Contextual explanation rules for neural clinical classifiers

Madhumita Sushil, Simon Šuster, Walter Daelemans
Computational Linguistics and Psycholinguistics Research Center, University of Antwerp, Belgium
madhumita.sushil.k@gmail.com

LIMITATIONS OF EXISTING APPROACHES FOR EXPLANATION RULE LEARNING

Explanations over unigrams

Conjunctive explanation rules over unigrams **lose sequential information**.

Input Embeddings not supported

Rule-based explanation pipelines for high-dimensional NLP tasks **do not support embeddings as inputs**.

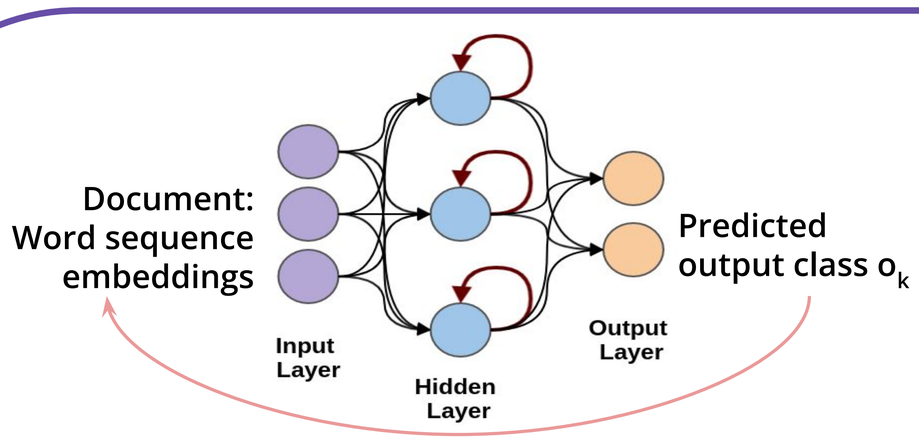
Hierarchical explanations

Some rule-based explanation methods **output hierarchical explanations**.

RESEARCH QUESTION

How can we induce explanation rules that encode phrase-level sequential semantics?

PROPOSED TECHNIQUE TO EXPLAIN RNNs



1. **Input saliency,** $G = \frac{\partial o_k}{\partial I}$
2. **Compute word importance =** ***dot(I,G)***
3. **Compute skipgram importance =** ***mean(word_imp)***
4. **Retain the most important skipgrams**

   no *signs* ***of infection*** ***found*** *.* document1, class *non-septic*

   infection *is* ***positive***, ***found*** *evidence .* document2, class *septic*
5. **Discretize skipgram importance**

   ++ High positive impact on output probability
   + Low positive impact on output probability
   −− High negative impact on output probability
   − Low negative impact on output probability
   ⊘ Absent in the input sequence
6. **Rules as explanations**

if *no of infection* is ++ and *found* is - then *septic*
else: *non-septic*

DATASETS FOR EVALUATION

Synthetic dataset:

17 sentences per document sampled from MIMIC-III
- Containing an ***infection_term***
- Containing a ***measurement_term***
- Containing neither of the terms

Synthetic labeling rule:
- If ***infection_term*** is not negated *and* min two ***measurement_terms*** are not negated:
  - Class *septic* 49%
  - Class *non-septic* otherwise

Real datasets: MIMIC-III sepsis classification using:
1) discharge note 2) last non-discharge note

RESULTS - EXPLANATION ACCURACY %

	Synthetic	+Discharge	-Discharge
Classification F1	.97	.68	.60
Baseline Fidelity* F1	.76	.62	Did not converge
Baseline num rules	63	825	NA
UNRAVEL Fidelity F1	.99	.98	.77
UNRAVEL num rules	32	16	196

*Rules trained directly from the original input

RESULTS - EXAMPLE EXPLANATION RULES

hyperglycemia = ++ AND
evidence infection . = ⊘ AND
no infection . = ⊘ AND
negative infection = ⊘ AND
fungal infection other = ⊘ AND
altered = ++
→ **septic** (✓17466/17466)

to exclude = AND
infection = ++ AND
no infection = ⊘ AND
or of infection = ⊘ AND
of infection in the = ⊘ AND

Synthetic Dataset

sepsis major surgical = ++ → ***septic*** (✓209/209)

complaint : sepsis = ⊘ AND
chief hypotension major = ++
→ ***septic*** (✓169/169)

MIMIC-III Dataset (+discharge)

https://github.com/clips/rnn_expl_rules