# BETTER WORD REPRESENTATIONS WITH RECURSIVE NEURAL NETWORKS FOR MORPHOLOGY


MATEUSZ ZAN
S1113967

# OUTLINE

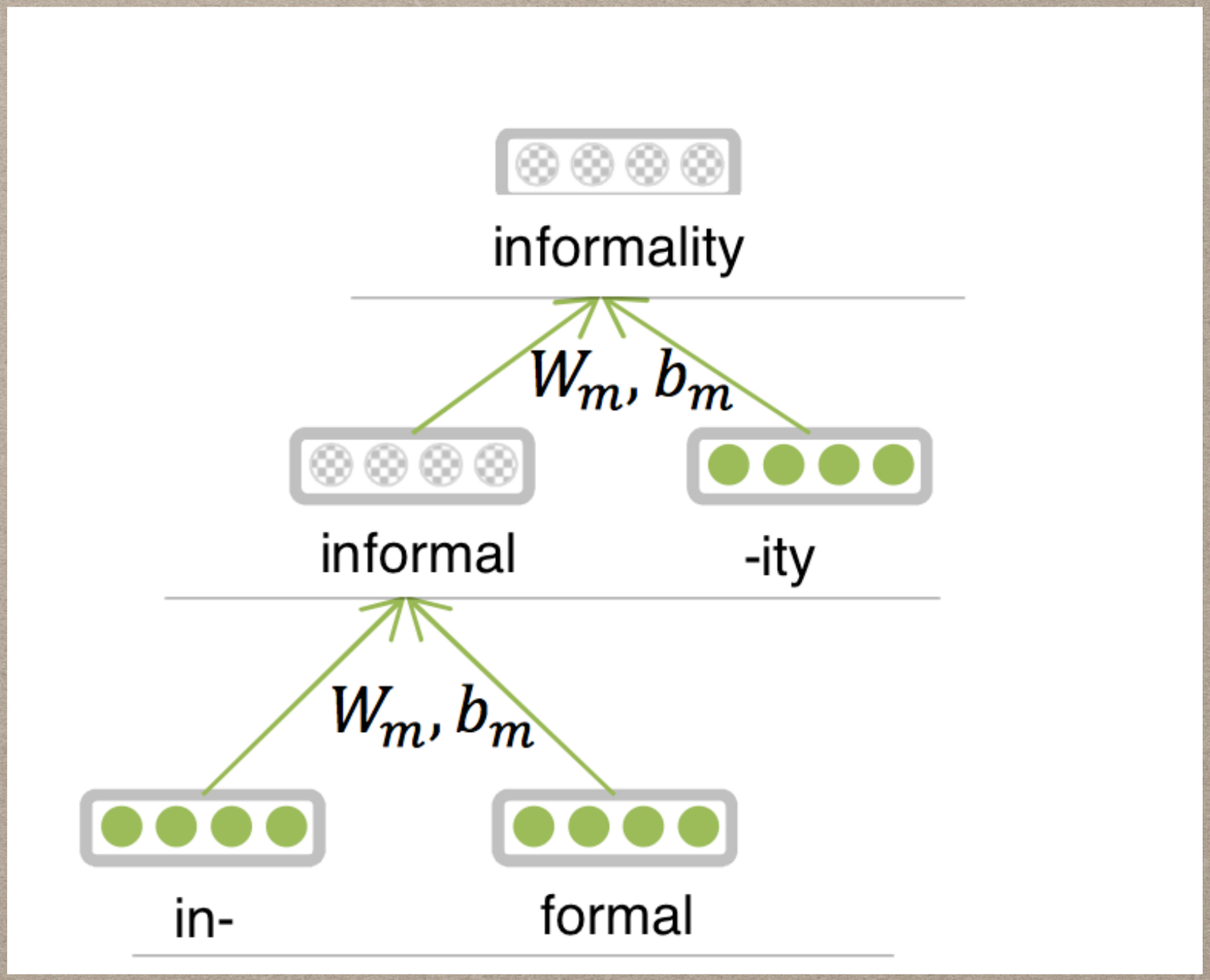
- Key points of interest
- Motivation
- Possible applications
- Methodology (2 different approaches)
- Experiment (Word Similarity Task)
- Results and Evaluation

# KEY POINTS OF INTEREST

- Morphology - study of patterns of word formation for a natural language
- Recurrent Neural Networks- statistical learning algorithm used in many NLP tasks
- Using RNNs, we aim to model morphology- be able to build better word representations for complex words, gradually building more complex words using morphemes

# MOTIVATION

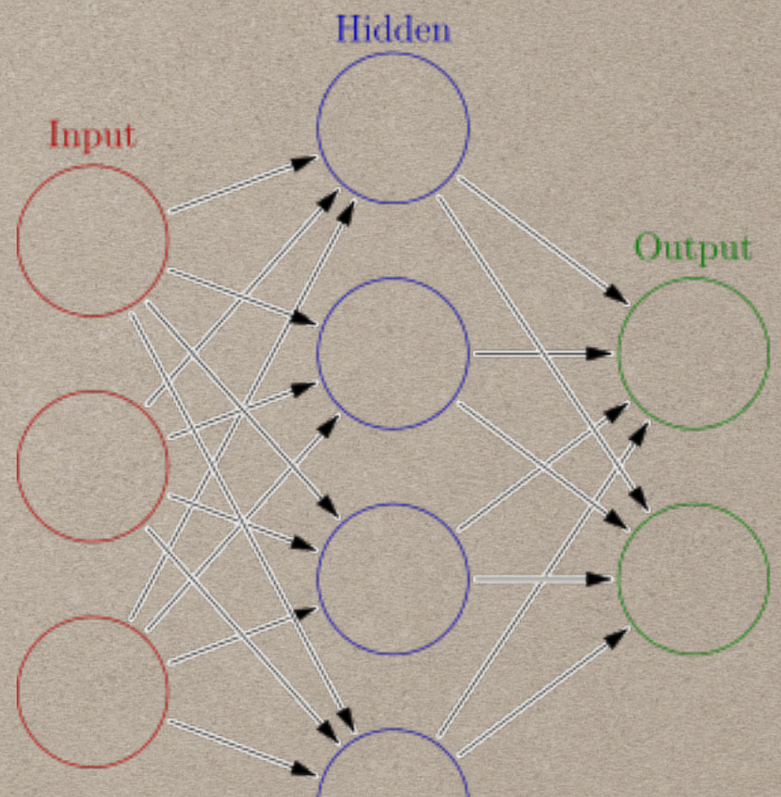
- *'formal'*
- *'in-formal-ity'*



# POSSIBLE APPLICATIONS

- POS Tagging- certain suffixes correspond to different parts of speech
- Parsing (Morphological parsing)
- semantic role labelling - breaking up words into smaller parts can help with labelling semantics

# METHODOLOGY




- Quick recap of neural networks!
- In this problem, we use a variant of Neural Network: **Morphological Recursive Neural Network**
- The Morphological Recursive Neural Network works on morpheme level, rather than on the word level

# RNN MODEL

- Morphemes encoded as vectors (dimension d) in an embedding matrix **We** (of size **d x |M|**, where **M** is set of all morphemes)
- Words are built gradually, by combing morphological parts. Parent vector p is constructed by combining stem vector with affix vector

$$p = f(W_m[x_{\text{stem}}; x_{\text{affix}}] + b_m)$$

- We have a model θ = **{We,Wm,bm}** and we want to learn the parameters

# TWO DIFFERENT APPROACHES

- Two approaches of the Neural Network implementation considered:
- Context-insensitive Morphological RNN
- Context-sensitive Morphological RNN
- The difference between the two is that the latter considers the contextual information (the other words in a sentence, etc.), while the other one doesn’t use that information

# CONTEXT-INSENSITIVE MODEL

- Model considers how words can be constructed simply from morphemic representation
- Given the reference words, the goal is to construct new words to match the reference as closely as possible
- Structure is the same as a basic RNN model, explained above
- For learning, a cost function s is defined measuring the Euclidean distance between output and reference vector
- Objective function we try to minimise is the sum of cost for N words:

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} s\left(x_i\right) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

# CONTEXT-SENSITIVE MODEL

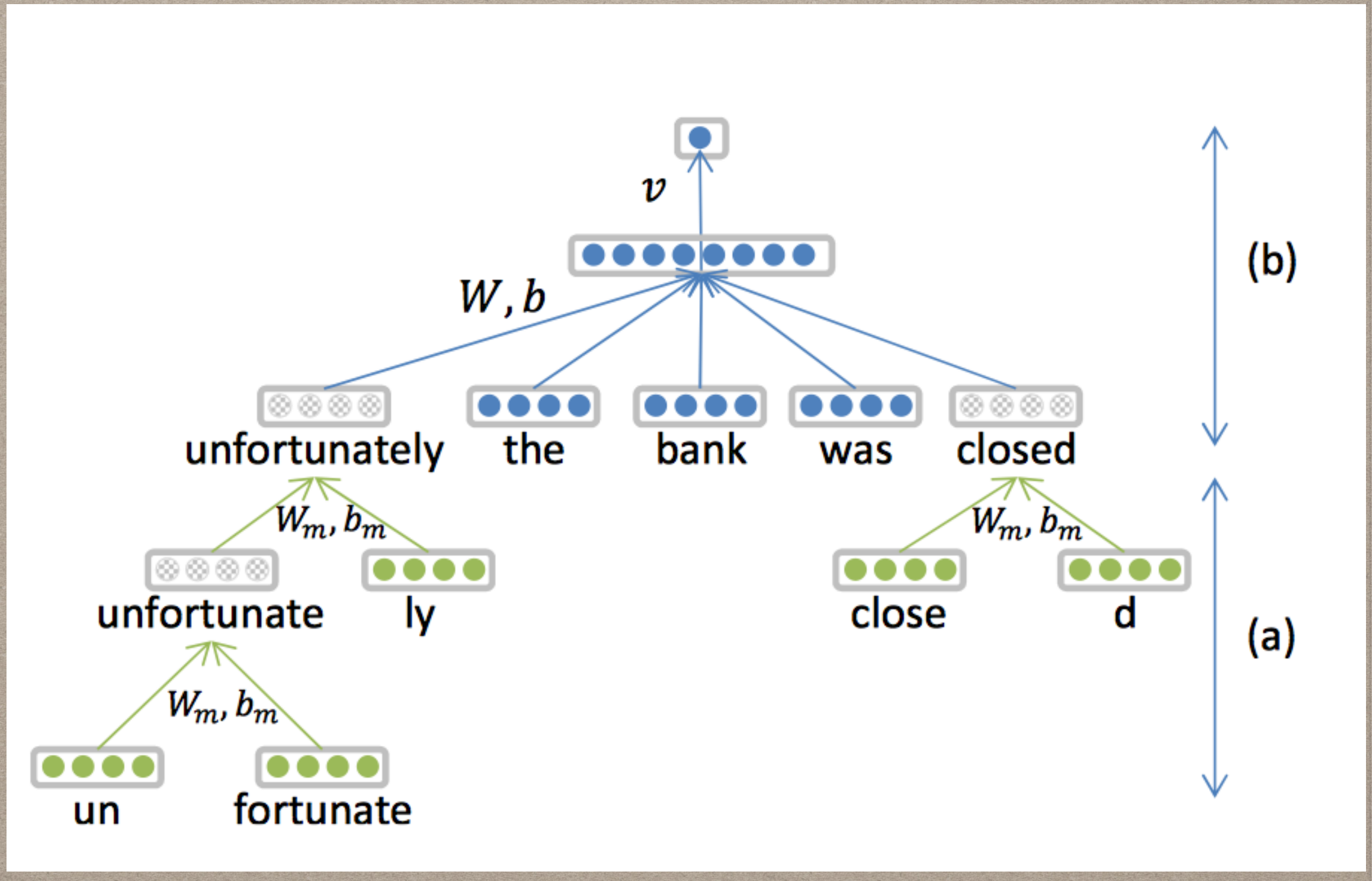
- Tries to address limitations of the previous model by considering the context in which the word appears
- 2 layers- MorphoRNN and word-based neural language model
- n-grams scored using formula

$$s(n_i) = \boldsymbol{v}^\top f(\boldsymbol{W}[\boldsymbol{x}_1; \ldots; \boldsymbol{x}_n] + \boldsymbol{b})$$

- Objective function to optimize parameters

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} \max\{0, 1 - s(n_i) + s(\overline{n}_i)\}$$

- Model parameters are θ = {We;Wm;bm;W;b;}.


v
W, b
unfortunately
the
bank
was
closed
(b)
W_m, b_m
unfortunate
ly
close
d
(a)
W_m, b_m
un
fortunate

# PARAMETER OPTIMISATION (LEARNING)

- Algorithm considers two passes: **forward-pass** and **backward-pass**
- For the latter, we are interested in minimising objective function, to optimise parameters (back-propagation)
- cimRNN

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{N} \frac{\partial s\left(x_i\right)}{\partial \boldsymbol{\theta}} + \lambda \boldsymbol{\theta}$$

- csmRNN

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i:1-s(n_i)+s(\overline{n}_i)>0} -\frac{\partial s\left(n_i\right)}{\partial \boldsymbol{\theta}} + \frac{\partial s\left(\overline{n}_i\right)}{\partial \boldsymbol{\theta}}$$

# COLLECTION OF AFFIXES- MORFESSOR

- Morfessor- morphological segmentation toolkit
- For this investigation, we assume input of form pre* stm suf*
- Words in data are split using the toolkit, and affixes are stored, presented in a table

| **Prefixes** | **Suffixes** |
|---|---|
| 0 *al* all *anti auto co* *counter cross de dis* electro end *ex* first five focus four half high hy- per ill *im in inter ir* jan jean long low market mc micro mid *multi* neuro newly no *non* off one *over post pre pro re* sec- ond *self semi* seven short six state *sub super* third three top *trans* two *un* *under* uni well | *able al ally* american *ance* *ate ation* backed bank based born controlled *d* dale down *ed en er es* field ford free *ful* general head *ia ian ible ic* in *ing isation* *ise ised ish ism ist ity ive* *ization ize ized izing* land led *less* ling listed *ly* made making man *ment ness* off on out owned related *s* ship shire style ton town up us ville wood |

# EXPERIMENTAL SETUP

- Two embeddings (C&W- Collobert, HSMN- Huang)
- Various datasets: WS353, MC, RG, SCWS*, RW- various datasets to avoid overfitting
- Rare Word dataset - RW

| | All words | Complex words |
|---|---|---|
| WS353 | 0 \| 0 / 9 / 87 / 341 | 0 \| 0 / 1 / 6 / 10 |
| MC | 0 \| 0 / 1 / 17 / 21 | 0 \| 0 / 0 / 0 / 0 |
| RG | 0 \| 0 / 4 / 22 / 22 | 0 \| 0 / 0 / 0 / 0 |
| SCWS* | 26 \| 2 / 140 / 472 / 1063 | 8 \| 2 / 22 / 44 / 45 |
| RW | 801 \| 41 / 676 / 719 / 714 | 621 \| 34 / 311 / 238 / 103 |

# WORD SIMILARITY TASK

- In this task, we compare similarity scores given by models and human annotators
- To measure relationship, Spearman's rank correlation is considered
- Results compared with human annotators rankings

# RESULTS AND EVALUATION

| | WS353 | MC | RG | SCWS* | RW |
|---|---|---|---|---|---|
| HSMN | 62.58 | 65.90 | 62.81 | 32.11 | 1.97 |
| +stem | 62.58 | 65.90 | 62.81 | 32.11 | 3.40 |
| +cimRNN | 62.81 | 65.90 | 62.81 | 32.97 | 14.85 |
| +csmRNN | **64.58** | **71.72** | **65.45** | **43.65** | **22.31** |
| C&W | 49.77 | 57.37 | 49.30 | 48.59 | 26.75 |
| +stem | 49.77 | 57.37 | 49.30 | **49.05** | 28.03 |
| +cimRNN | 51.76 | 57.37 | 49.30 | 47.00 | 33.24 |
| +csmRNN | **57.01** | **60.20** | **55.40** | 48.48 | **34.36** |

- HMSN gives much better performance for datasets with frequent words (WS353, MC, RG)
- C&W performs much better with the rare words datasets (SCWS*, RW)
- csmRNN outperforms cimRNN in every case!

# RESULT AND EVALUATION (CONTD.)

- Syntactically, cimRNN enforces structural agreement - ‘JJ-ness’ and ‘fearlessness’
- Considering semantics, words that share same stem are clustered together, not good!
- csmRNN model seems to have a good balance between the two features, mainly thanks to using the context information

# THANKS FOR ATTENTION!

- Any questions?