Joint Sentence Classification of Medical Paper Abstracts with Neural Networks and Syntactic Features

Carlos Sathler

May 3, 2018

SCHOOL OF INFORMATICS AND COMPUTING

INDIANA UNIVERSITY

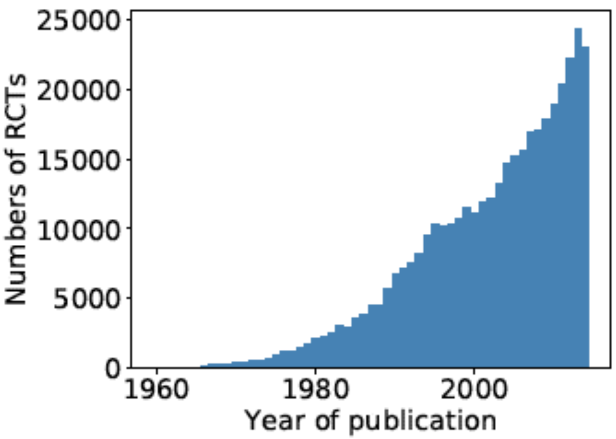
Department of Information and Library Science
Bloomington

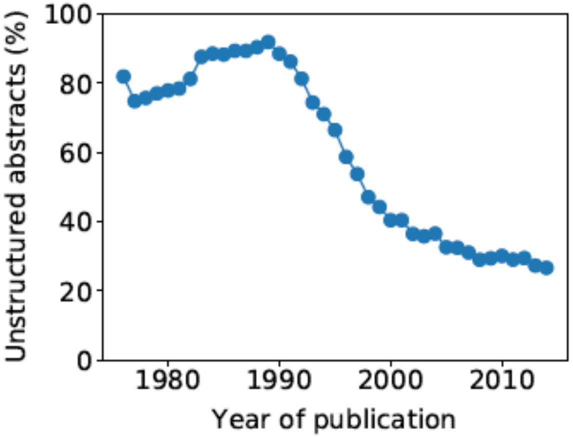
Agenda

- Motivation
- Background and related work
- Dataset
- Approach
- Experiments
- Results
- Conclusions

Motivation

- Evidence Based Medicine ”requires new skills of the physician, including efficient literature searching”*





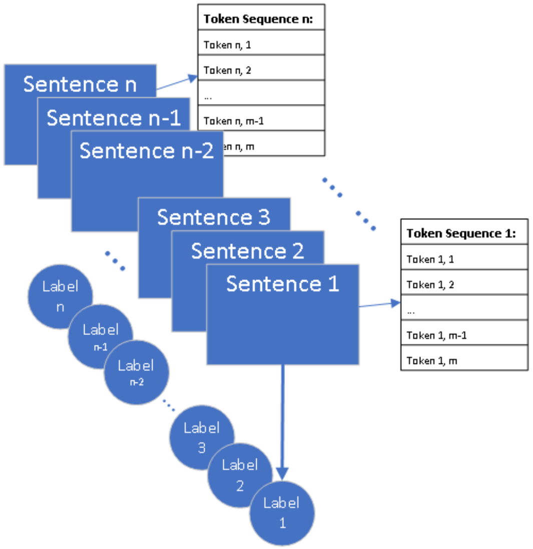
* Evidence-Based Medicine Working Group. "Evidence-based medicine. A new approach to teaching the practice of medicine." *Jama* 268.17 (1992): 2420.

Source of plots: Dernoncourt, Franck, and Ji Young Lee. "PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts." *arXiv preprint arXiv:1710.06071* (2017).

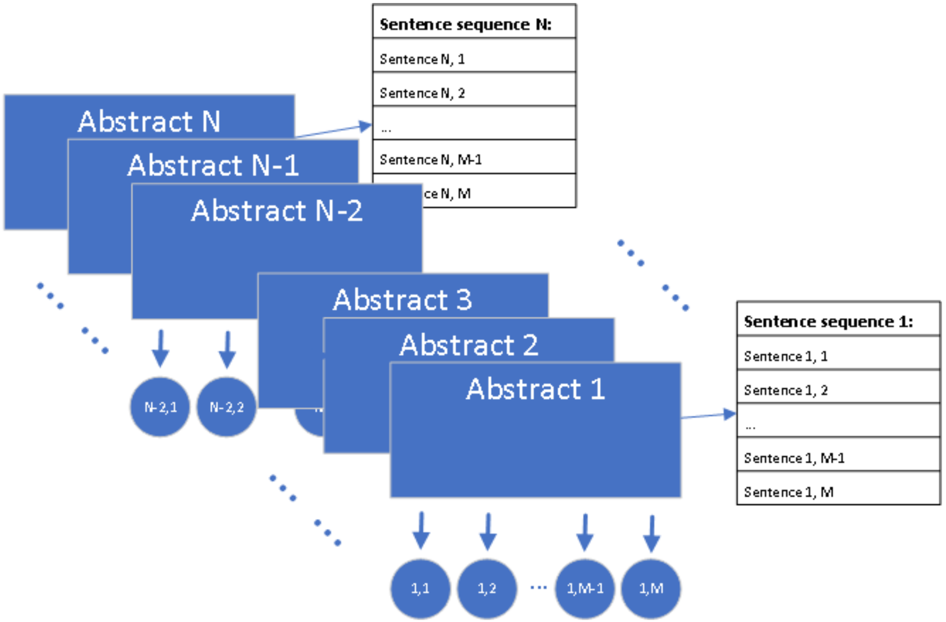
Background and related work

- Moschitti, Alessandro, and Roberto Basili. "Complex linguistic features for text classification: A comprehensive study." *European Conference on Information Retrieval*. Springer, Berlin, Heidelberg, 2004.
- Dernoncourt, Franck, Ji Young Lee, and Peter Szolovits. "Neural Networks for Joint Sentence Classification in Medical Paper Abstracts." *arXiv preprint arXiv: 1612.05251* (2016).
- Dernoncourt, Franck, and Ji Young Lee. "PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts." *arXiv preprint arXiv: 1710.06071* (2017).

Joint sentence classification



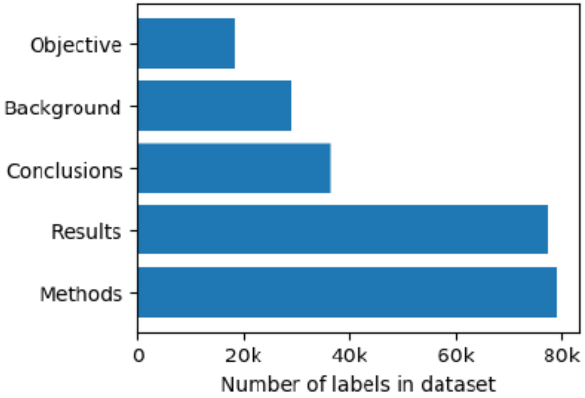
Sentence Classification

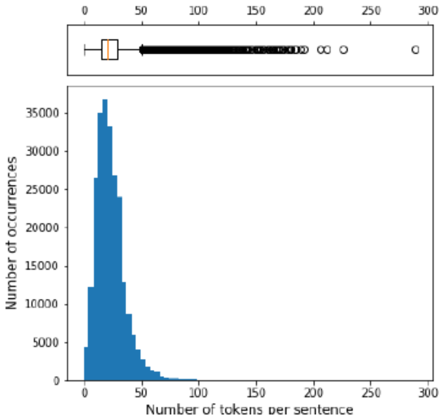


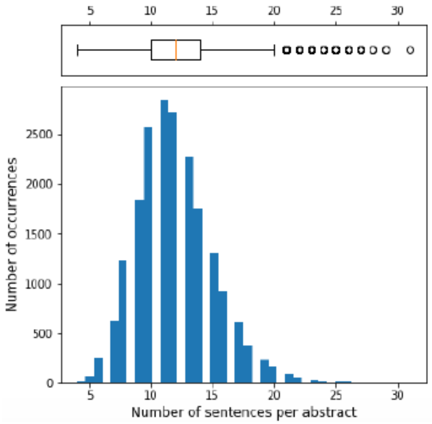
“Joint” Sentence Classification

Dataset: PubMed 200k RCT*

partition	abstract_id	seq	text	label
train	4293578	0	To investigate the efficacy of 6 weeks of daily low-dose oral prednisolone in improving pain , mobility , and systemic low-grade inflammation in the short term and whether the effect would be sustained at 12 weeks in older adults with moderate to severe knee osteoarthritis (OA) .	OBJECTIVE
train	4293578	1	A total of 125 patients with primary knee OA were randomized 1:1 ; 63 received 7.5 mg/day of prednisolone and 62 received placebo for 6 weeks .	METHODS
train	4293578	2	Outcome measures included pain reduction and improvement in function scores and systemic inflammation markers .	METHODS
train	4293578	3	Pain was assessed using the visual analog pain scale (0-100 mm) .	METHODS
train	4293578	4	Secondary outcome measures included the Western Ontario and McMaster Universities Osteoarthritis Index scores , patient global assessment (PGA) of the severity of knee OA , and 6-min walk distance (6MWD) .	METHODS



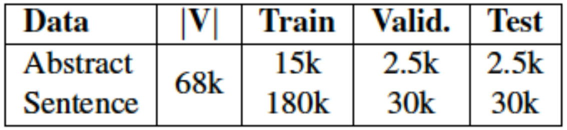




* Dernoncourt, Franck, and Ji Young Lee. "PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts." *arXiv preprint arXiv:1710.06071* (2017).

Approach

- Use delivered partitions



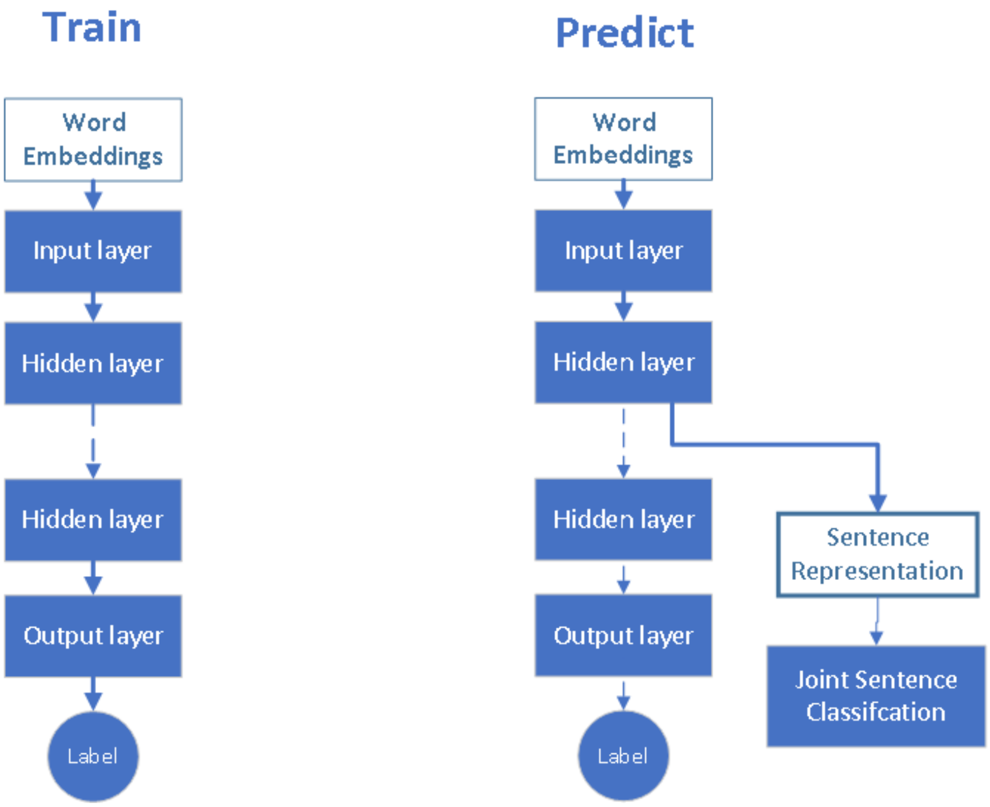
Data	|V|	Train	Valid.	Test
Abstract	68k	15k	2.5k	2.5k
Sentence		180k	30k	30k

- Reproduce best result reported by Dernoncourt and Lee (bi-LSTM, MLP)
- Extend model by adding NLP features
  - POS tags (spaCy)
  - Constituent parse trees (Stanford CoreNLP)
- Compare

Benchmark

1. Perform sentence classification using word embeddings on LSTM followed by MLP neural net.
2. Perform prediction using Step 1 model and use output from hidden layer to create sentence representations.
3. Perform joint sentence classification using sentence representations obtained from Step 2 on bi-LSTM neural net.

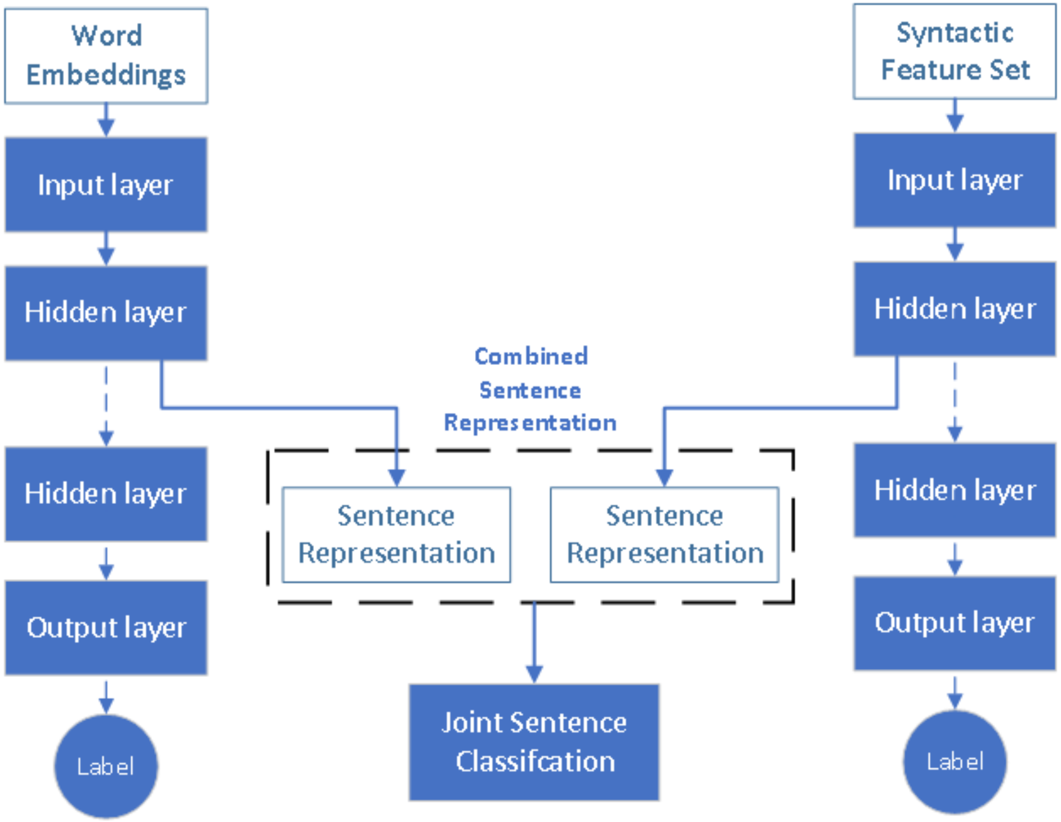
Sentence representation



Extended model

1. Perform sentence classification using word embeddings on LSTM followed by MLP neural net.
2. Perform sentence classification using onehotencoded NLP syntactic features extracted from text.
3. Perform predictions for Steps 1 and 2 and use outputs from hidden layers to create combined sentence representations.
4. Perform joint sentence classification using combined sentence representations obtained from Step 3.

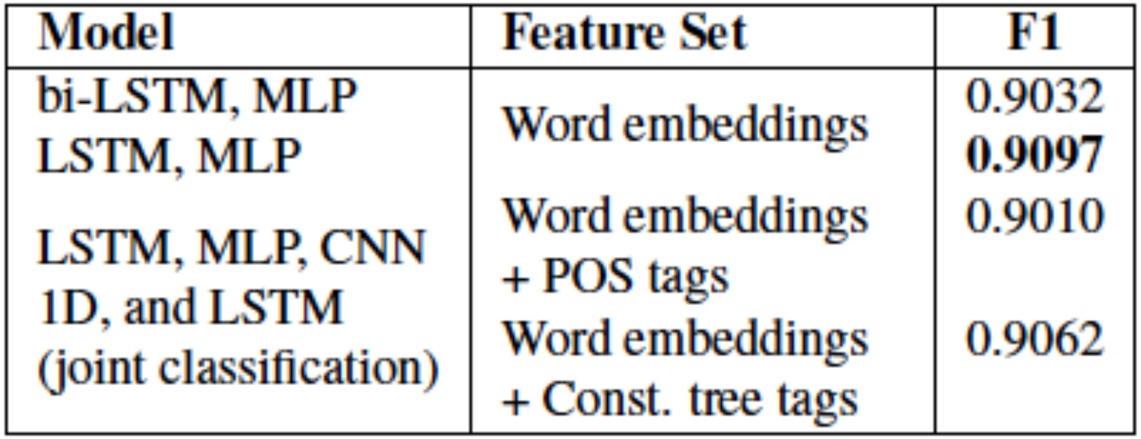
Combined sentence representation



Experiments

- Focus on joint classification after F1-score = 0.90 was achieved
- Type of network for joint sentence classification: LSTM, bi-LSTM, CNN 1D
- Combination of vector representations

Results



Model	Feature Set	F1
bi-LSTM, MLP	Word embeddings	0.9032
LSTM, MLP		**0.9097**
LSTM, MLP, CNN 1D, and LSTM (joint classification)	Word embeddings + POS tags	0.9010
	Word embeddings + Const. tree tags	0.9062

Conclusion

- POS tag is easy to extract but parse tree tags are hard to extract - time consuming!
- No benefit found. New features slightly degraded performance of models using only word embeddings.
- Do not recommend further research on benefits of incorporating POS tag and parse tree tag for sentence classification using this dataset.

Questions?

Email csathler@iu.edu or cssathler@gmail.com