Leveraging Deep Neural Networks And Semantic Similarity Measures For Medical Concept Normalization In User Reviews

Miftahutdinov Z.Sh.
Tutubalina E.V.
Kazan Federal University

1 june 2018

Medical concept normalization – mapping a disease mention to a concept in a controlled vocabulary.

Examples:

inflammation in my neck –> C0263854, Cervical arthritis

very painful joints –> C3864084, Arthralgia

can't sleep –> C0393758, Insomnia

high BP –> C0020538, Increased venous pressure

Problem Description

Possible Bottlenecks

- Social-media language
- Ambiguity
- Vocabulary variations
- Abbreviations
- Variety of target vocabularies

MetaMap

- mapping to UMLS
- rule based

DNorm

- mapping to MEDIC
- pairwise learning to rank

Dataset

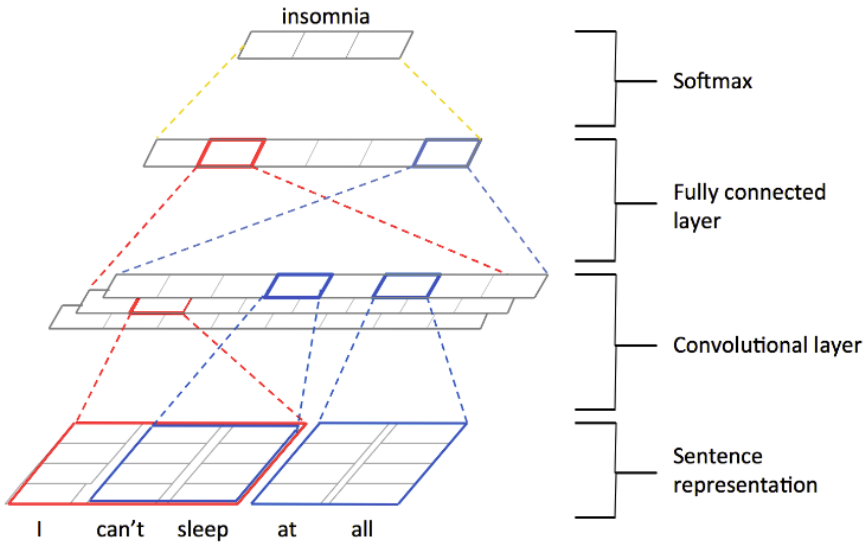
- Twitter
- mapping to MedDRA concepts
- Training set 6650 phrases, 472 concepts
- Test set 2500 phrases, 254 concepts

Team Results

Team	Model	Accuracy (%)
gnTeam	Logistic Regression	87.7
	Bi-GRU	85.5
	Ensemble	88.5
UKNLP	Hierarchical Char-LSTM	87.2
	Hierarchical Char-CNN	87.7

Background

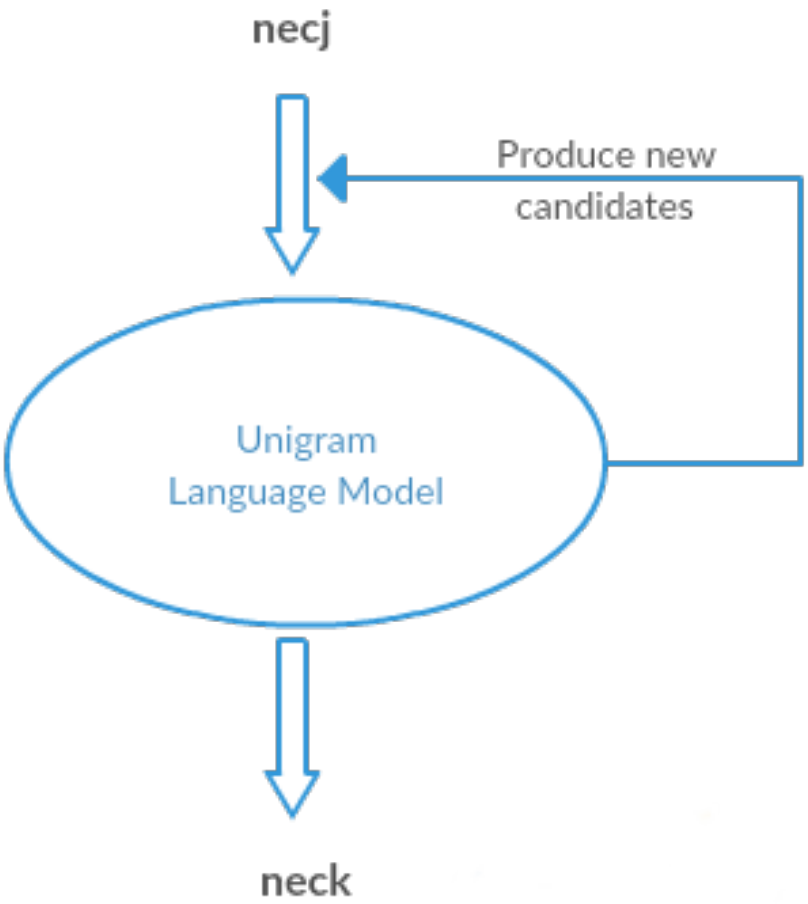
Limsopatham and Collier



- Data from askapatient
- mapping to SNOMED
- 8411 phrases
- 1029 unique codes
- 81% Accuracy

Proposed Model

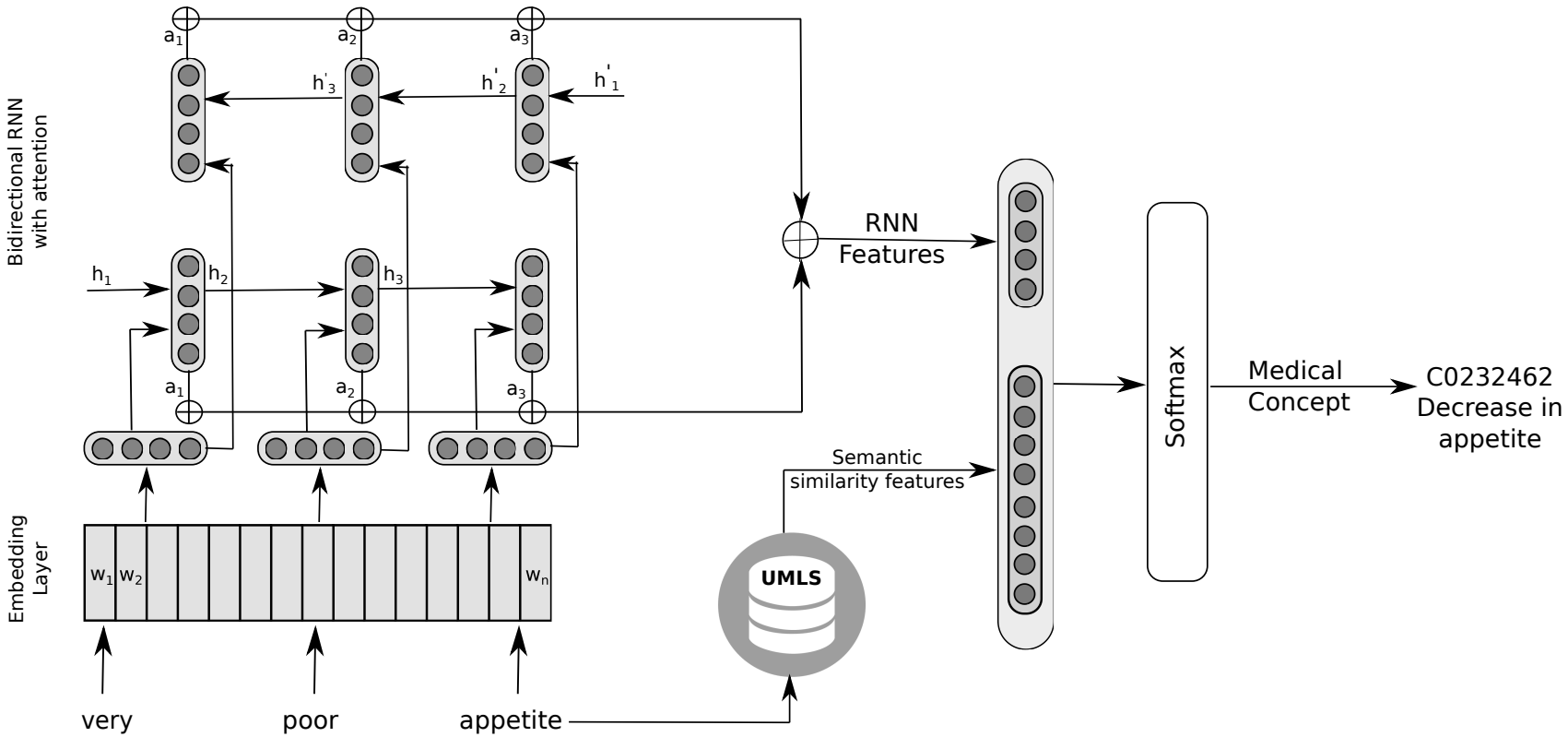
Spell checking



- Levenshtein distance
- Up to 3 misspellings

Proposed Model

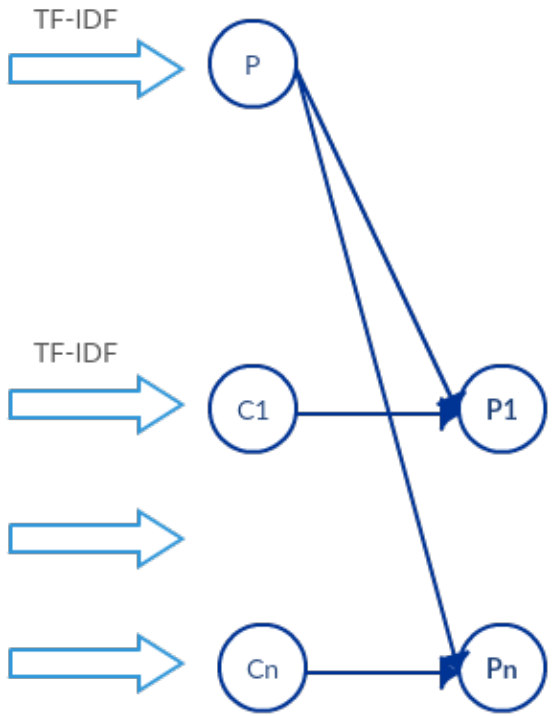
Neural Network Architecture



Proposed Model

Semantic Similarity Features

Concept ID	Concept Description
C0000727	abdomen acute syndrome abdominal acute
...	...
C3531965	viral pericarditis with pericardial effusion



Results

Model	Accuracy (%)
DNorm	73.39
CNN	81.41
RNN	79.98
GRU+At., TFIDF	85.71
New Folds	
CNN	46.19
LSTM	64.51
GRU	63.05
LSTM + Attention	65.73
GRU + Attention	67.08
LSTM + Attn, TFIDF	67.63
GRU + Attn, TFIDF	69.92

Limitations and Future Work

- SMM4H dataset
- Take context into account
- Ranking algorithms