Natural Language Processing

Approaches to sequence labelling

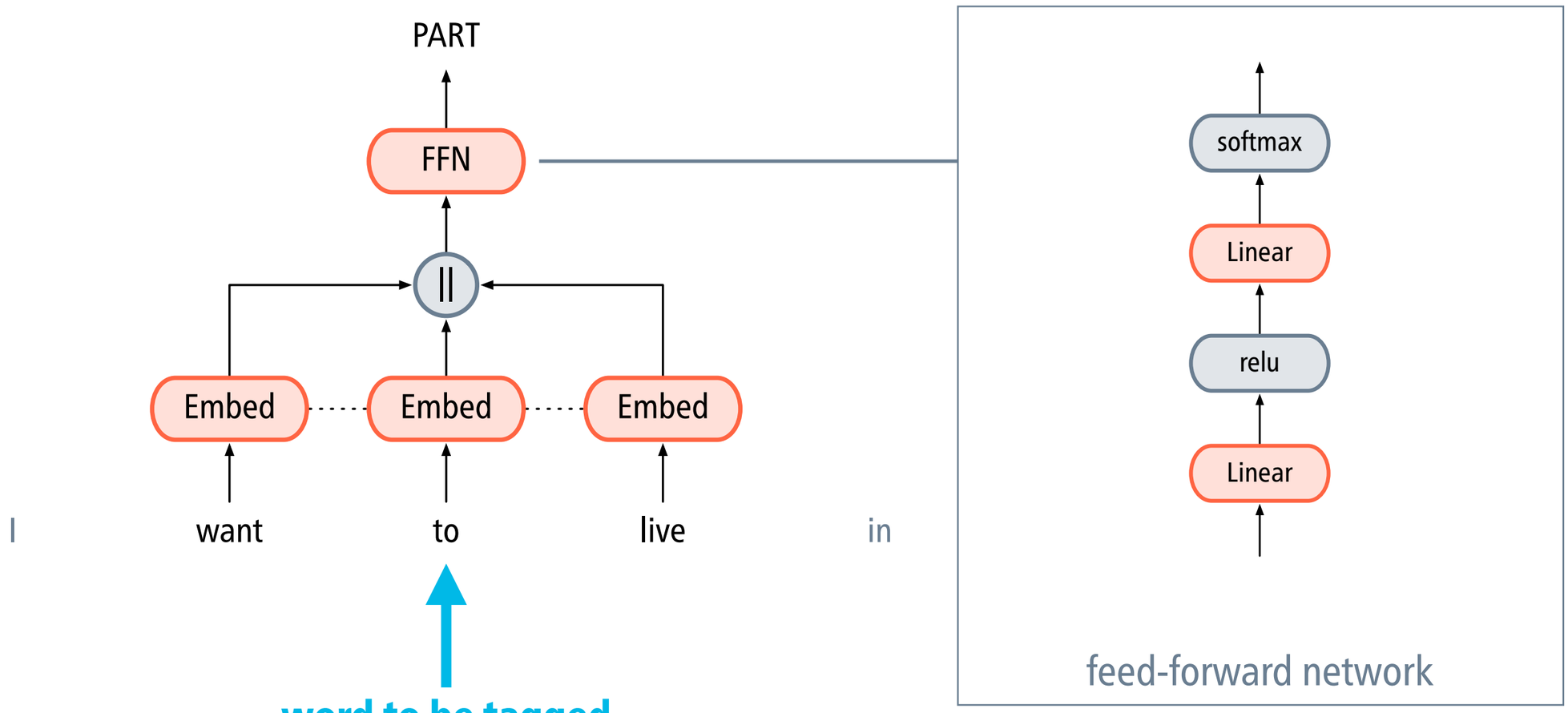
Marco Kuhlmann
Department of Computer and Information Science

LINKÖPING UNIVERSITY

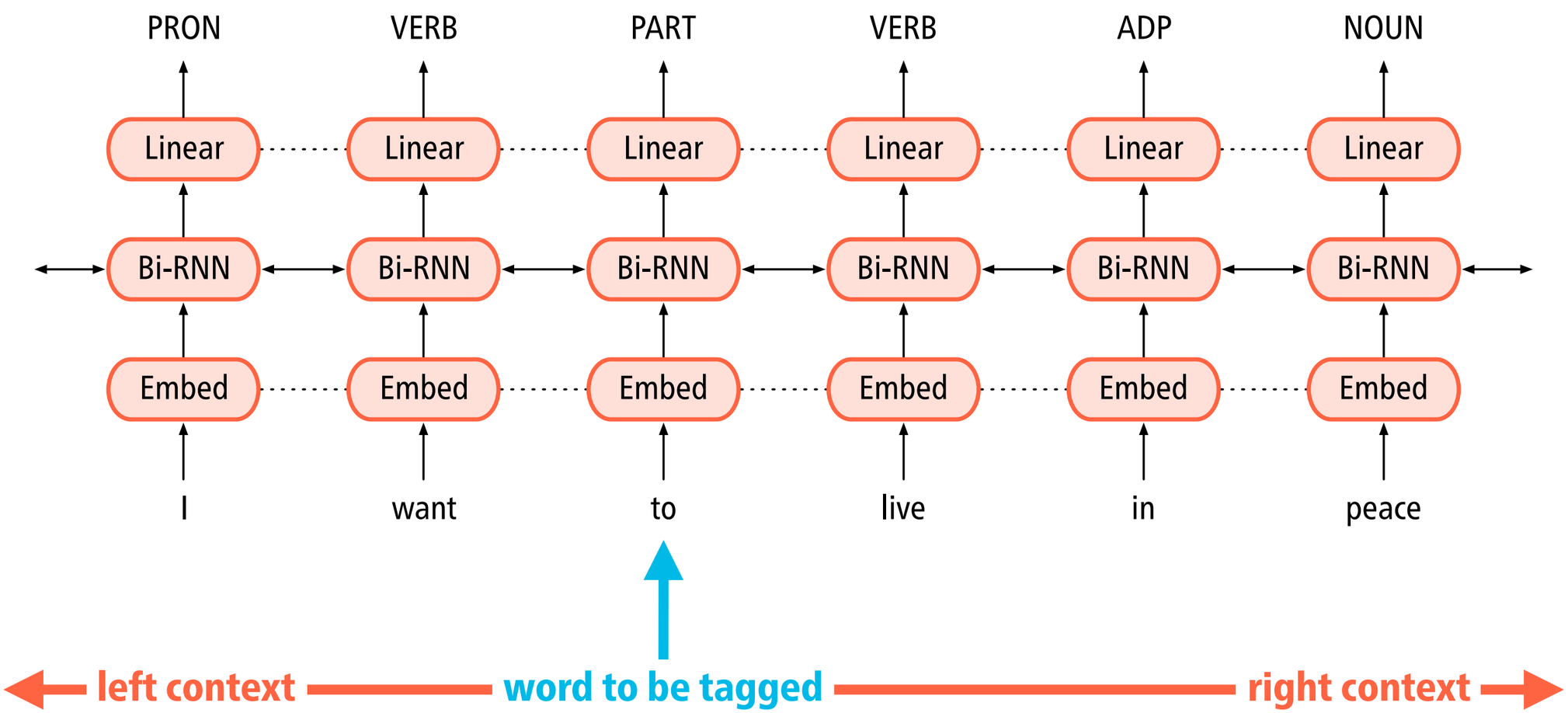
Approaches to sequence labelling

Local search	**Global search**
sequence of independent classification problems	combinatorial optimisation over the full search space
expressive feature models (RNNs, attention)	restricted feature models (factorised feature functions)
no need for specialised algorithms; fast	requires specialised algorithms; slow

Fixed-window model



Bidirectional RNN model

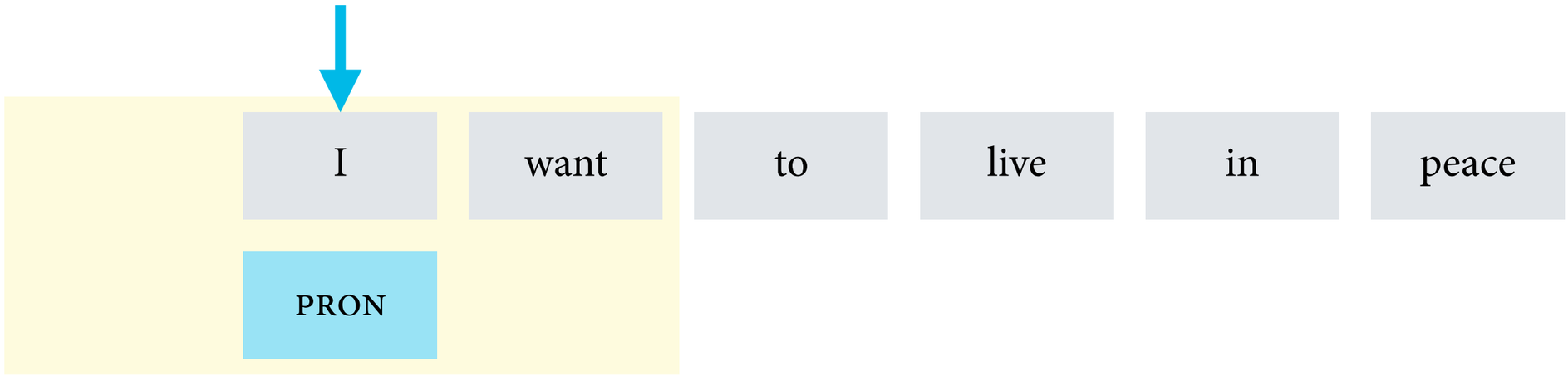


Labels are interdependent

I	want	to	live	in	peace
PRON	VERB	PART	VERB	ADP	NOUN
PRON	VERB	ADP	VERB	ADP	NOUN

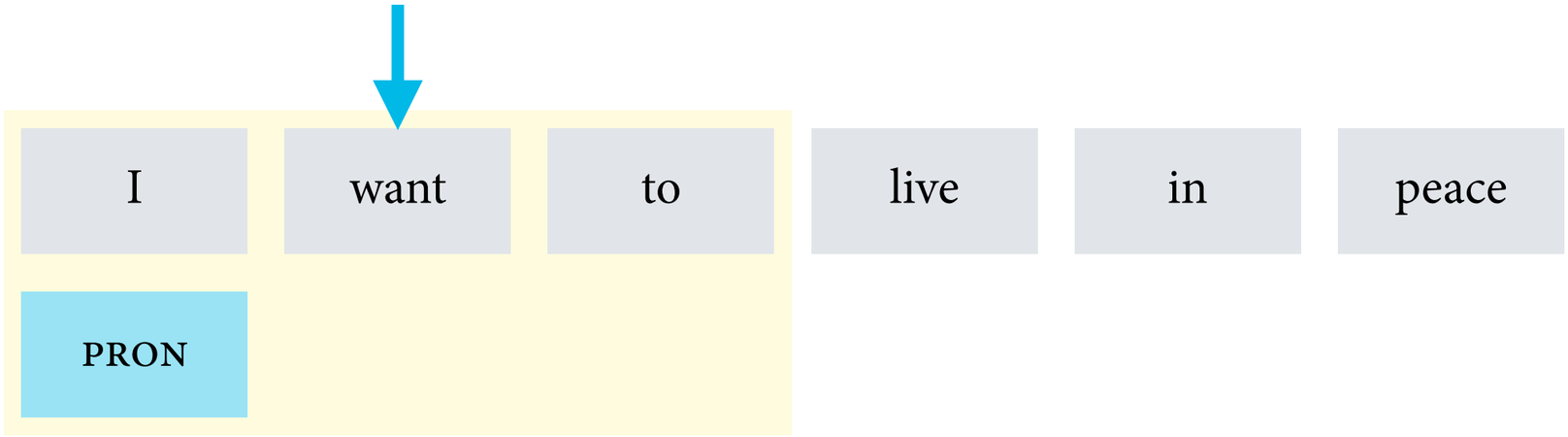
Some combinations of part-of-speech tags
are more likely than others.

Autoregressive tagging with a fixed-window model



The model predicts the tag for the first word in the sentence.
Features are extracted from a context window.

Autoregressive tagging with a fixed-window model



The prediction at the second position can use features defined over the tags that have already been predicted.

Training autoregressive models

- At test time, we run the model incrementally, and feed it with its own predicted labels.
- At training time, we feed the model with the gold-standard label. This regime is called **teacher forcing**.
- Teacher forcing can be problematic, because the model does not learn to deal with its own prediction errors.

  difference between training time and prediction time (exposure bias)

Approaches to sequence labelling

Local search	**Global search**
sequence of independent classification problems	combinatorial optimisation over the full search space
expressive feature models (RNNs, attention)	restricted feature models (factorised feature functions)
no need for specialised algorithms; fast	requires specialised algorithms; slow

Dealing with combinatorial explosion

- The number of candidate sequences is exponential in the length of the input sequence, so naive optimisation is doomed to fail.
- To make the search problem tractable, we will restrict ourselves to **factorised scoring functions**.
- Factorised scoring functions will allow us to use specialised algorithms to solve the optimisation problem in polynomial time.

  Example: Viterbi algorithm

Factorised scoring function

candidate output sequence

sequence length

$$\text{score}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = \sum_{i=1}^{|\boldsymbol{x}|} \text{score}_1(\boldsymbol{x}, i, y_i; \boldsymbol{\theta}) + \sum_{i=1}^{|\boldsymbol{x}|} \text{score}_2(\boldsymbol{x}, i, y_{i-1}, y_i; \boldsymbol{\theta})$$

input sequence

score for a single label

score for a pair of labels

Maximum Entropy Markov Model (MEMM)

$$\text{score}(\boldsymbol{x}, i, y_{i-1}, y_i) \;=\; \text{score}_1(\boldsymbol{x}, i, y_i) + \text{score}_2(\boldsymbol{x}, i, y_{i-1}, y_i)$$

$$P(\boldsymbol{y} \mid \boldsymbol{x}) \;=\; \prod_{i=1}^{|\boldsymbol{x}|} \frac{\exp\big(\text{score}(\boldsymbol{x}, i, y_{i-1}, y_i)\big)}{\sum_{y'} \exp\big(\text{score}(\boldsymbol{x}, i, y_{i-1}, y')\big)}$$

candidate label

Algorithmic problems for MEMMs

- **Training:** To train a model, we want to minimise the negative log likelihood on gold-standard examples $((\boldsymbol{x}, y_{i-1}), y_i)$.

  standard softmax regression problem

- **Decoding:** For a trained model, we want to find the most probable label sequence $\boldsymbol{y}$, given the input sequence $\boldsymbol{x}$.

  involves the search over exponentially many candidate sequences → Viterbi

The label bias problem

Zhang and Teng (2021)

Sequence id	Sequence
1	ACA
2	ACA
3	AA
4	AAB
5	ABB
6	CCC

Global probability

$\frac{2}{6}$

$\frac{1}{6}$

Factorised probability

How often is A followed by C?

How often is C followed by A?

$$\frac{5}{6} \cdot \frac{2}{6} \cdot \frac{2}{4} = \frac{5}{36}$$

How often does A start a sequence?

$$\frac{5}{6} \cdot \frac{2}{6} \cdot \frac{1}{1} = \frac{5}{18}$$

How often is A followed by B?

How often is B followed by B?

Conditional random field (CRF)

Lafferty et al. (2001)

$$\text{score}(\boldsymbol{x}, \boldsymbol{y}) = \sum_i^{|\boldsymbol{x}|} \text{score}(\boldsymbol{x}, i, y_{i-1}, y_i)$$

$$P(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{\exp(\text{score}(\boldsymbol{x}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}'} \exp(\text{score}(\boldsymbol{x}, \boldsymbol{y}'))}$$

candidate label sequence