Are we there yet? Exploring clinical knowledge of BERT models

Madhumita Sushil, Simon Šuster and Walter Daelemans

University of Antwerp
CLIPS | Centre for Computational Linguistics and Psycholinguistics

THE UNIVERSITY OF MELBOURNE

BERT for medical language inference

MedNLI: How well do BERT models perform?

MedNLI: Given a pair of sentences about patient health, identify entailment/contradiction/neutral relations between them.

Model	MedNLI-dev	MedNLI-test
BERT-base-uncased	82.1	77.8
BERT-base-cased	79.9	78.8
BERT-base-cased + PMC + PubMed (BioBERT v1.0)	84.3	82.5
BERT-base-cased + Pubmed 1M (BioBERT v1.1)	**84.8**	**82.9**
SciBERT-base-uncased (SciBERT vocab)	81.5	82.2

What type of examples do the models fail on?

Manual analysis of 50 errors in dev set

with BioBERT v1.1 model

Error Type	Count (of 50)
Insufficient domain knowledge	20
Spurious correlations / dataset bias	6
Difficult instance	5
Incorrect numeric inference	4
Incorrect negation	3
Incorrect tense resolution	2
Incorrect temporal sequence inference	2
Modifier ignored	2
Incorrect abbreviation understanding	2
Lexical (P,H) overlap	2
Insufficient commonsense knowledge	1

Error Examples

Error Type	Example
Insufficient domain knowledge	P: ... she was treated with Benadryl ... H: Patient has had an allergic reaction ~~Entailment~~ *Neutral*
Spurious correlations / dataset bias	P: She spoke with her oncology team ... H: The patient has cancer. ~~Neutral~~ *Entailment*
Incorrect numeric inference	P: ... an ejection fraction of 69% with normal wall motion. H: patient has normal cardiac output ~~Entailment~~ *Contradiction*
Incorrect negation resolution	P: ... no identified sepsis risk factors. H: ... has multiple risk factors for sepsis ~~Contradiction~~ *Entailment*

Error Examples

Error Type	Example
Incorrect tense resolution	P: ... he had a CT of the chest and CTA of his coronary arteries ... H: patient will go for coronary angiography ~~*Neutral*~~ *Entailment*
Incorrect temporal inference	P: ... biopsy ... showed signs of rejection ... subsequently did well. H: The patient had transplant failure ~~*Contradiction*~~ *Entailment*
Modifier ignored	P: Left common femoral dorsalis pedis bypass graft. H: Patient has CAD ~~*Neutral*~~ *Entailment*
Incorrect abbreviation understanding	P: Her ... PO intake have been normal. H: She has been NPO since midnigh ~~*Contradiction*~~ *Neutral*
Insufficient commonsense knowledge	P: ... status post high speed motor vehicle crash ... H: Patient has recent trauma ~~*Entailment*~~ *Neutral*

Augmenting clinical knowledge in BERT models

Knowledge graphs vs. textual resources

Knowledge graphs:

- \+ Concretely defined relationships
- \- Expensive to construct
- \- Incomplete
- \- Unavailable for low-resource languages

Textual resources:

- \+ Easy to obtain for different tasks, languages
- \- Difficult to parse concrete relations
- \- Difficult to identify what's relevant

Knowledge graphs vs. textual knowledge resources

Knowledge graphs:

\+ Concretely defined relationships

\- Expensive to construct

\- Incomplete

\- Unavailable for low-resource languages

Textual resources:

\+ Easy to obtain for different tasks, languages

\- Difficult to parse concrete relations

\- Difficult to identify what's relevant

Our research focus: Textual corpora with fundamental knowledge

Medical Textbook
3.6M tokens

Medical subset of Wikipedia
40M tokens

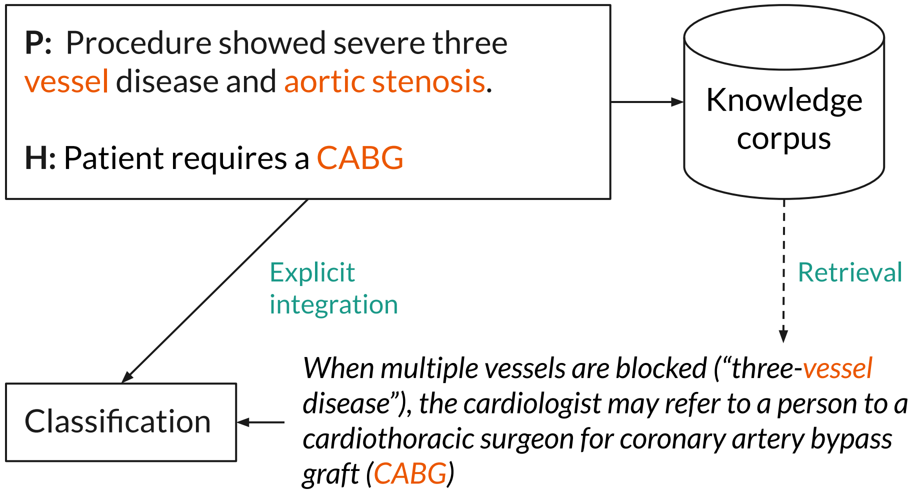
Knowledge integration techniques

Implicit: Further masked language modeling

Explicit: Top *k* sentences from external corpora that establish a relation between P, H entities are appended before classification

1. **Static retrieval:** BM25 + dependency paths are used to find relevant sentences
2. **Dynamic retrieval:** End-to-end retriever and classifier training

   Weighted dot product between instance, context embeddings used for retrieval; weights learned during training. Additional retrieval loss (optional).



Results

No significant difference in results with knowledge augmentation.

Wikimed seems to be better knowledge source than Medbook, potentially due to its larger size.

Model	MedNLI-dev	MedNLI-test
BioBERT v1.1	84.8	82.9
He et al. (2020): BioBERT v1.1 + disease	NA	82.2
Sharma et al. (2019) (+UMLS)	NA	79.0
BioBERT v1.1 + Wikimed MLM	84.2	83.3
BioBERT v1.1 + Medbook MLM	83.2	80.1
BioBERT v1.1 + Wikimed (static)	83.9	83.1
BioBERT v1.1 + Medbook (static)	83.8	82.5
BERT-base-uncased	82.1	77.8
BERT-base-uncased + jointly trained Wiki retriever	79.4	78.5
BERT-base-uncased + trained Wiki retriever + retrieval loss	79.1	77.9

Example retrieval

Method	Text
Example	P: Infusion stopped and she was treated with Benadryl 50 mg x 1, prednisone 40 mg x 1, ativan 1 mg. H: Patient has had an allergic reaction
Gold retrieval	Benadryl is a brand name for a number of different antihistamine medications used to stop allergies, including diphenhydramine, acrivastine and cetirizine.
Static retrieval	Prednisone is used for many different autoimmune diseases and inflammatory conditions, including asthma, COPD, CIDP, rheumatic disorders, allergic disorders, ..., and as part of a drug regimen to prevent rejection after organ transplant.
Dynamic retrieval	Gemeprost (16, 16-dimethyl-trans-delta2 PGE methyl ester) is an analogue of prostaglandin E. It is used as a treatment for obstetric bleeding. It is used with mifepristone to terminate pregnancy up to 24 weeks gestation. Vaginal bleeding, cramps, nausea, vomiting, loose stools or diarrhea, headache, muscle weakness; dizziness; flushing; chills; backache; dyspnoea; chest pain; palpitations and mild pyrexia. Rare: Uterine rupture, severe hypotension, coronary spasms with subsequent myocardial infarctions. ...

Potential reasons for failure despite success in QA

More difficult task setup than span-identification:

- Finding a passage that somewhat looks like question isn't sufficient; context needs to explicitly define relationship between (P, H) entities
- No good heuristics to determine which (P, H) entity pairs should be considered
- Extremely large search space in external corpora

Conclusions

BioBERT models, although good, still make several errors on examples requiring domain knowledge for inference.

State-of-the-art solutions lead to unreliable knowledge augmentation for language inference.

Efforts need to be concentrated towards developing methods to augment fundamental domain knowledge from textual corpora to solve the problem of advanced knowledge-based reasoning.