Normalising Medical Concepts from Personal Health Messages in Social Media

UNIVERSITY OF CAMBRIDGE

Language Technology Lab

Nut Limsopatham and Nigel Collier
nl347@cam.ac.uk, nhc30@cam.ac.uk

Overview

Text normalisation that can recognise when laymen's terms in social media messages refer to a particular medical concept may lead to automatic monitoring and inferences on community health conditions. To deal with discrepancy between the type of language used in social media and medical ontologies, we investigate the use of machine translation and neural networks to learn the transition between the language in social media and medical ontologies.

1. Introduction

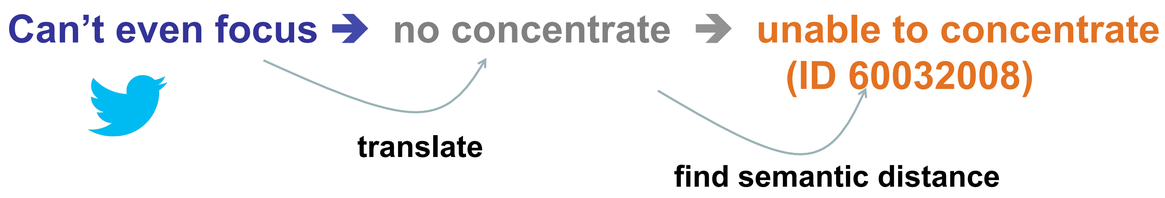
Aim: "Automatically identify mentions of medical conditions, in terms of medical concepts, in social media messages"

Social media message	SNOMED ID	Description
No way I'm getting any sleep 2nite	193462001	Insomnia
Still tired as shit	84229001	Fatigue
Can't even focus forreal	60032008	Unable to concentrate
DRUG makes u skinny	89362005	Weight loss

Table 1: Examples of mappings from social media messages to SNOMED-CT

2. Phrase-based Machine Translation (P-MT) [2]

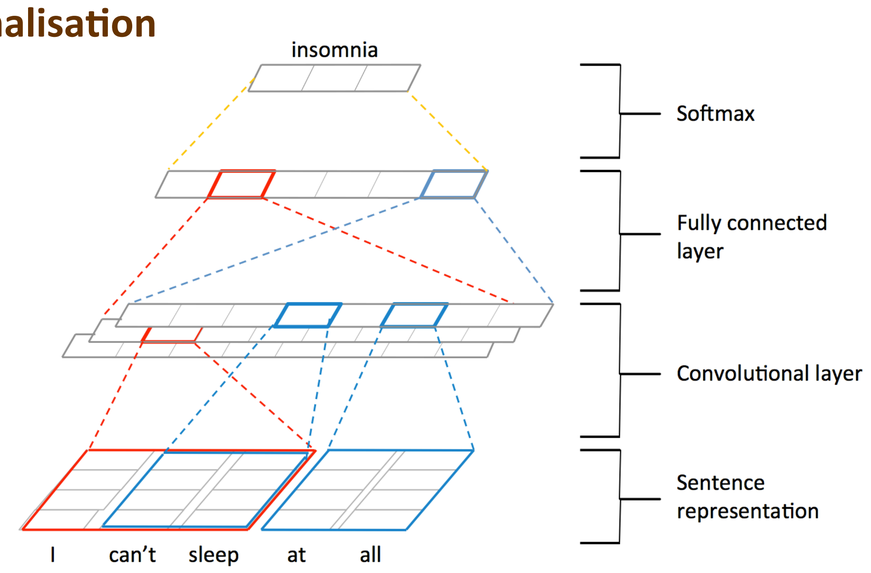
- We use phrase-based MT to **translate social media text to formal medical text**, then map the translated text to a medical concept



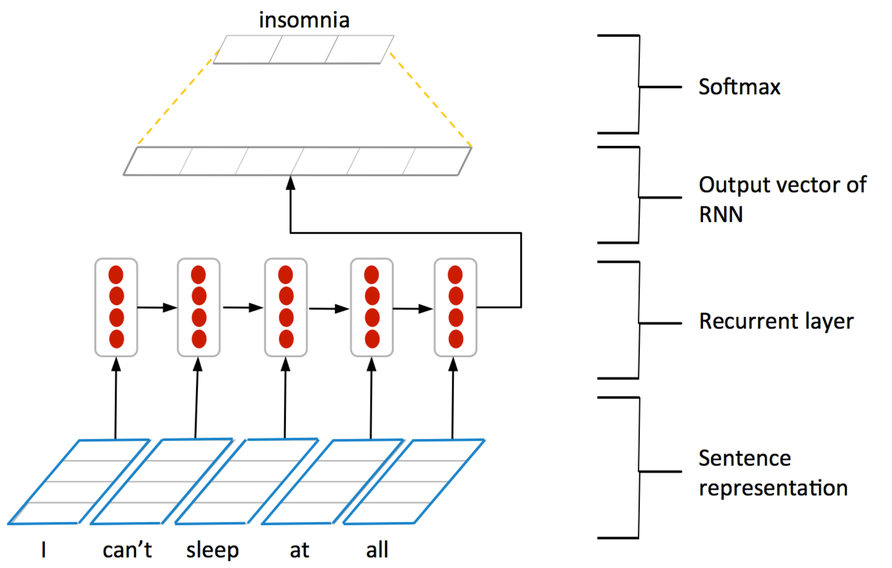
3. DNNs for Medical Concept Normalisation [3]

- *Hypothesis: "Deep neural networks could capture semantic similarity between texts from social media messages and medical concepts "*

➢ Convolutional Neural Networks (CNN) for Concept Normalisation



➢ Recurrent Neural Networks (RNN) for Concept Normalisation



4. Experimental Setup

- For evaluation, we use 3 different test collections, which contain social media texts from Twitter and AskAPatient.com [1]

Dataset	# Queries	# Target concepts	Data source
TwADR-S	201	58	Twitter messages
TwADR-L	1,436	2,220	Twitter messages
AskAPatient	8,662	1,036	Posts from askapatient.com

Table 2: Information regarding the three used datasets

- The datasets are publically available and can be downloaded from **https://doi.org/10.5281/zenodo.55013**
- We evaluate the proposed approaches based on the **accuracy** measure, using **10-fold cross-validation**

Baselines:

- **BM25 [4]** – A term-matching-based approach
- **EmbSim** – Cosine similarity of word vector representations
- **DNorm [1]** – A learning-to-rank-based approach, which achieved SOTA performance on several concept normalisation tasks

5. Experimental Results

- DNorm, which is a SOTA concept normalisation system, is the most effective baseline.
- P-MT performed comparable to DNorm
- Both CNN and RNN markedly and consistently outperform all of the baseline for all the three datasets
  - ➢ CNN significantly outperforms DNorm by up to 44%

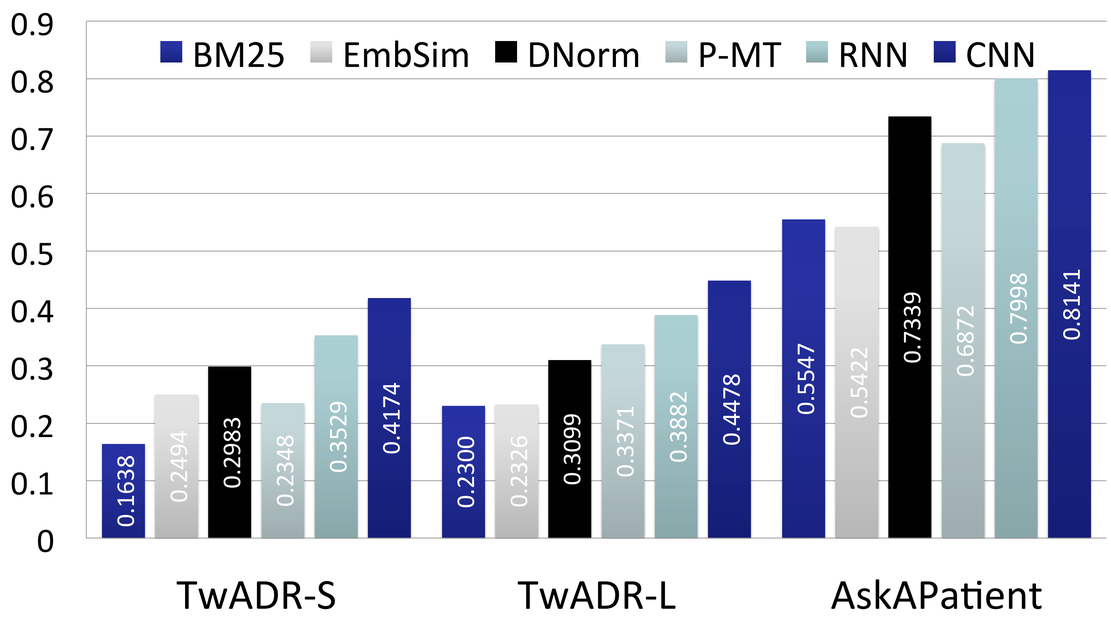


Figure 1: Accuracy performance on the three datasets

6. Conclusions

- Medical concept normalisation on social media texts is a difficult task, because of the discrepancy between the type of language used in social media and medical ontologies
- CNN and RNN can effectively capture semantics of social media texts, which help to bridge the discrepancy gap

References

[1] R. Leaman, R. Dogan, Z. Lu. *DNorm: disease name normalization with pairwise learning to rank*. Bioinformatics, 29(22), 2013.
[2] N. Limsopatham, N. Collier. *Adapting phrase-based machine translation to normalise medical terms in social media messages*. In EMNLP 2015.
[3[N. Limsopatham, N. Collier. *Normalising Medical Concepts in Social Media Texts by Learning Semantic Representations*. In ACL 2016.
[4] S. Robertson, H. Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc., 2009.