Understanding Hidden Memories of Recurrent Neural Networks

Yao Ming, Shaozu Cao, Ruixiang Zhang, Zhen Li, Yuanzhe Chen, Yangqiu Song, Huamin Qu.

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

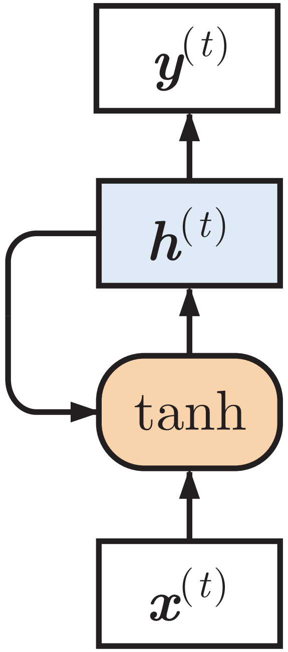
What is a Recurrent Neural Network?

Introduction

What is Recurrent Neural Networks (RNN)?

A deep learning model used for:


Machine Translation, Speech Recognition, Language Modeling, …



A vanilla RNN

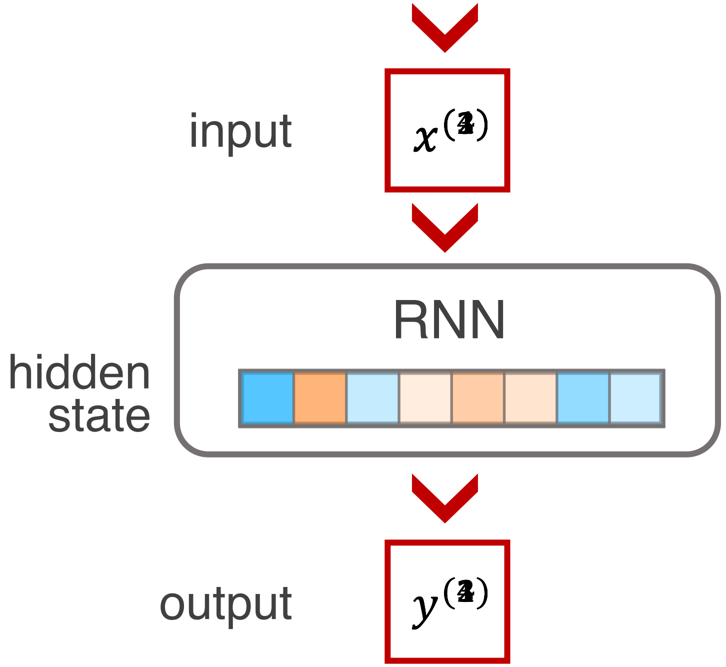
Introduction

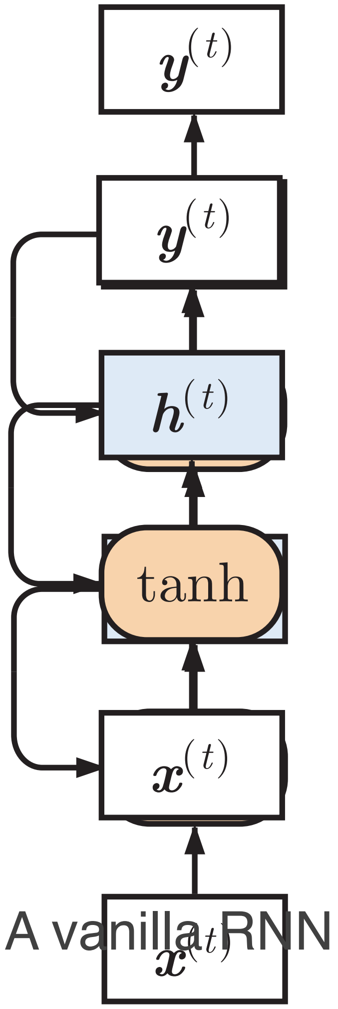
What is Recurrent Neural Networks (RNN)?

A vanilla RNN takes an input $\boldsymbol{x}^{(t)}$, and update its hidden state $\boldsymbol{h}^{(t-1)}$ using:

$$\boldsymbol{h}^{(t)} = \tanh\left(\boldsymbol{W}\boldsymbol{h}^{(t-1)} + \boldsymbol{V}\boldsymbol{x}^{(t)}\right)$$

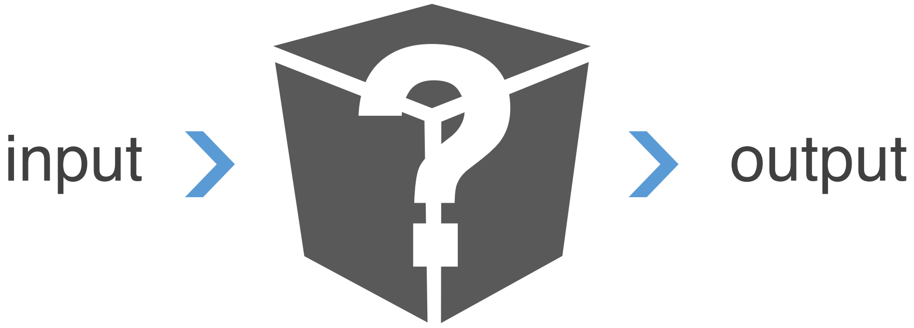
Visual Analytics Science & Technology





A vanilla RNN

A 2-layer RNN

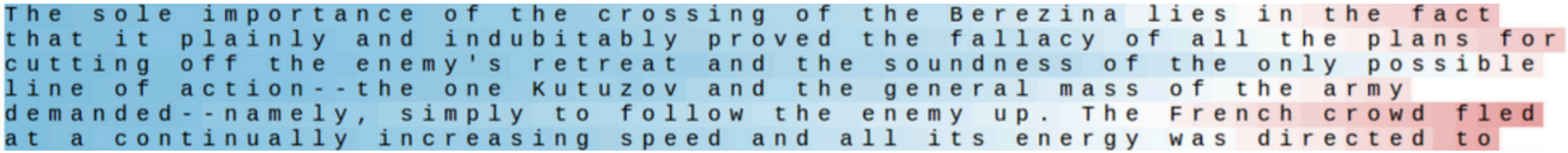


What has the RNN learned from data?

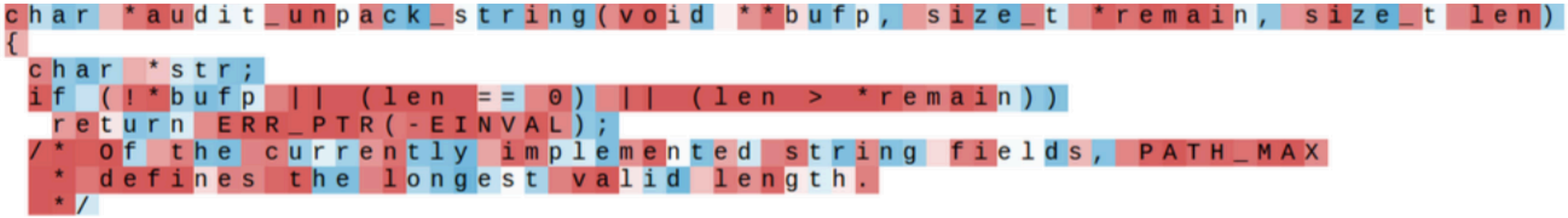
Motivation

What has the RNN learned from data?

A. map the value of **a single hidden unit** on data (Karpathy A. et al., 2015)



A unit sensitive to position in a line.

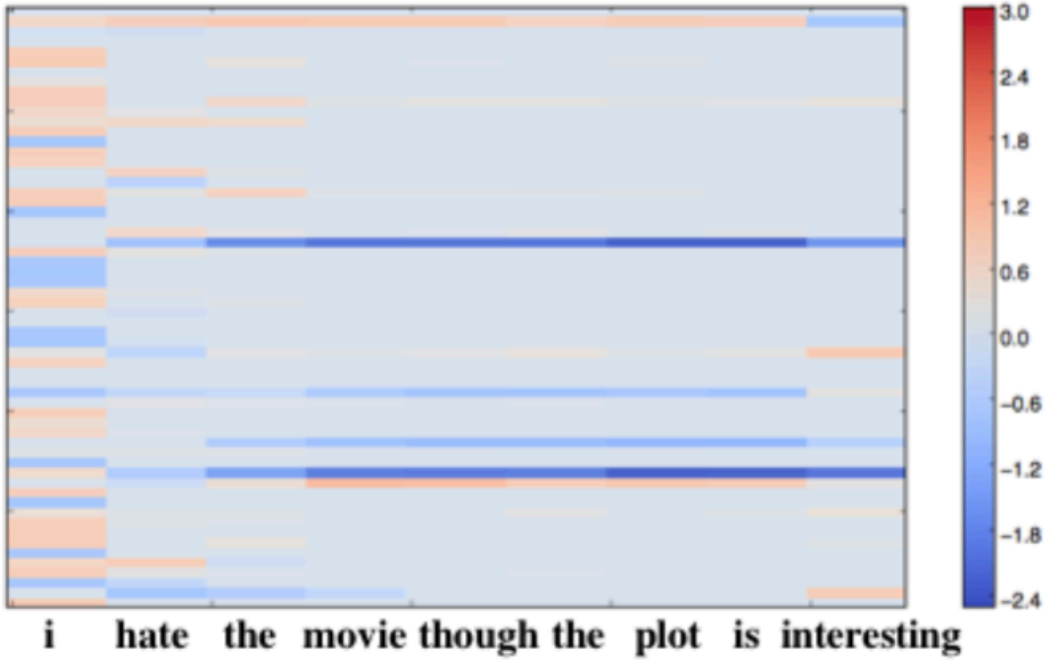


A lot more units have no clear meanings.

Motivation

What has the RNN learned from data?

B. matrix plots (Li J. et. al., 2016)



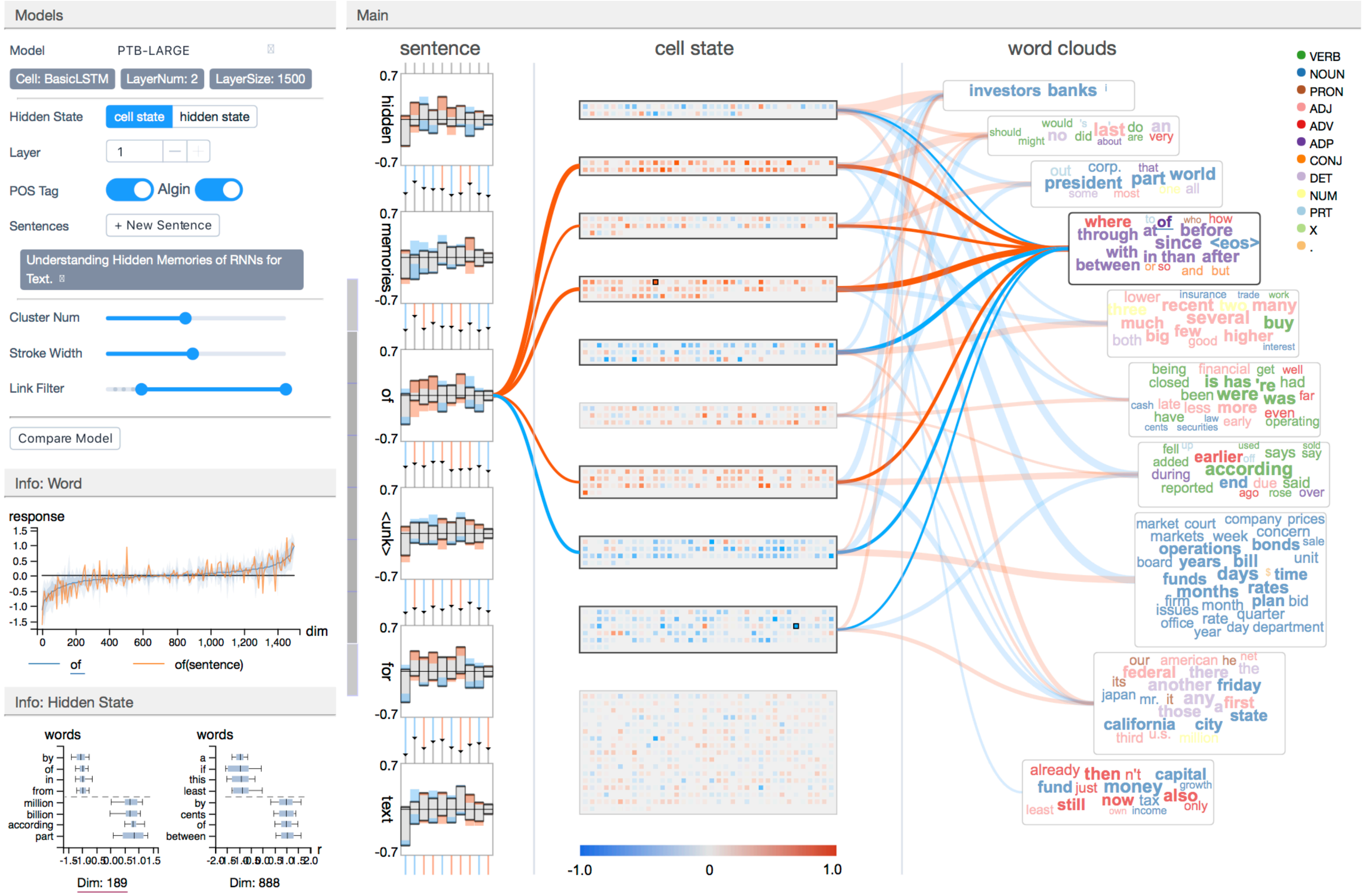
Each column represents the value of the hidden state vector when reads a input word

Scalability!

Machine Translation: 4-layer, 1000 units/layer (Sutskever I. et al., 2014)

Language Modeling: 2-layer, 1500 units/layer (Zaremba et al., 2015)

Our Solution - RNNVis



Our Solution

Explaining individual hidden units

Bi-graph and co-clustering

Sequence evaluation

Solution

Explaining an individual hidden unit using its most salient words

How to define salient?

Model's response to a word w at step t: the update of hidden state $\Delta\boldsymbol{h}^{(t)}$

$$\Delta\boldsymbol{h}^{(t)} = \left[\Delta h_i^{(t)}\right], i = 1, \dots, n.$$

Larger $\mathrm{abs}(\Delta h_i^{(t)})$ implies that the word w is more salient to unit i.

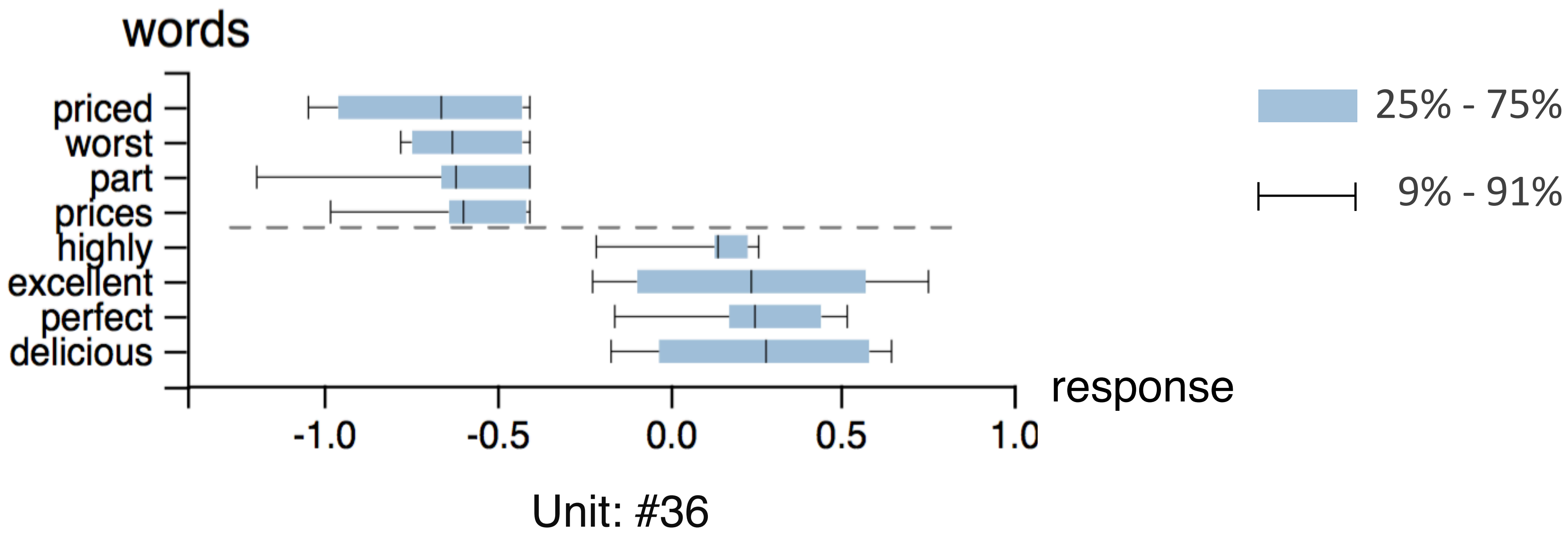
Since $\Delta h_i^{(t)}$ can vary given the same word w, we use the expectation:

$$\mathrm{E}\left(\Delta\boldsymbol{h}^{(t)} \mid w_t = w\right)$$

Can be estimated by running the model on dataset and take the mean.

Solution

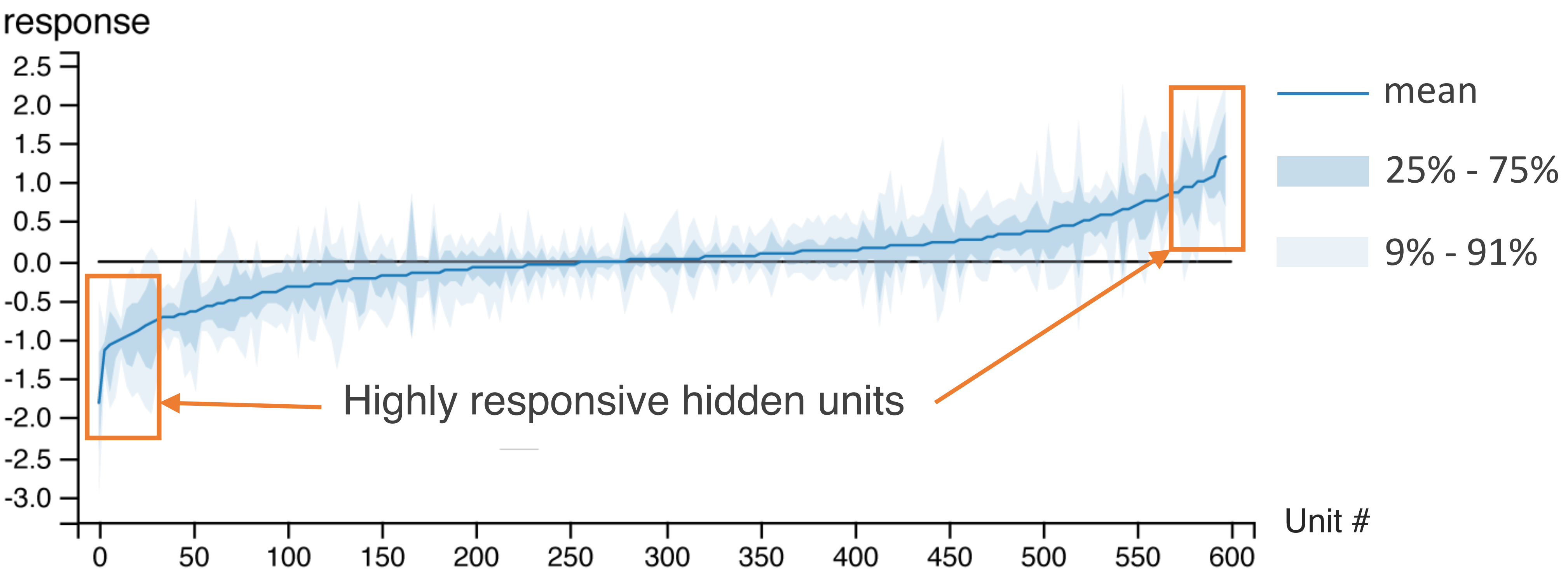
Explaining an individual hidden unit using its most salient words



Top 4 positive/negative salient words of unit 36 in an RNN (GRU) trained on Yelp review data.

Solution

Explaining an individual hidden unit using its most salient words



Distribution of model's response given the word "he".
Units reordered according to the mean. (an LSTM with 600 units)

Solution

Explaining an individual hidden unit using its most salient words

Investigating one unit/word at a time…

P: Too much user burden!

S: An overview for easier exploration

Solution

Explaining individual hidden units

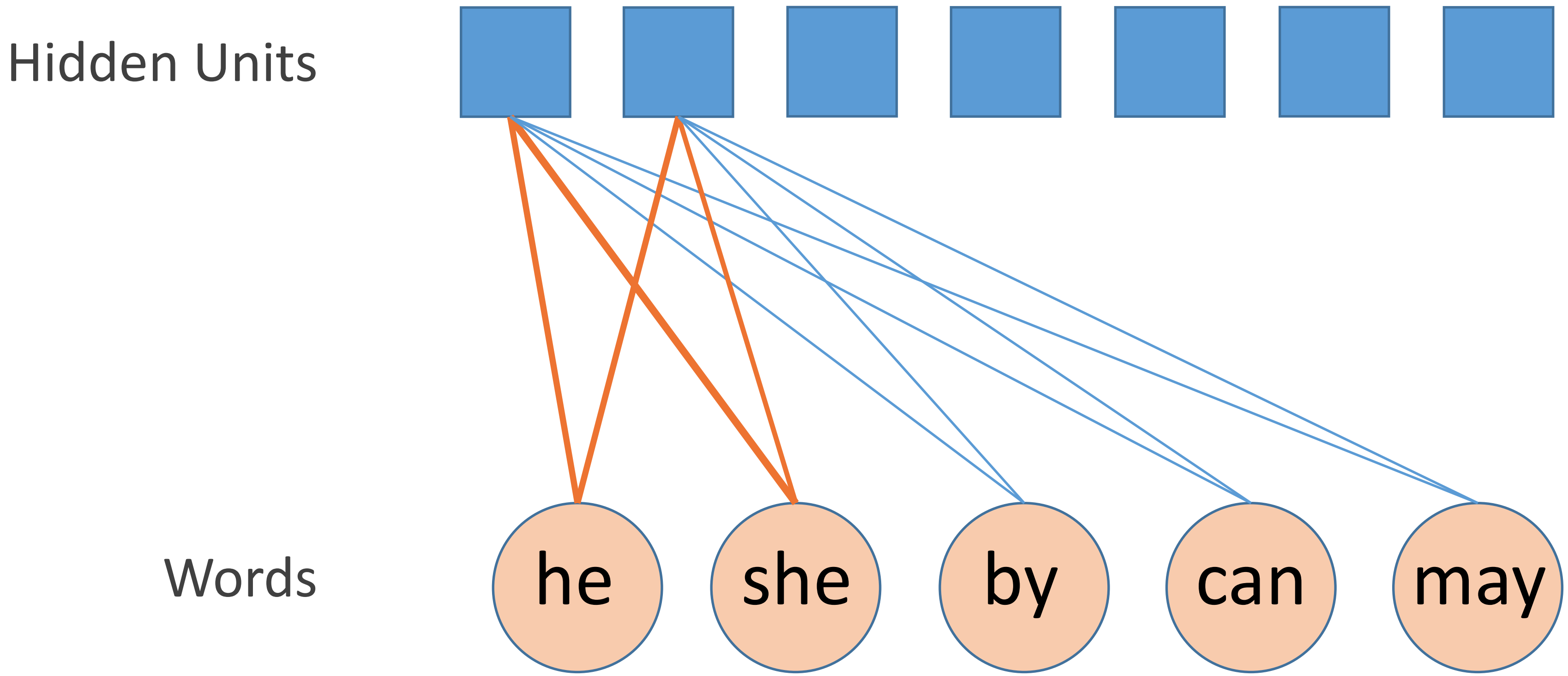
Bi-graph and co-clustering

Sequence evaluation

HKUST

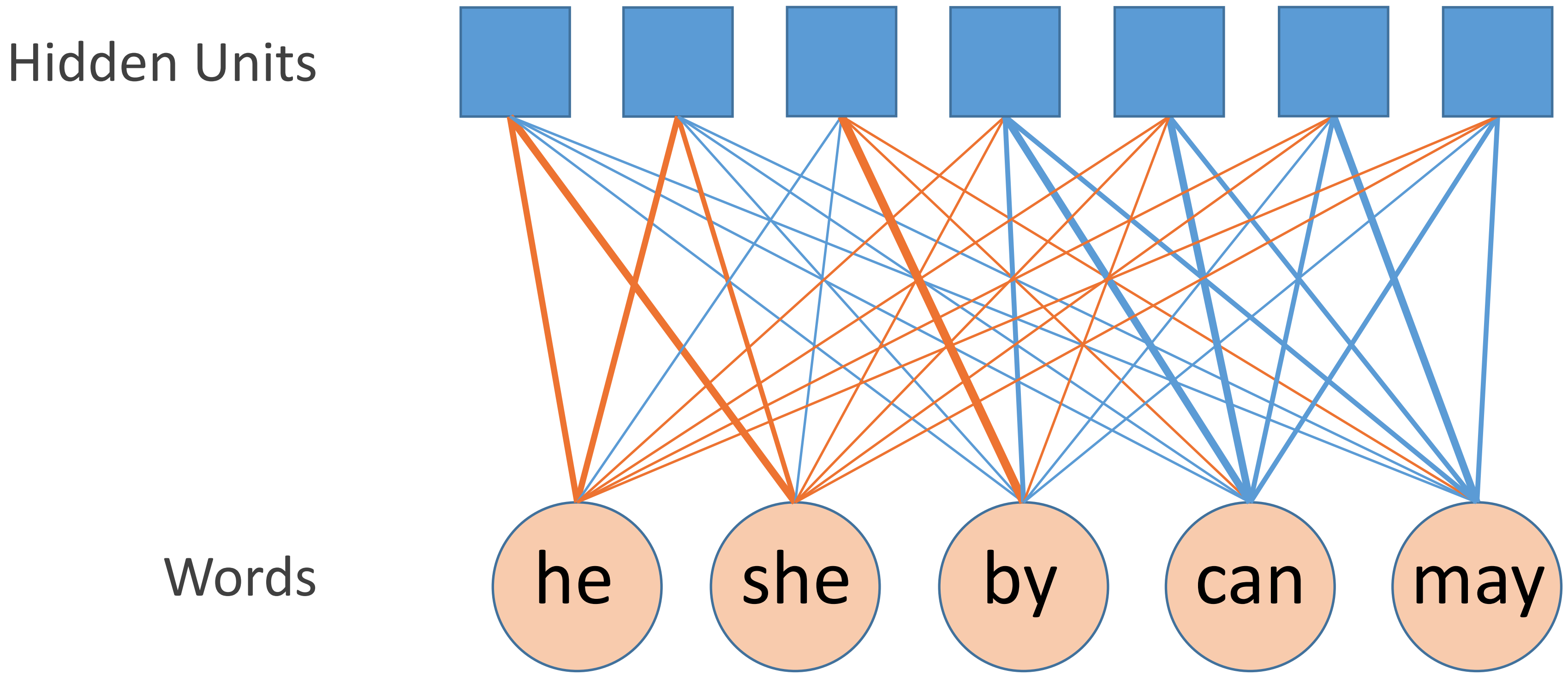
Solution

Bi-graph Formulation



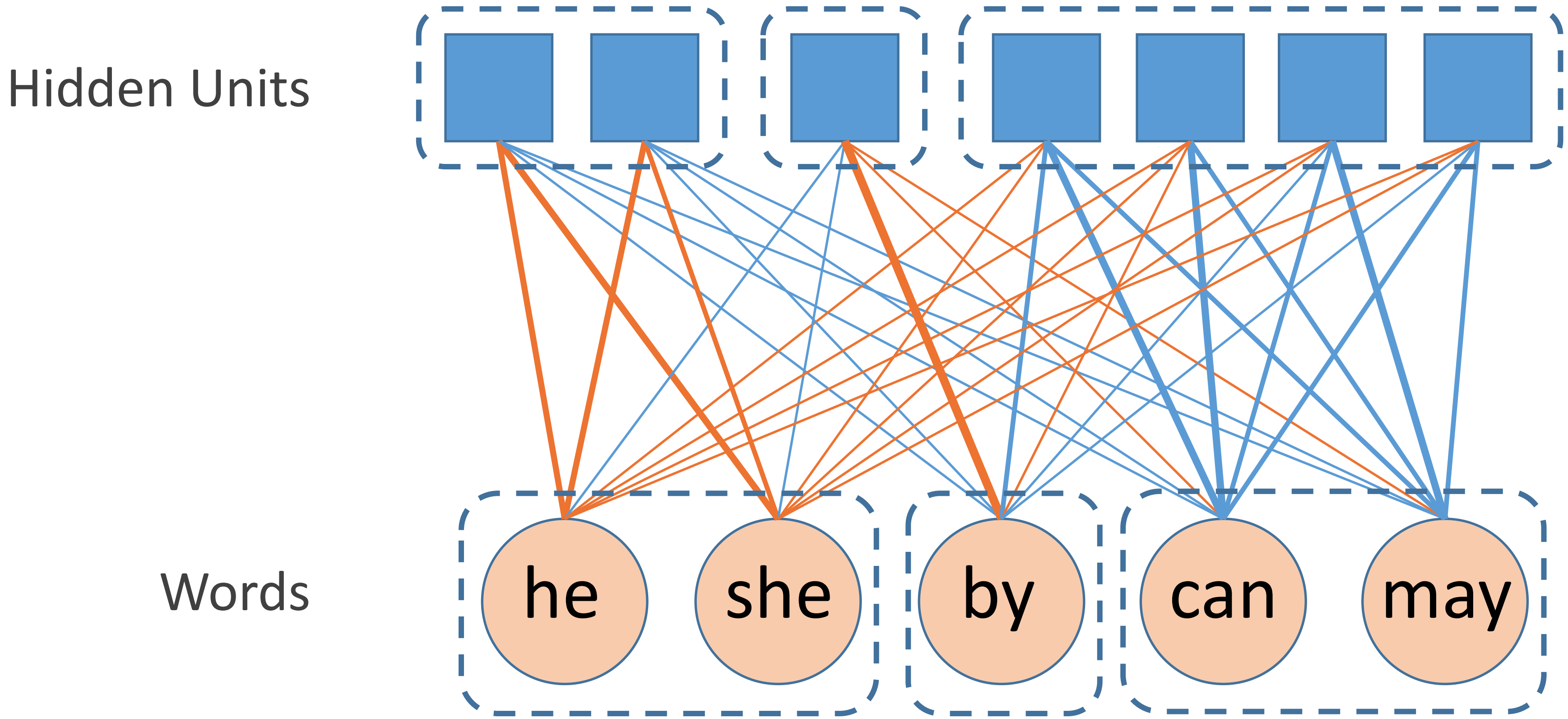
Solution

Bi-graph Formulation



Solution

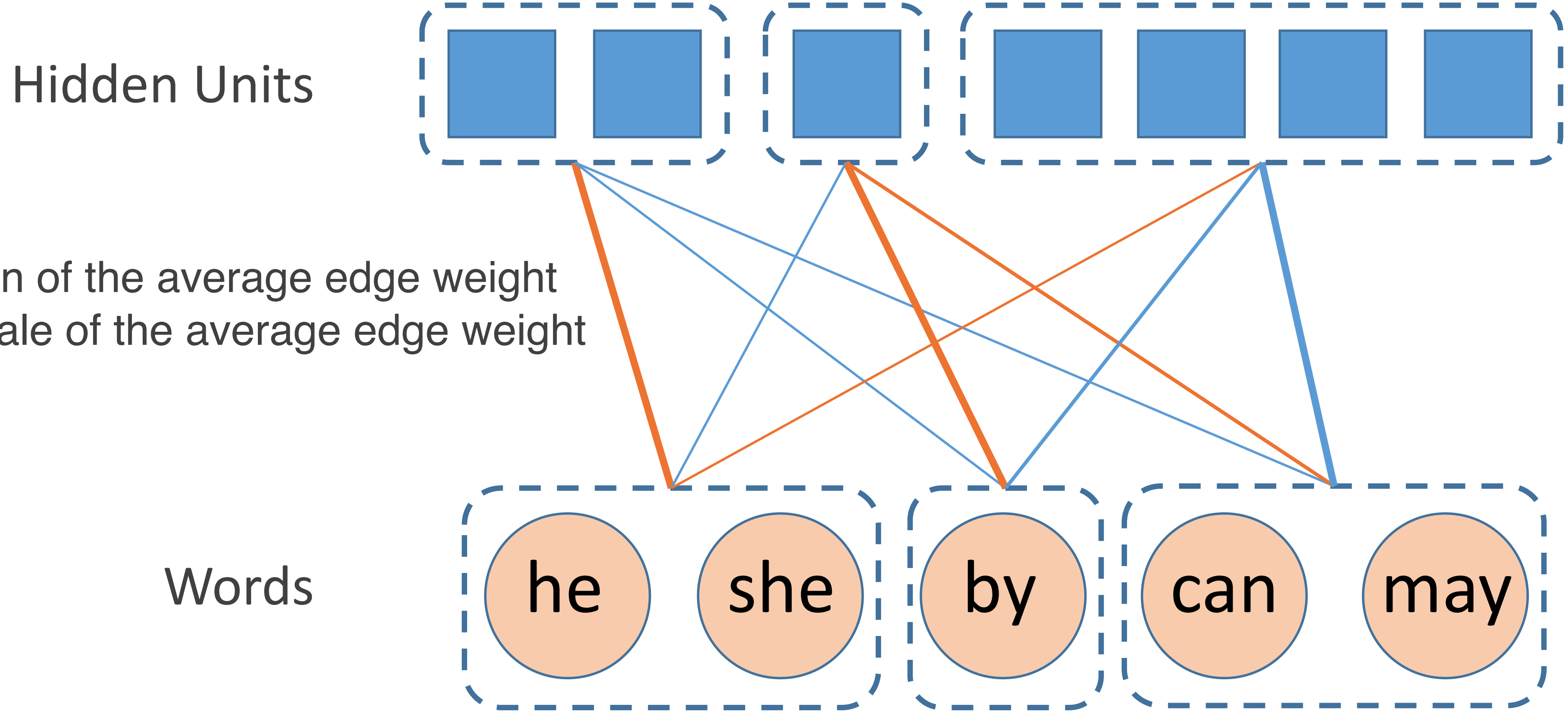
Co-clustering



Algorithm* Spectral co-clustering (Dhillon I. S., 2001)

Solution

Co-clustering – Edge Aggregation



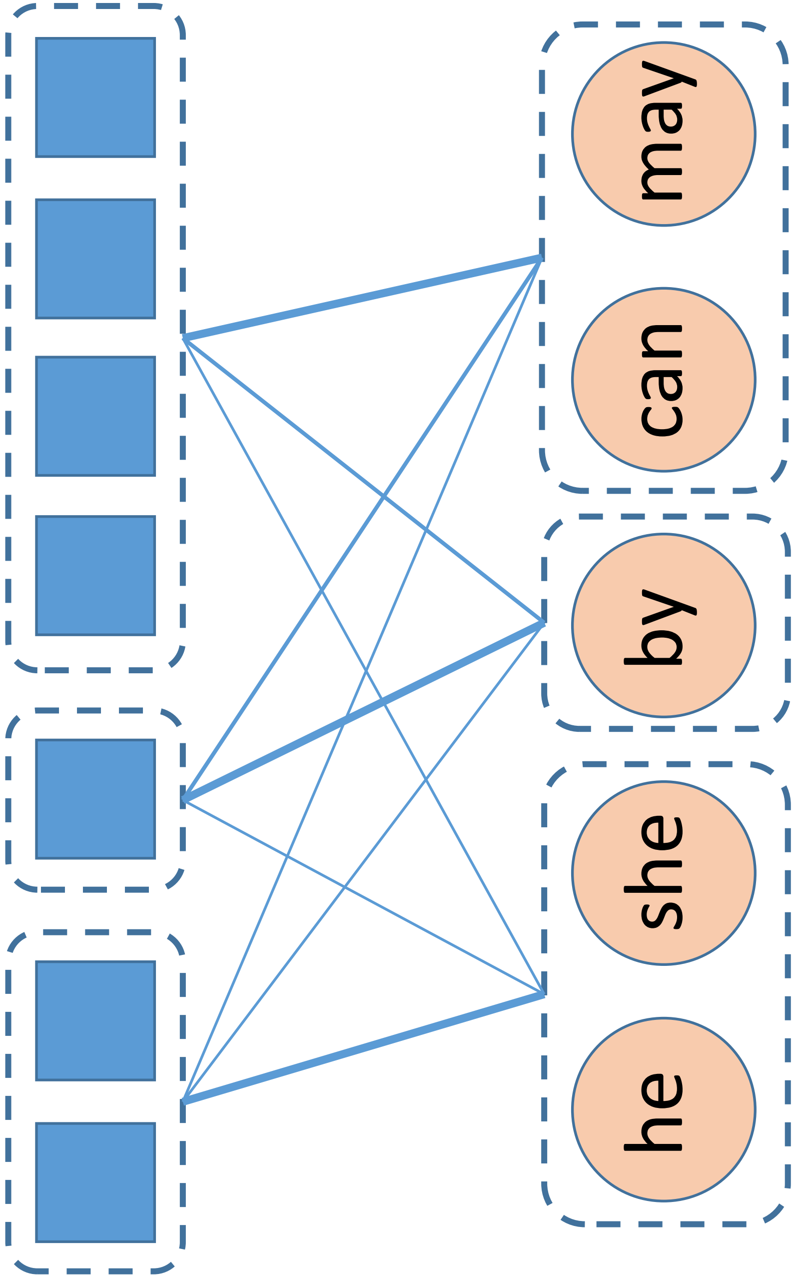
Color: sign of the average edge weight
Width: scale of the average edge weight

Solution

Co-clustering - Visualization

Hidden Units

Words



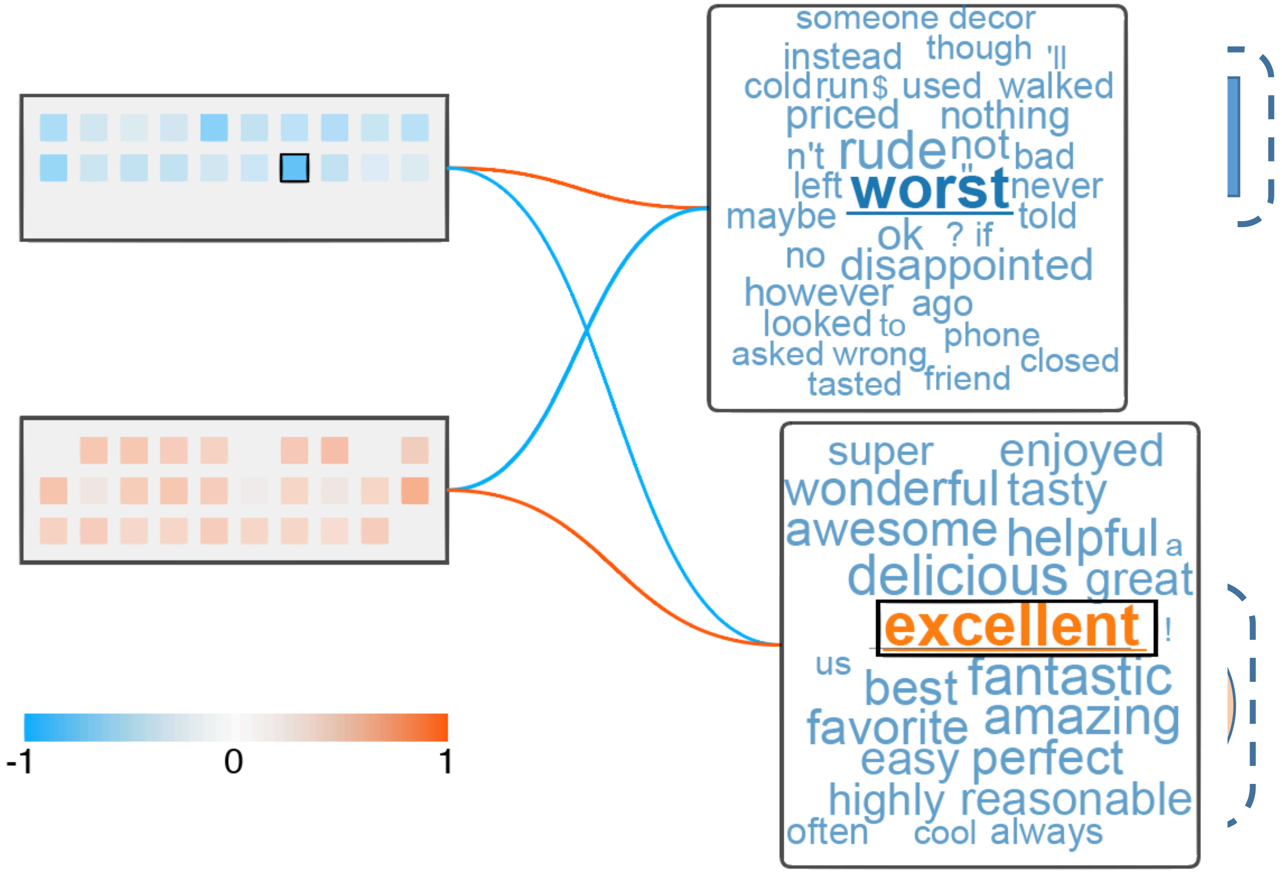
Solution

Co-clustering - Visualization

Hidden Units

Color: each unit's salience to the selected word

Words



Hidden Units Clusters (Memory Chips)

Words Clusters (Word Clouds)

Solution

Explaining individual hidden units

Bi-graph and co-clustering

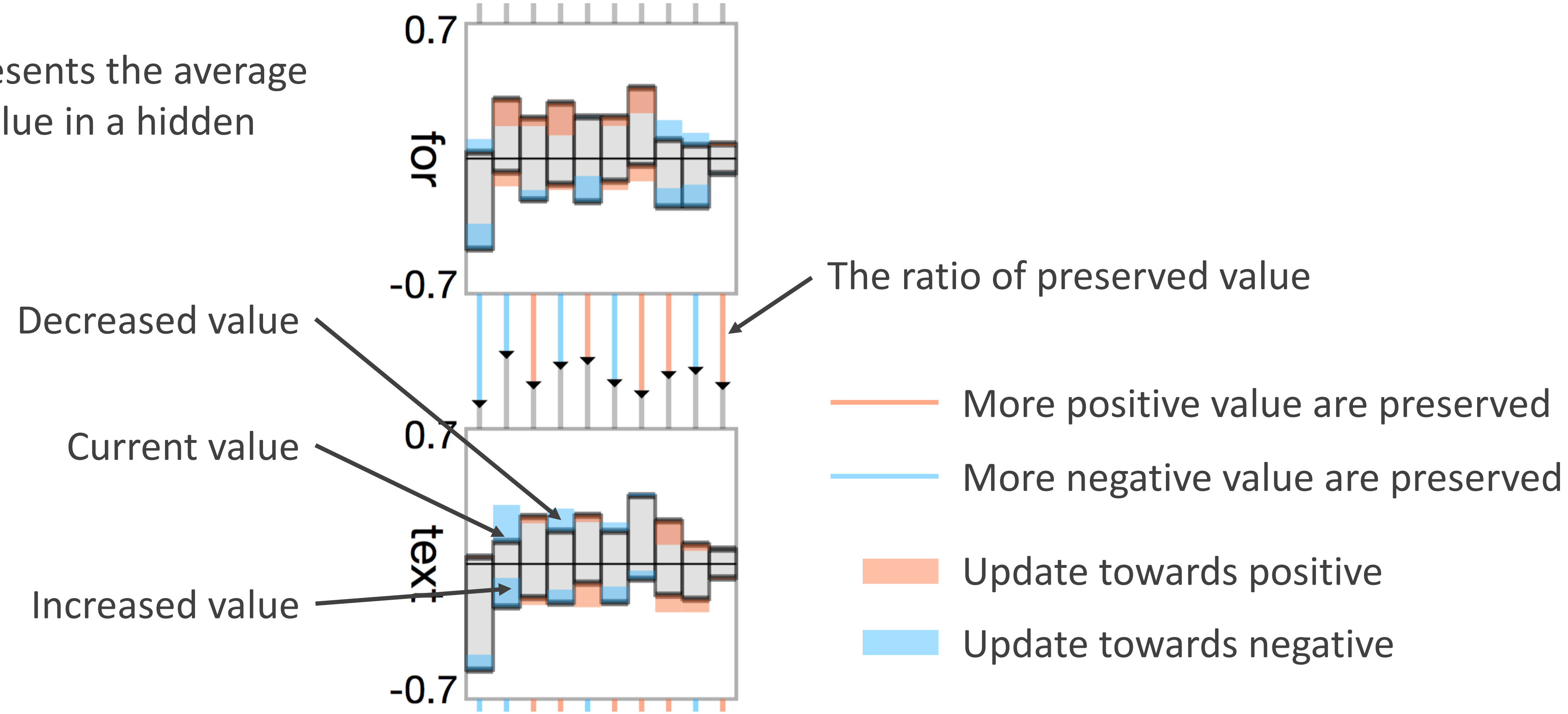
Sequence evaluation

Solution

Glyph design for evaluating sentences

Each glyph summarizes the dynamics of hidden unit clusters when reading a word

Each bar represents the average scale of the value in a hidden units cluster



Case Studies

How do RNNs handle sentiments?

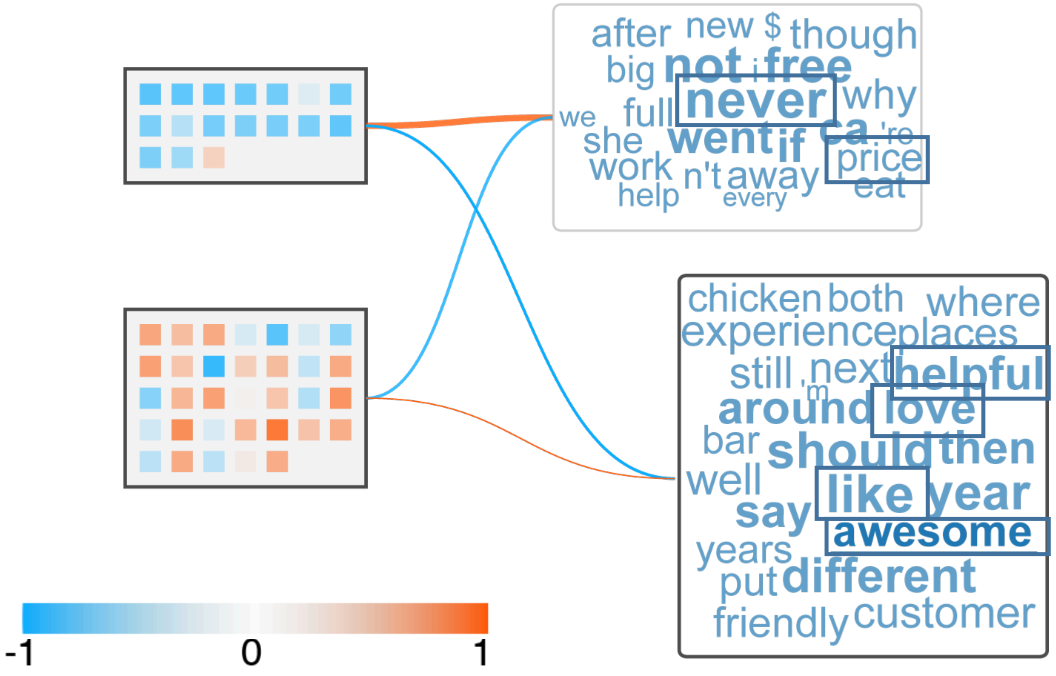
The language of Shakespeare

HKUST

Case Study – Sentiment Analysis

Each unit has two sides

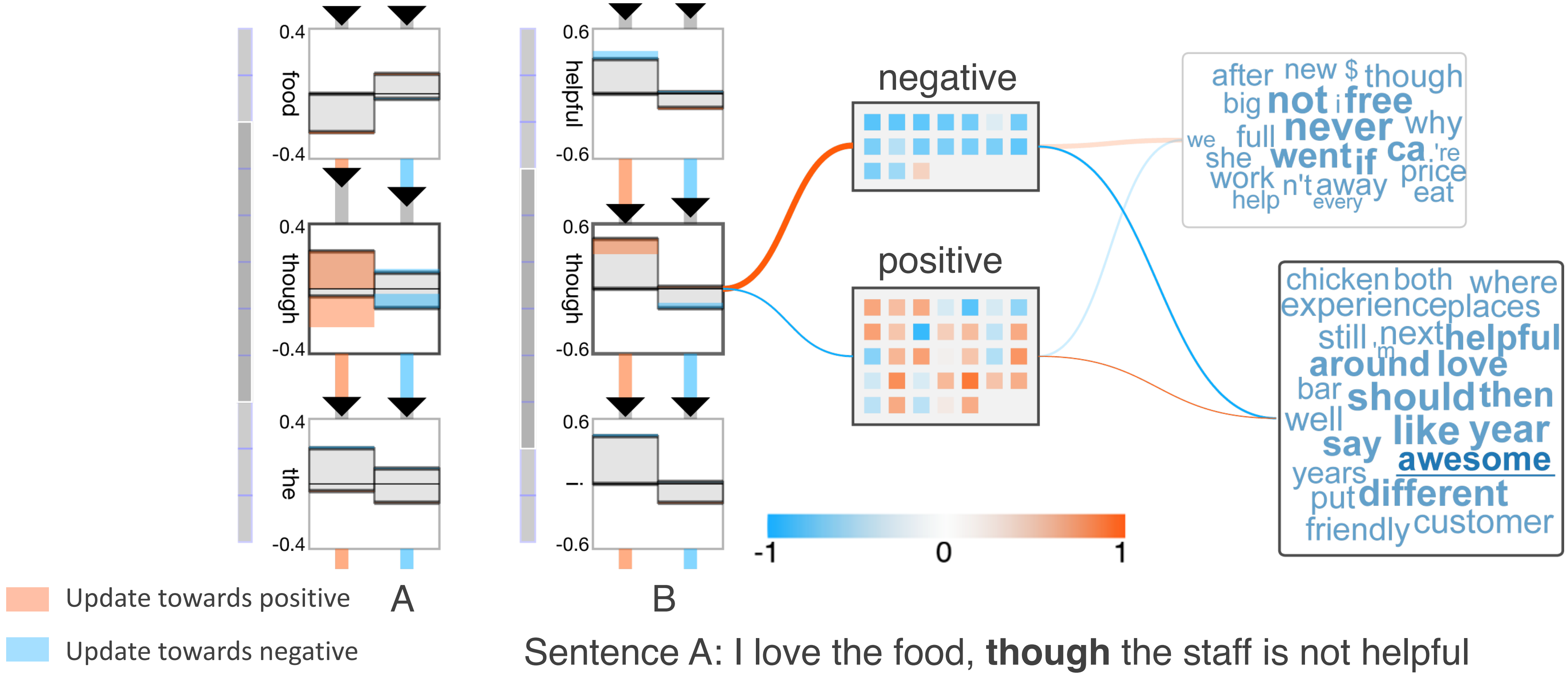
Single-layer GRU with 50 hidden units (cells), trained on Yelp review data



Case Study – Sentiment Analysis

RNNs can learn to handle the context

Single-layer GRU with 50 hidden units (cells), trained on Yelp review data



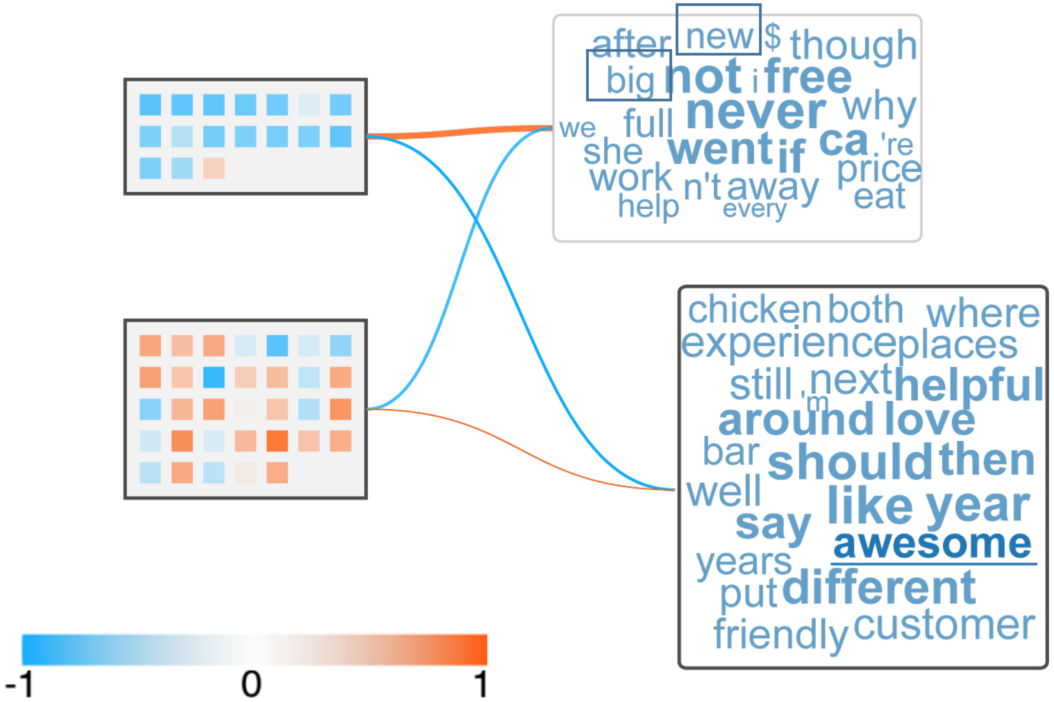
Sentence A: I love the food, **though** the staff is not helpful

Sentence B: The staff is not helpful, **though** I love the food

Case Study – Sentiment Analysis

Clues for the problem

Single-layer GRU with 50 hidden units (cells), trained on Yelp review data.

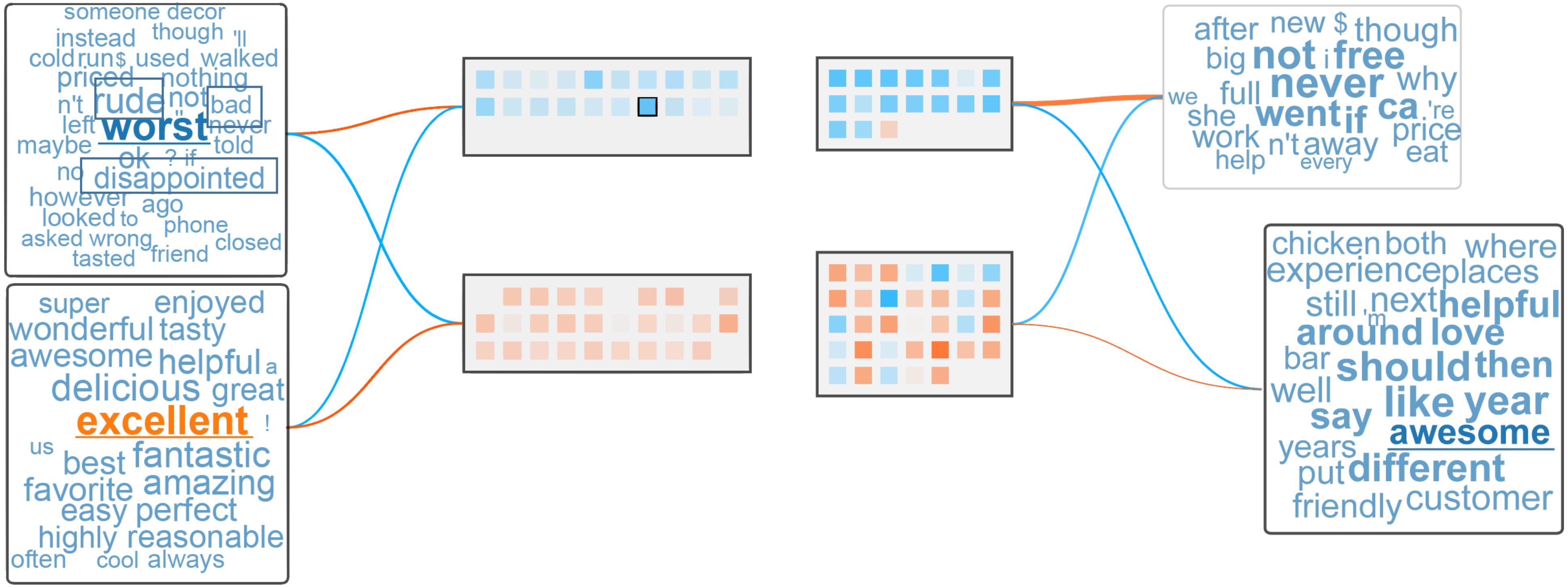


Problem: the data is not evenly sampled.

Case Study – Sentiment Analysis

Visual indicator of the performance

Single-layer GRUs with 50 hidden units (cells), trained on Yelp review data.



Accuracy (test): 91.9%

Balanced Dataset

Accuracy (test): 88.6%

Unbalanced Dataset

Case Studies

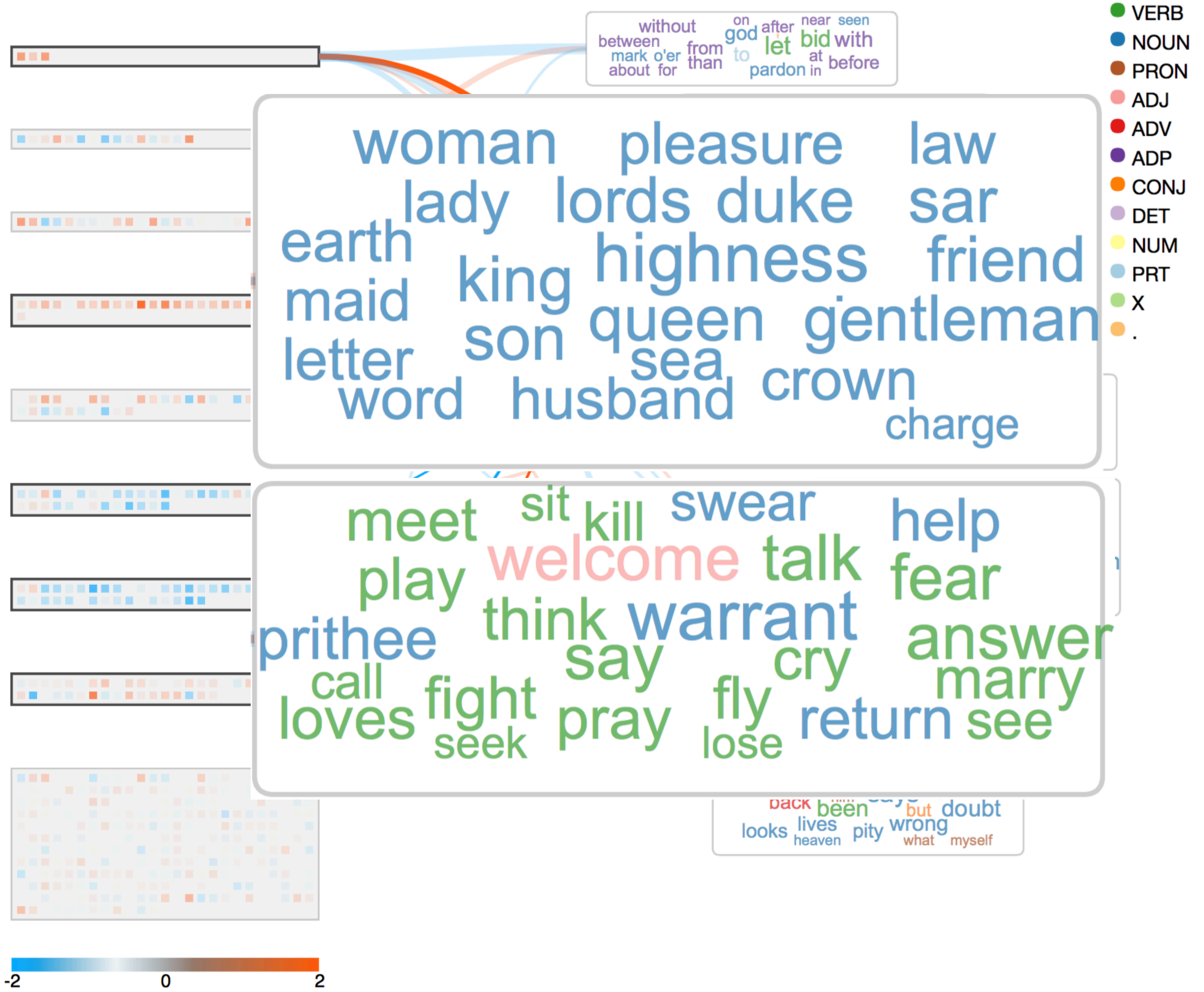
How do RNNs handle the sentiments?

The language of Shakespeare

HKUST

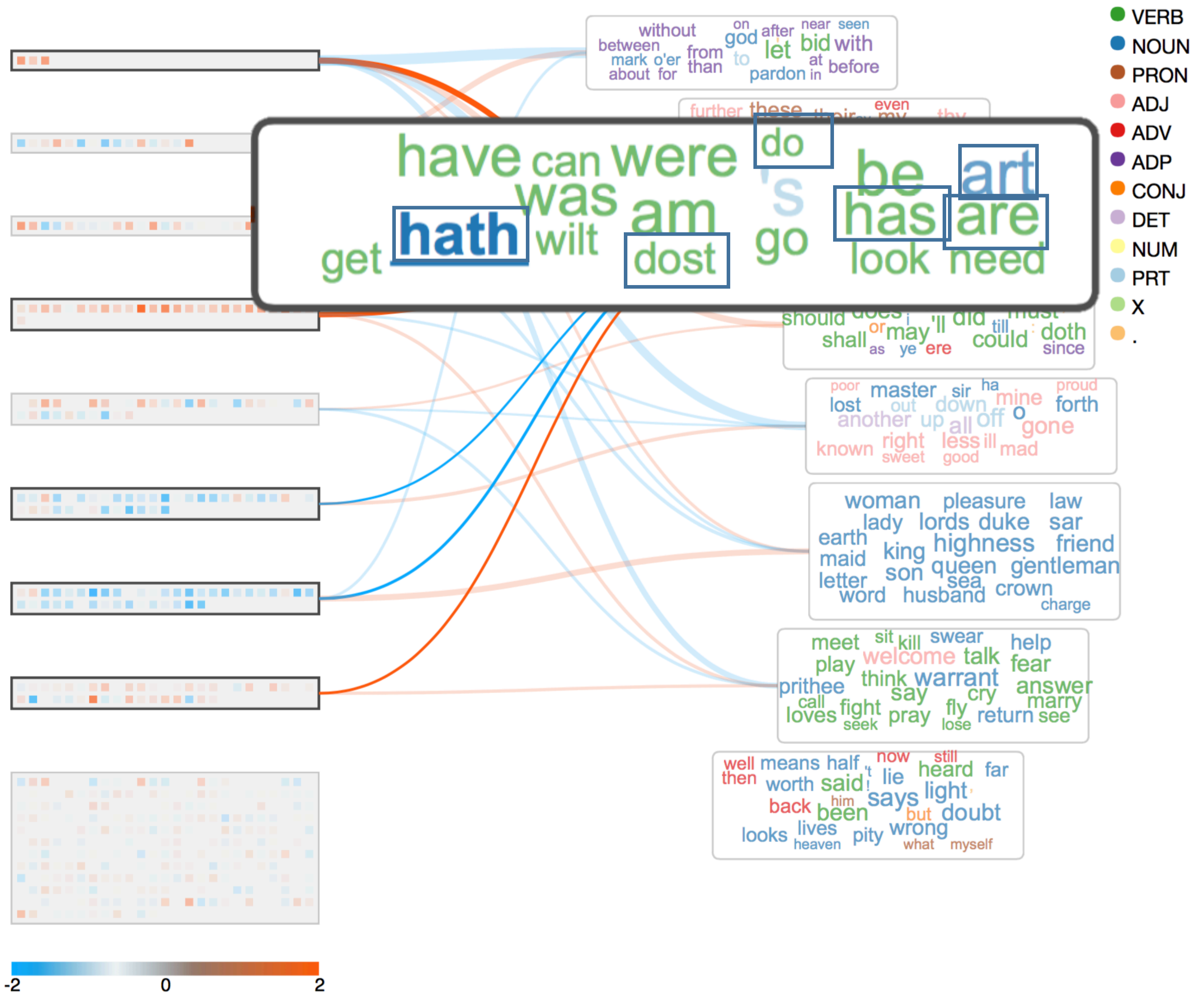
Case Study – Language Modeling

The language of Shakespeare – A mixture of the old and the new



Case Study – Language Modeling

The language of Shakespeare – A mixture of the old and the new



Discussion & Future Work

- Clustering. The quality of co-clustering? Interactive clustering?
- Glyph-based sentence visualization. Scalability?
- Text data. How about speech data?
- RNN models. More advanced RNN-based models like attention models?

Thank you!

Contact: Yao Ming, ymingaa@connect.ust.hk
Page: www.myaooo.com/rnnvis
Code: www.github.com/myaooo/rnnvis

HKUST

Technical Details

Explaining individual hidden units - Decomposition

The output of an RNN at step t is typically a probability distribution:

$$p_i = \text{softmax}\left(\boldsymbol{U}\boldsymbol{h}^{(t)}\right) = \frac{\exp\left(\boldsymbol{u}_i^T \boldsymbol{h}^{(t)}\right)}{\sum_j \exp(\boldsymbol{u}_j^T \boldsymbol{h}^{(t)})}$$

where $\boldsymbol{U} = \left[\boldsymbol{u}_i^T\right], i = 1,2, \dots, n$, is the output projection matrix.

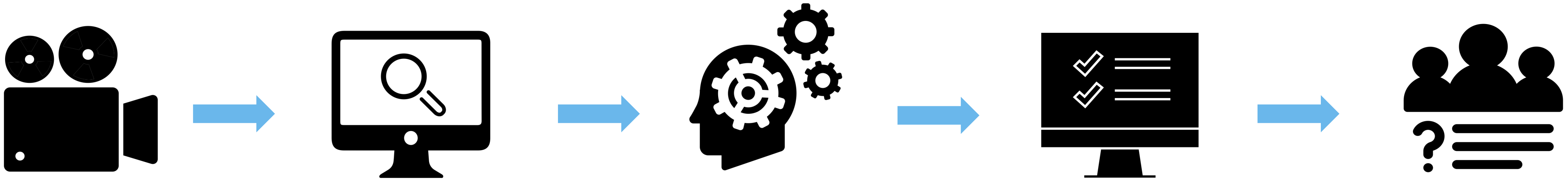
The numerator of p_i can be decomposed to:

$$\exp\left(\boldsymbol{u}_i^T \boldsymbol{h}^{(t)}\right) = \exp\left(\sum_{\tau=1}^{t} \boldsymbol{u}_i^T \left(\boldsymbol{h}^{(\tau)} - \boldsymbol{h}^{(\tau-1)}\right)\right) = \prod_{\boldsymbol{\tau}=\mathbf{1}}^{\boldsymbol{t}} \exp(\boldsymbol{u}_i^T \Delta \boldsymbol{h}^{(t)})$$

Here $\exp(\boldsymbol{u}_i^T \Delta \boldsymbol{h}^{(t)})$ is the multiplicative contribution of input word w_t, the update of hidden state $\Delta \boldsymbol{h}^{(t)}$ can be regard as the model's response to w_t.

Evaluation

Expert Interview



1. Show a tutorial video
2. Explore the tool
3. Compare two models
4. Answer questions
5. Finish a survey

HKUST

Challenges

What are the challenges?

1. The complexity of the model

    - Machine Translation: 4-layer LSTMs, 1000 units/layer (Sutskever I. et al., 2014)
    - Language Modeling: 2-layer LSTMs, 650 or 1500 units/layer (Zaremba et al., 2015)

2. The complexity of the hidden memory

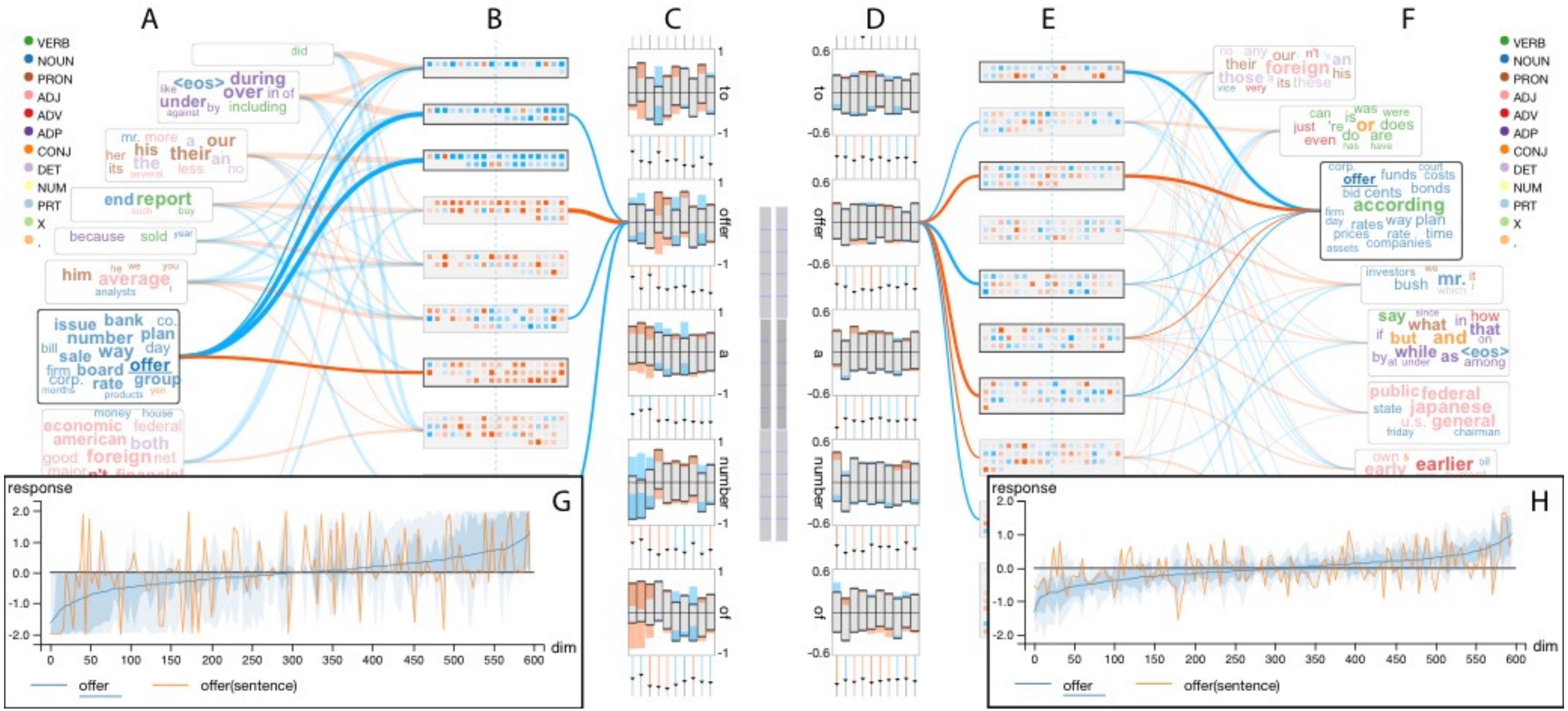
    - Semantic information are distributed in hidden states of an RNN.

3. The complexity of the data

    - Patterns in sequential data like texts are difficult to be analyzed and interpreted

Other Findings

Comparing LSTMs and vanilla RNNs



Left (A-C): co-cluster visualization of the last layer of an RNN. Right (D-F): visualization of the cell states of the last layer of an LSTM. Bottom (GH): two models' responses to the same word "offer".

Contribution

- A visual technique for understanding what RNNs learned.
- A VA tool that ablates the hidden dynamics of a trained RNN.
- Interesting findings with RNN models.