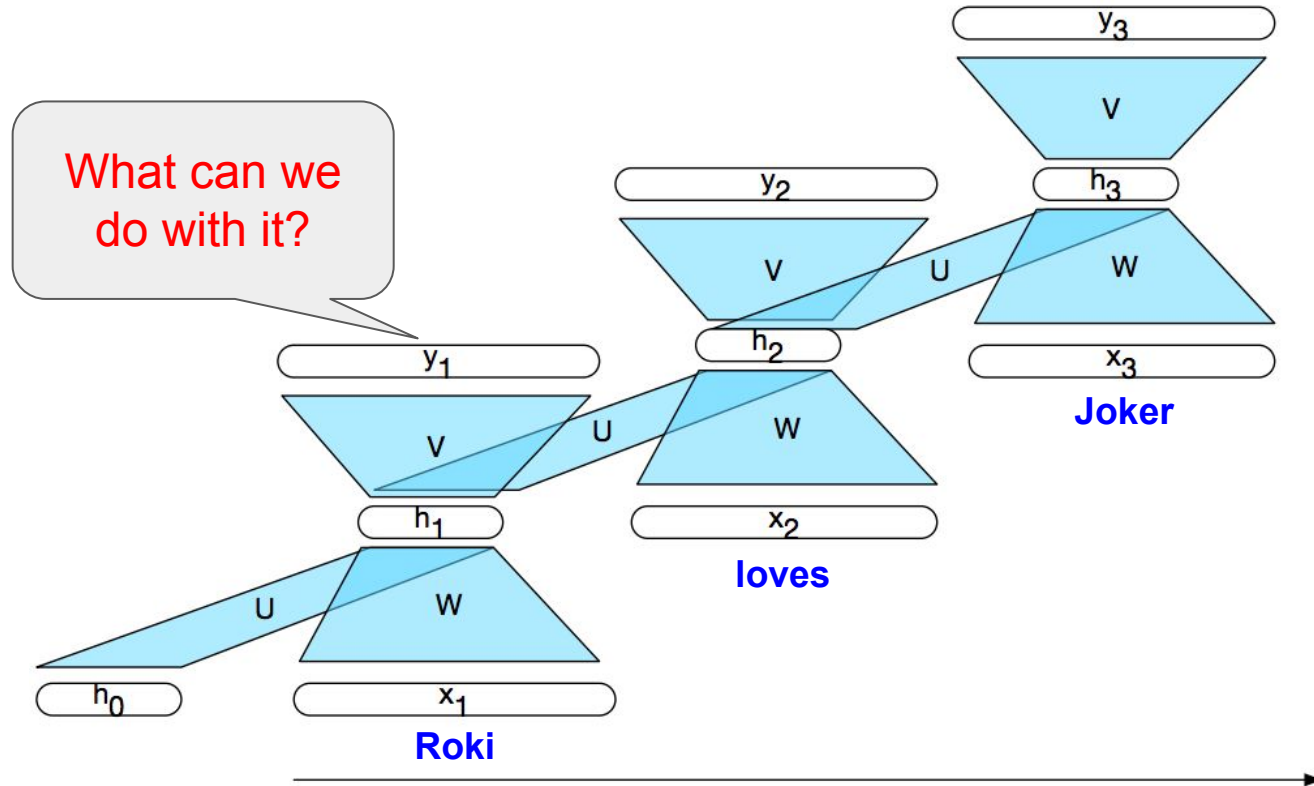


Week 4: Named Entity Recognition (Cont.)

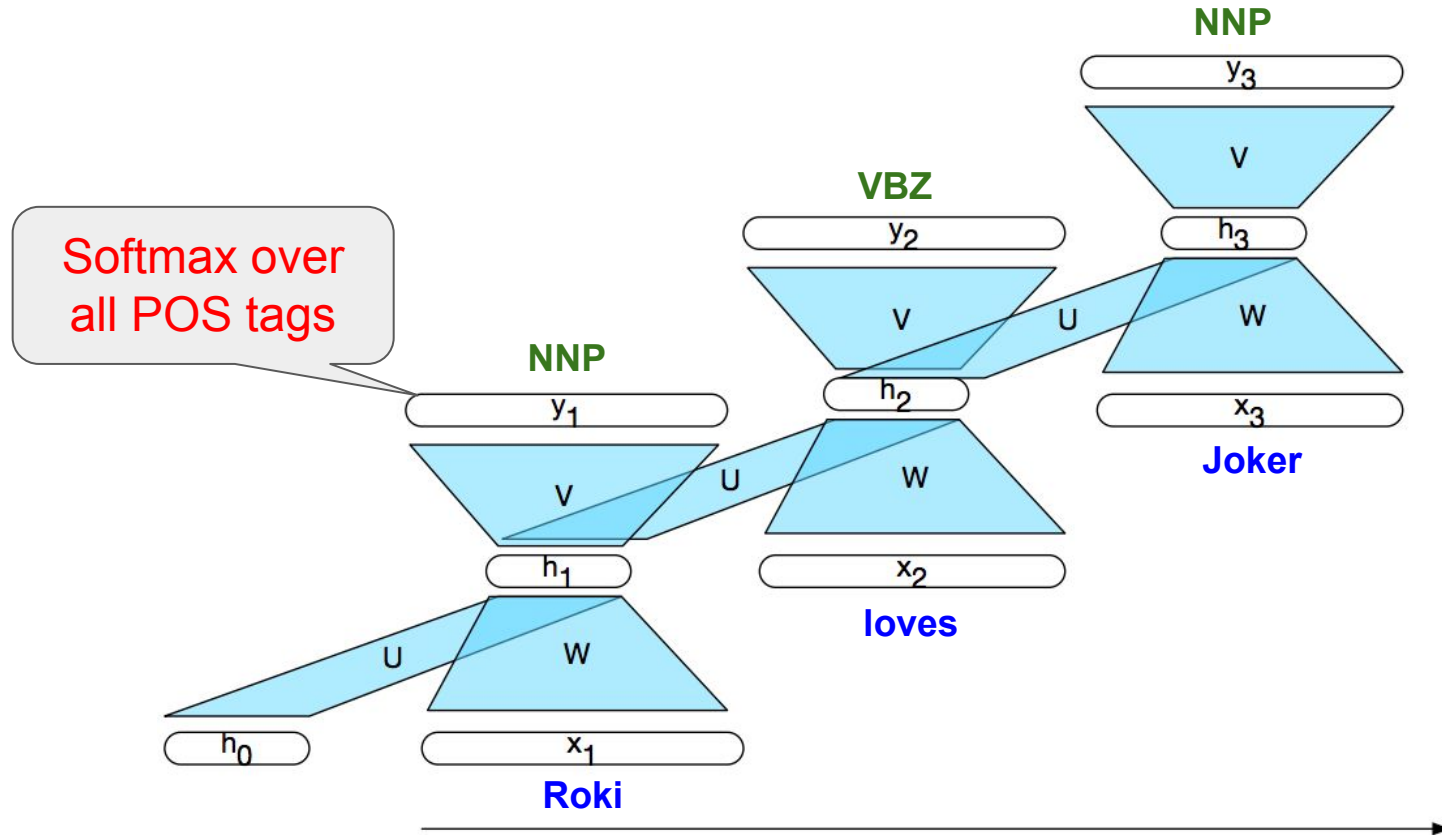
Nhung Nguyen
slides courtesy of NaCTeM

Sequence processing with recurrent neural networks

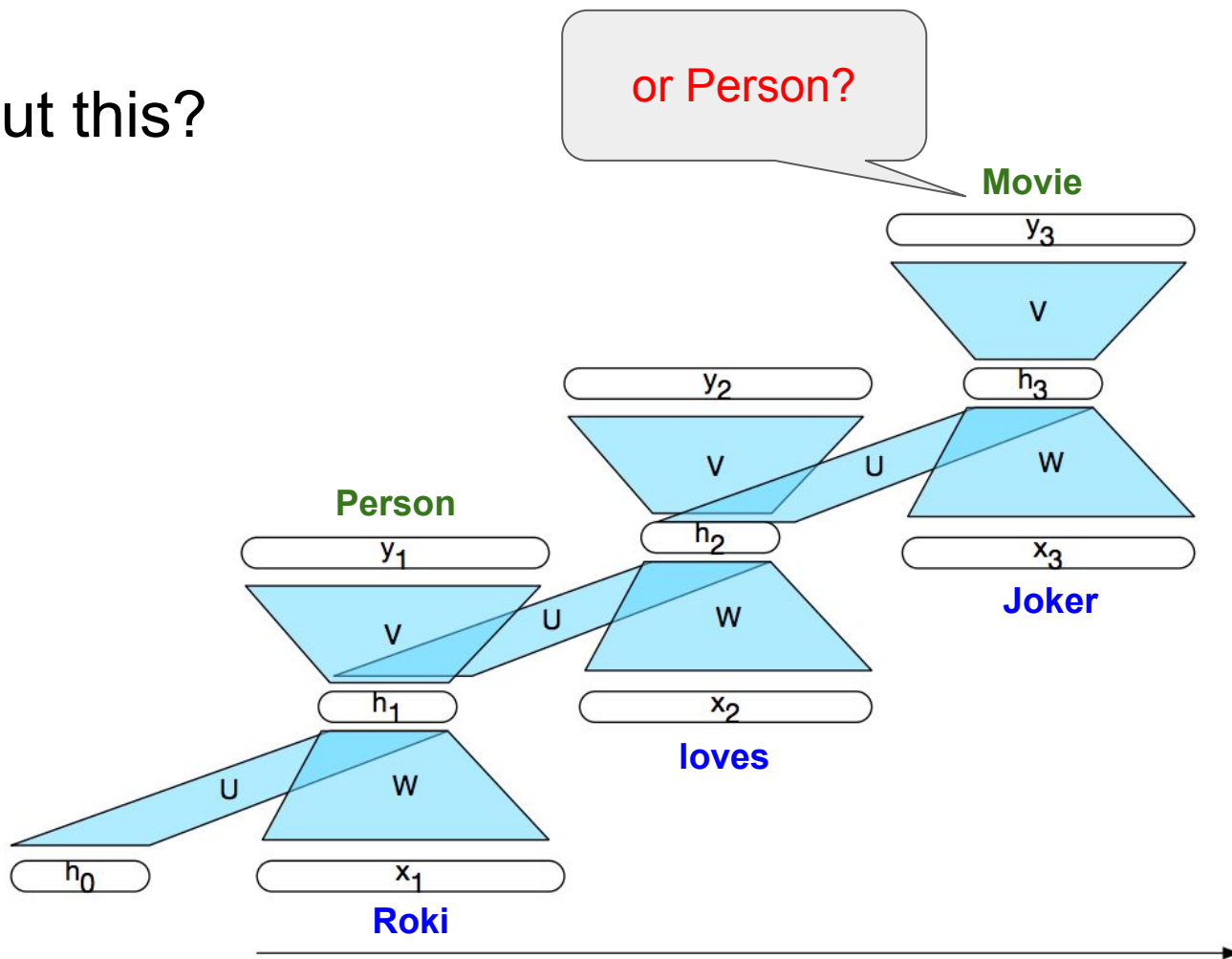
Look back to RNN



Part-of-speech tagging



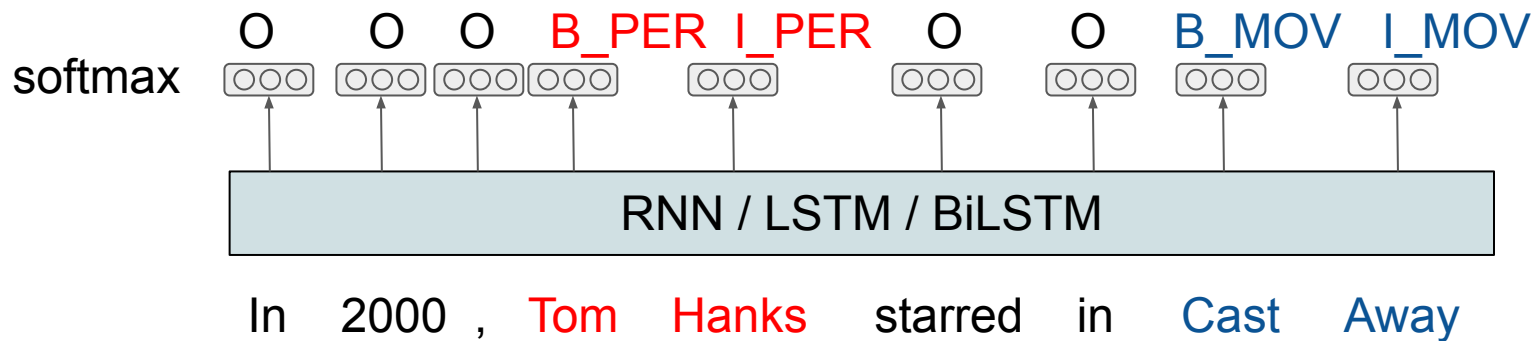
How about this?



Local approach using softmax

- Predicting tags independently

$$Pr(tag|token) = softmax(\mathbf{W}\mathbf{h}_{token} + \mathbf{b})$$



- Why do we use a recurrent net?

Local approach using softmax (cont.)

- Predicting tags independently

$$Pr(tag|token) = softmax(\mathbf{W}\mathbf{h}_{token} + \mathbf{b})$$

- Training a local model is to minimize negative log likelihood

$$L(\theta) = \sum_{S \in D} \sum_{(tag, token) \in S} -\log Pr(tag|token)$$

where D is a training set (a set of tagged sentences)

Global approach

- Predicting all tags at once

$$Pr(tag_{1:n}|token_{1:n}) = \frac{\exp\{f(tag_{1:n}, token_{1:n})\}}{\sum_{tag'_{1:n} \in T} \exp\{f(tag'_{1:n}, token_{1:n})\}}$$

- Training a local model is to minimize negative log likelihood

$$L(\theta) = \sum_{(tag_{1:n}, token_{1:n}) \in D} -\log Pr(tag_{1:n}|token_{1:n}; \theta)$$

Global approach (Cont.)

- Using linear-chain CRF
- The feature function can be replaced:

From a BiLSTM

$$f(tag_i, tag_{i-1}, token_i) = \mathbf{W}h_i + \mathbf{b}$$

Local vs global approaches

- For sequence labelling tasks, e.g., POS and NER, global approaches are better than local ones
- Why?
 - An *I* tag can only appear after an *I* or a *B* (never an *O*)
 - There are often more *O*s than *B*s and *I*s, ...

CRF vs. Neural networks

CRF

- Feature engineering
- Do not need pre-trained vectors
- Models are roughly interpretable
- *Perform well with datasets that have many NE categories(*)*

Neural networks

- Do not need features
- Need **pre-trained** vectors from big language models
- Models use implicit features (created by hidden layers) → **not easy to interpret**
- *Perform not so well with datasets that have many NE categories(*)*

(*) This observation is only based on my personal experiences

Summary

- There are several approaches to NER, **global approach** is better than local one
- **CRF**, a traditional approach to NER, usually produces state-of-the-art performance, but we **need feature engineering**
- **Neural networks** have more advantages in terms of feature engineering but they **have limitations in interpretability** and resources.