Present	0.941	0.937	**0.98**
Absent	0.915	0.846	**0.993**
Possible	0.625	0.6	**0.717**

Evaluation

Comparison with current solutions

	Present	Absent	Possible
Conditional Softmax Shared Decoder (Bhatia et al., 2019)	-	0.905	-
Bidirectional LSTM with Attention (Chen, 2019)	0.950	0.927	0.637
BioBERT + Discharge Summaries (ours)	**0.984**	**0.988**	**0.836**

Evaluation

F1 Scores of MIMIC-III, BioScope

	Present	Absent	Possible
BioScope	-	0.8446	0.5930
MIMIC			
Discharge Summaries	0.9513	0.9389	0.6330
Physician Letters	0.9292	0.8923	0.5926
Nurse Letters	0.9673	0.9	0.7097
Radiology Reports	0.9501	0.9766	0.6914

Evaluation

Error Analysis - i2b2

Predicted Label

True label	Present	Absent	Possible
Present	0.98	0	0.02
Absent	0.01	0.99	0
Possible	0.13	0.01	0.86

Evaluation

Error Analysis - i2b2

Typos → *probalbe, appeas*

Wrongly labeled test samples → *likely* and *concerning* should be labeled as Possible; *not*, *no* and *resolved* should be labeled as Absent

Overall labeling inconsistency → *appeared to be, concerning for and consistent with* sometimes labeled as Present, sometimes as Possible

Model weaknesses

→ sensitive to longer dependencies

May be either viral or secondary to resolving abdominal pain with resultant hematoma

→ detecting wrong classes where no key phrases are present

His hospital course was remarkable for ruling in for pneumonia

Evaluation

Error Analysis - MIMIC III

Data Processing → Lost information and samples missing context

Annotators' mistakes and disagreements → *possibly consistent with* should be labeled as Possible; *no* or *not* found to be Present

Labeling disagreements → *unlikely* labeled as Possible, but in i2b2 it's found to be Absent

Model weaknesses

→ sensitive to longer dependencies

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

→ listed entities

no hydrocephalus, subarachnoid hemorrhage, no fracture

Evaluation

BioScope

Unseen patterns → *hypothesise, raises the question, instead, cannot*

Disagreement between annotators → *apparent, assumed* labeled as Present in the i2b2 dataset; *estimated* labeled as Possible in BioScope

Model weaknesses → *may, would* and *was not* usually followed after a very long entity

Trouble with non-diseases

→ long entity

It strongly suggests the iORF is actually two adjacent genes → Present

→ same example with a disease

It strongly suggests pneumonia → Possible

5 Conclusion

Discussion and Future Work

Conclusion

Discussion

Annotation guideline perceived differently

→ adding more data to training set confused the model

Possible class found challenging

Model capability

→ surpassed current best solutions

→ can generalize to other EHRs, has trouble with other general purpose texts

Model reliability

Present and Absent class have high Recall scores (0.9795 and 0.9927)
Model is highly confident about its predictions

Techniques that are found helpful

- Testing - model should undergo rigorous tests in order to gain more trust
- Boundary conditions - specifying a set of boundary conditions and rules the data should fulfill
- Explainability - most mistakes are due to inconsistent labeling (mostly because of the Possible class)

Conclusion

A more detailed annotation guideline, a clear definition of the Possible class, and/or a strong supervision by clinicians

Adding more layers

The authors of (Liu, 2019) show that adding an additional layer at the beginning of BERT can improve its overall performance

Syntactic dependency

Should be important as the model will have an additional information about the dependency between the entities and cues

Including experts

Will benefit the annotation process

Interpretability

Try to explain the decision that BERT makes and what happens within its layers

DEMO

Thank you!

QUESTIONS?

Literature

S. Hehner, S. Biesdorf, and M. Möller. Digitizing healthcare–opportunities for germany, Mar 2020. URL https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/digitizing-healthcare-opportunities-for-germany#.

S. Perera, A. Sheth, K. Thirunarayan, S. Nair, and N. Shah. 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, DARE '13, page 21–26, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450324250. doi: 10.1145/2512410.2512427. URL https://doi.org/10.1145/2512410.2512427.

J. C. Feblowitz, A. Wright, H. Singh, L. Samal, and D. F. Sittig. Summarization of clinical information: A conceptual model. J. of Biomedical Informatics, 44(4):688–699, Aug. 2011. ISSN 1532-0464. doi: 10.1016/j.jbi.2011.03.008. URL https://doi.org/10.1016/j.jbi.2011.03.008.

J. Alammar. The illustrated transformer, Jun 2018a. URL https://jalammar.github.io/illustrated-transformer/

J. Alammar. The illustrated bert, elmo, and co. (how nlp cracked transfer learning), Dec 2018b. URL https://jalammar.github.io/illustrated-bert/

J. Alammar. Visualizing a neural machine translation model (mechanics of seq2seq models with attention), May 2018c. URL https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

J. Alammar. A visual guide to using bert for the first time, Nov 2019. URL https://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/.

O. Uzuner, B. South, S. Shen, and S. DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association : JAMIA, 18:552–6, 06 2011. doi: 10.1136/amiajnl-2011-000203.

Literature

W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics, 34 5:301–10, 2001.

P. Bhatia, B. Celikkaya, and M. Khalilia. 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, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1091. URL https://www.aclweb.org/anthology/P19-1091.

B. de Bruijn, C. Cherry, S. Kiritchenko, J. Martin, and X. Zhu. 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, sep 2011. doi: 10.1136/amiajnl-2011-000150. URL https://doi.org/10.1136%2Famiajnl-2011-000150.

L. Chen. Attention-based deep learning system for negation and assertion detection in clinical notes. International Journal of Artificial Intelligence and Applications, 10:1–9, 2019.

S. Arnold, F. Gers, T. Kilias, and A. Löser. Robust named entity recognition in idiosyncratic domains. Aug 2016.

M. Neumann, D. King, I. Beltagy, and W. Ammar. ScispaCy: Fast and robust models for biomedical natural language processing. In Proceedings of the 18th BioNLP Workshop and Shared Task, pages 319–327, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5034. URL https://www.aclweb.org/anthology/W19-5034.

M. McHugh. Interrater reliability: The kappa statistic. Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB, 22:276–82, 10 2012. doi: 10.11613/BM.2012.031.

Literature

S. Kamalodeen. How to write a discharge summary: Discharge letter, Jul 2020. URL https://geekymedics.com/how-to-write-a-discharge-summary/.

A. Johnson, T. Pollard, and R. Mark. Mimic-iii clinical database, Sep 2016a. URL https://physionet.org/content/mimiciii/1.4/.

A. Johnson, T. Pollard, and R. Mark. Mimic-iii clinical database demo (version 1.4), Apr 2019. URL http://dx.doi.org/10.13026/C2XW26.

G. Szarvas, V. Vincze, R. Farkas, and J. Csirik. The bioscope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. BioNLP, pages 38–45, 07 2008.

E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott. Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://www.aclweb.org/anthology/W19-1909.