# A Neural Language Model for Dynamically Representing the Meanings of Unknown Words and Entities in a Discourse


Sosuke Kobayashi

Preferred Networks

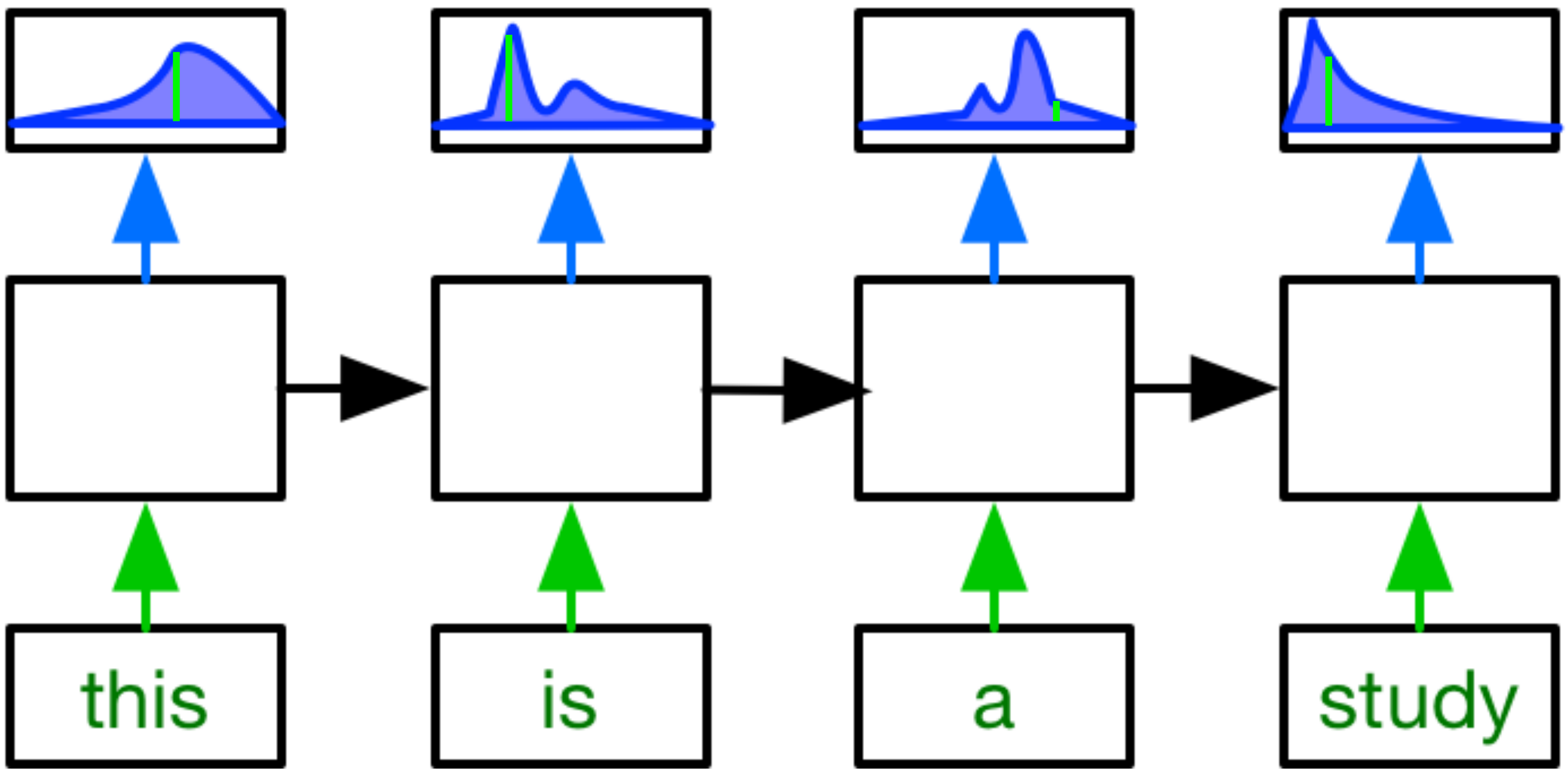
Naoaki Okazaki

Tokyo Institute of Technology

Kentaro Inui

Tohoku University / RIKEN

# RNN Language Model



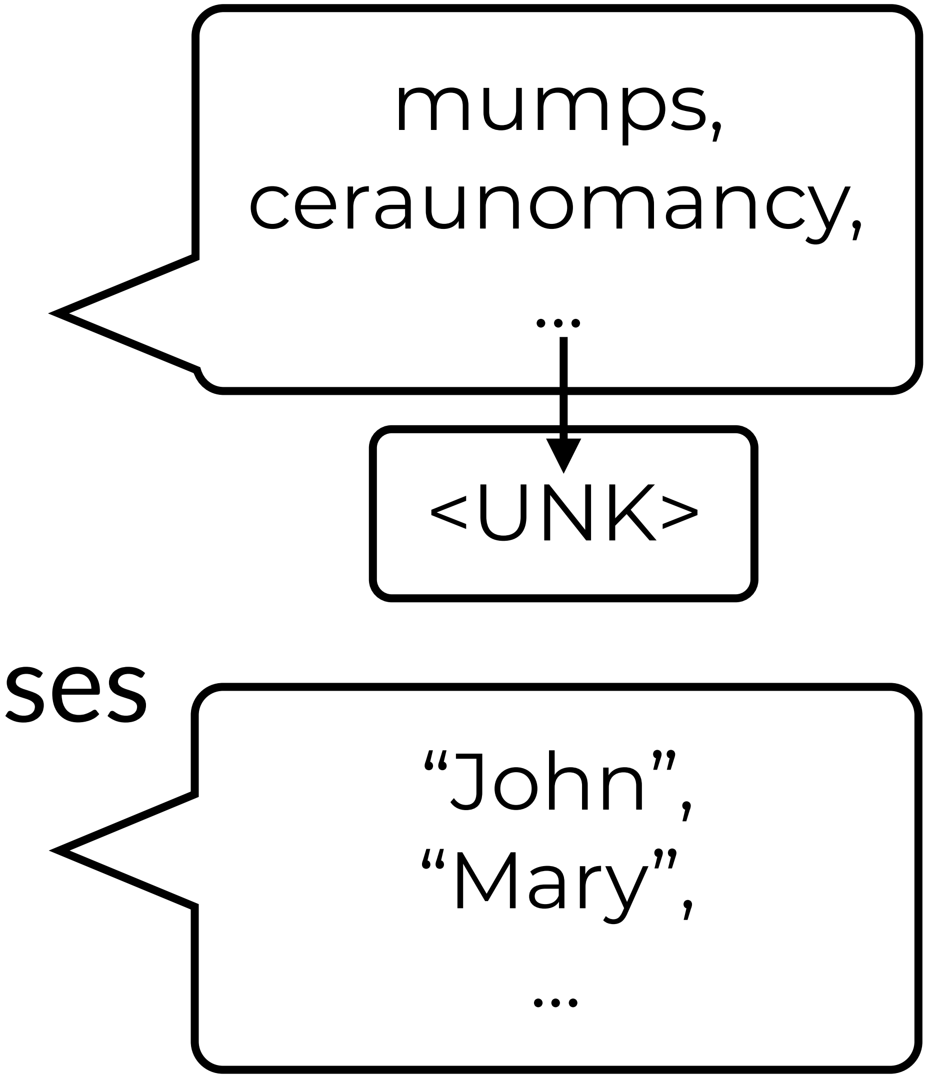

- Output matrix: calculate probability
- RNN matrices: encoding context
- Embedding matrix: represent word meaning

# Neural Language Model

- Embedding matrix
  Output matrix
  - Cannot cover all words
    → **Unknown words**
  - Referents differ by discourses
    → **Unknown entities**

# Neural Language Model

- Embedding matrix
  Output matrix
  - Cannot cover all words
    → **Unknown words**
  - Referents differ by discourses
    → **Unknown entities**



# Dynamic Entity Representation

[Kobayashi et al. NAACL 2016]

- Unknown's meaning representation cannot be obtained statically…

  ↓

  **Dynamically update meaning representation while reading text**

- Infer on-the-fly meanings from context

- "she contracted ***mumps***" → ***mumps*** is a disease?
- "***John*** loves Fender" → "***John***" is a guitarist?

# Usage: Input Embedding

- Language models
  encode context words and predict next words
  - Input word embeddings can be replaced
- **Dynamic modeling makes context informative**

- "… with him, John played [???]"
- with dynamic model:
  "… with him, **<John; guitarist>** played [???]"

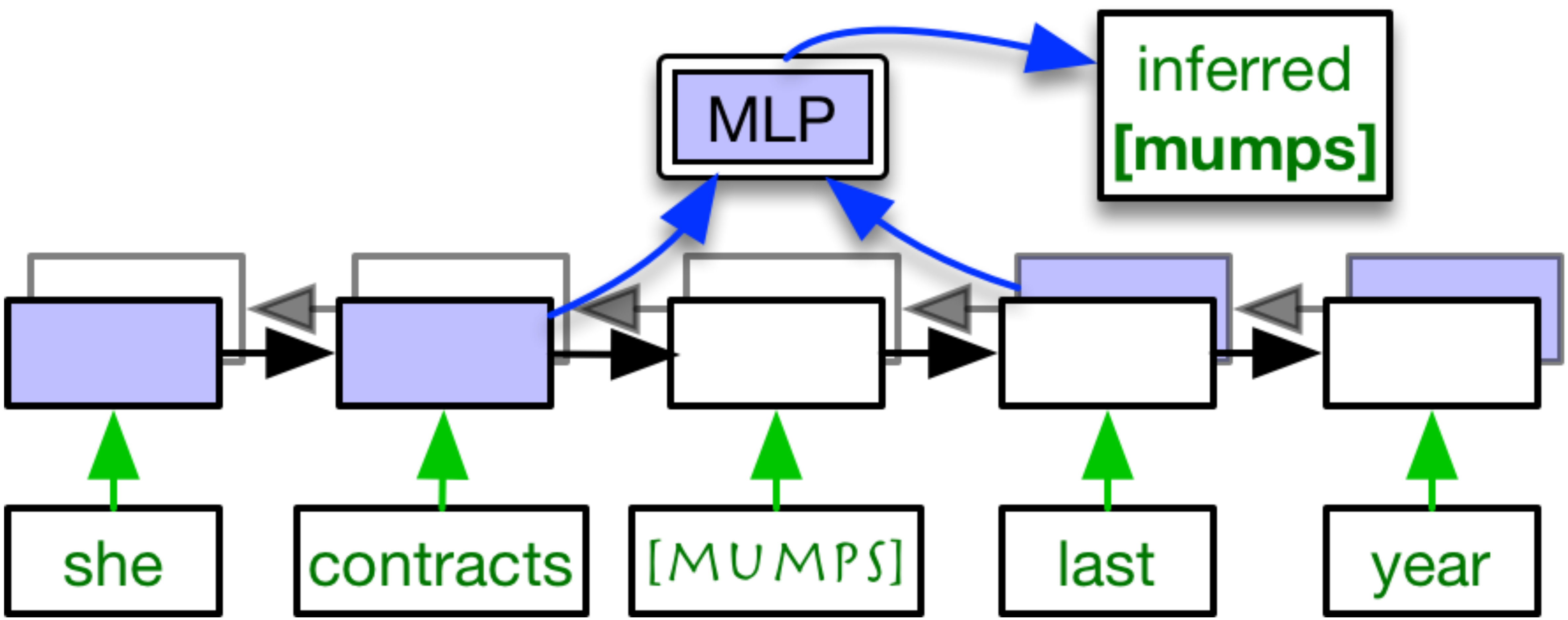
# Usage: Output Matrix

- Language models
  encode context words and predict next words
  - Output matrix's rows can be replaced
- **Dynamic modeling makes target informative**

  - "… she is a big fan of [???]"
    John? Mary?
  - with dynamic model:
    "… she is a big fan of [???]"
    **<guitarist>? <mother>?**

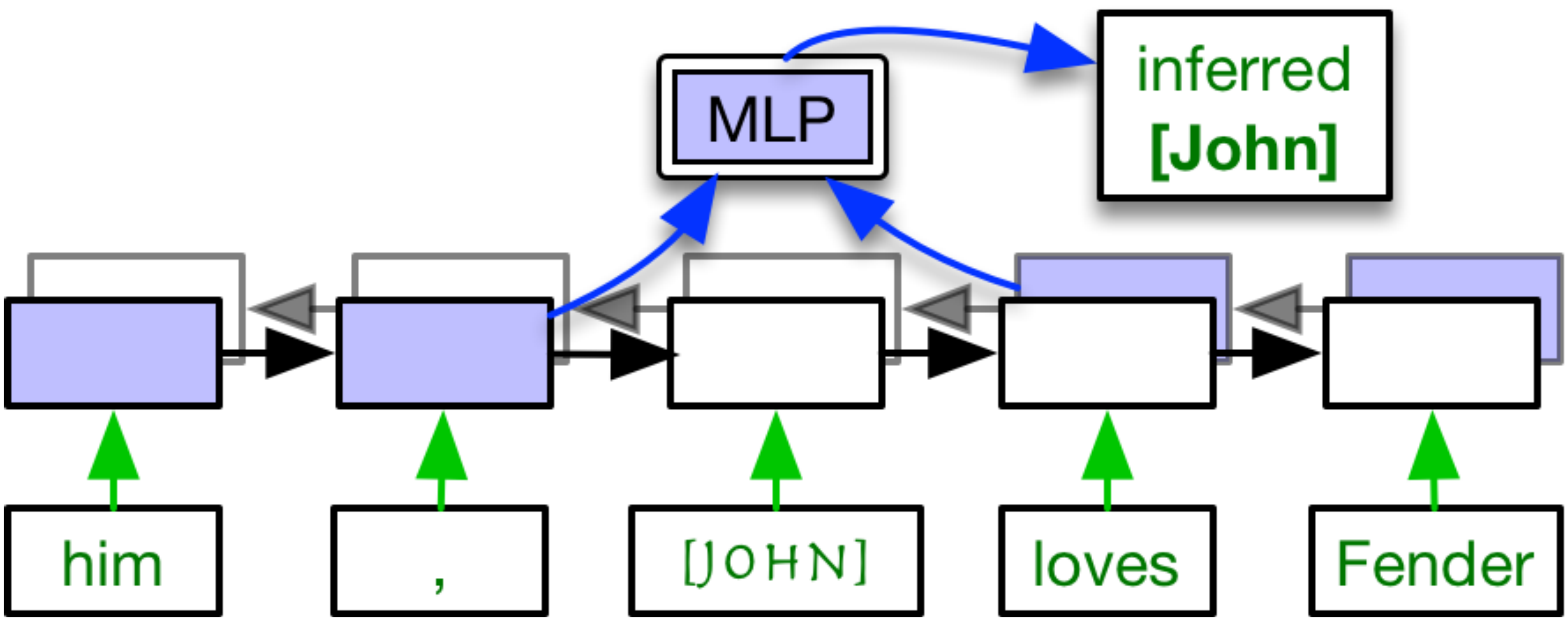
# Recipe: Context Encoding

- Encode context of the target word
  - e.g. bi-directional RNN



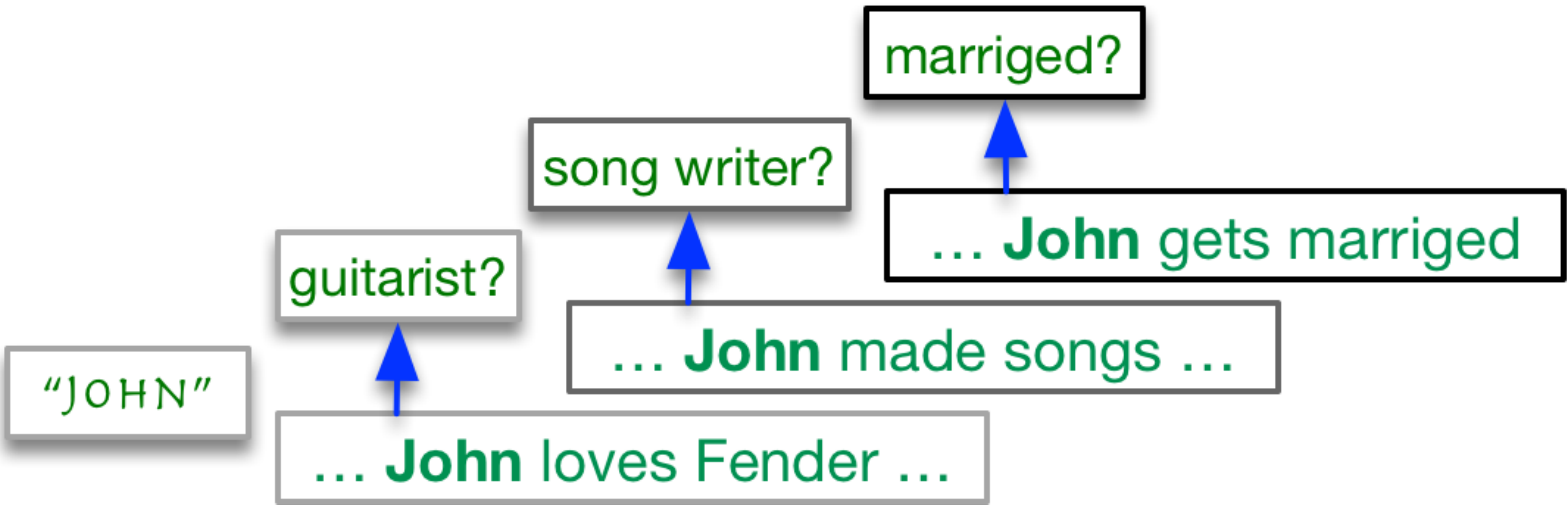
# Recipe: Context Encoding

- Encode context of the target word
  - e.g. bi-directional RNN



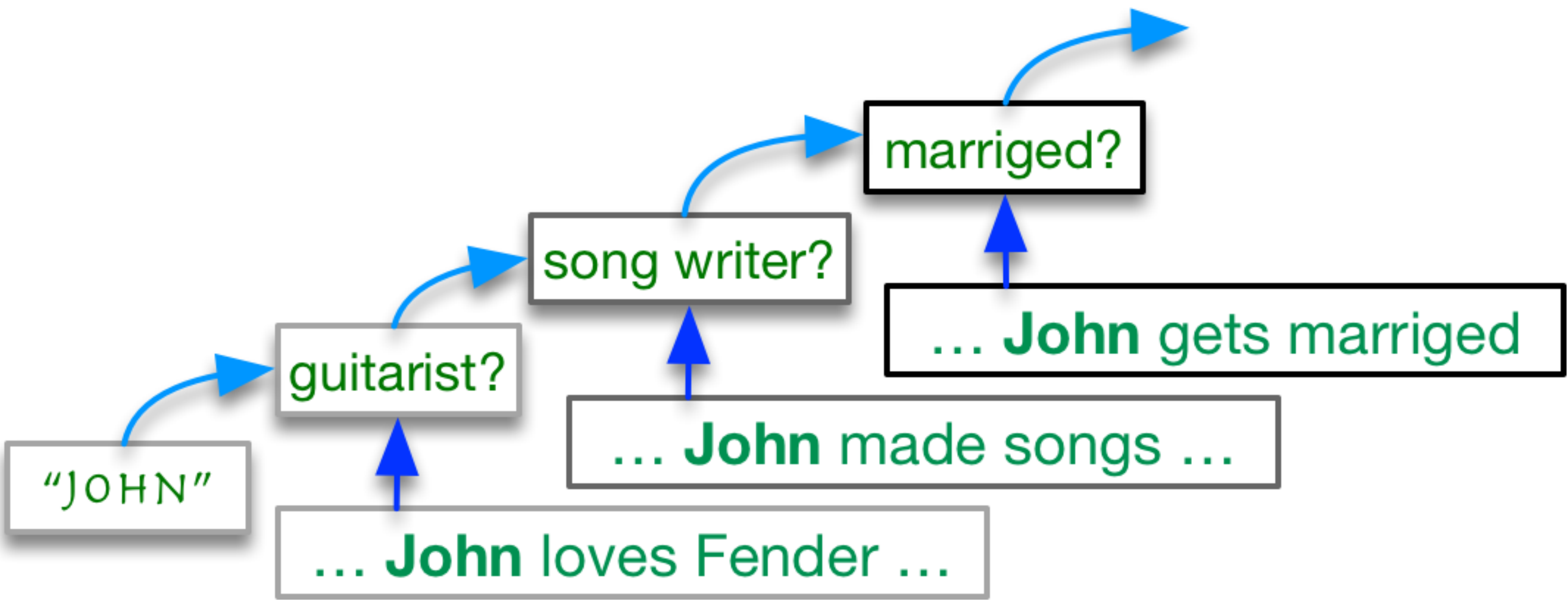
# Recipe: Context Merging

- Merge multiple contexts where the target occurs
  - e.g. RNN, max-pooling



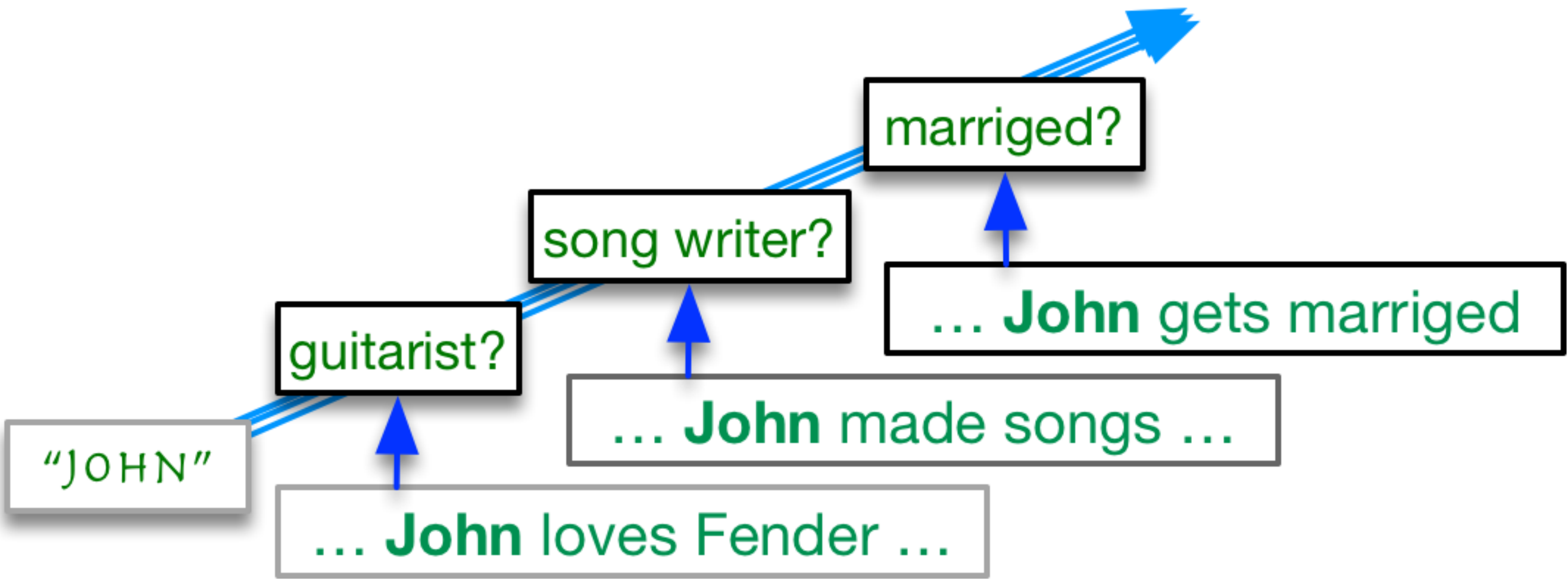
# Recipe: Context Merging

- Merge multiple contexts where the target occurs
  - e.g. **RNN**, max-pooling



# Recipe: Context Merging

- Merge multiple contexts where the target occurs
  - e.g. RNN, **max-pooling**



# Dataset for Evaluation

- Dataset for language modeling from OnteNotes
  - Coreferents are unified and anonymized

John, he, … → *[UNK1]* Mary, she, … → *[UNK2]*

| | |
|---|---|
| *RAW* | **John** loves guitars.<br>**Mary** did not prefer music.<br>But, many people are big fans of **him**. … |
| *OURS* | ***[UNK1]*** loves guitars.<br>***[UNK2]*** did not prefer music.<br>But, many people are big fans of ***[UNK1]***. … |

# Result: Language Modeling

- Dynamic modeling improves perplexity
  - Especially when entities reappear

| | | All tokens | | Reappearing entities | | Tokens following them | |
|---|---|---|---|---|---|---|---|
| | Baseline | 64.8 | | 48.0 | | 128.6 | |
| +Dynamic | input only | 62.8 | | 42.4 | | 109.5 | |
| | output only | 62.5 | | 35.9 | | 129.0 | |
| | input & output | 60.7 | 4.1 | 34.0 | 14.0 | 106.8 | 21.8 |

# Result: Language Modeling

- Dynamic modeling improves perplexity
  - Especially when entities reappear

| | | All tokens | Reappearing entities | Tokens following them |
|---|---|---|---|---|
| | **Baseline** | 64.8 | 48.0 | 128.6 |
| +Dynamic | input only | 62.8 | 42.4 | 109.5 |
| | output only | 62.5 | 35.9 12.1 | 129.0 |
| | input & output | **60.7** 4.1 | **34.0** 14.0 | **106.8** 21.8 |

"... she is a big fan of [???]"
<John; guitarist>

# Result: Language Modeling

- Dynamic modeling improves perplexity
  - Especially when entities reappear

| | All tokens | Reappearing entities | Tokens following them |
|---|---|---|---|
| **Baseline** | 64.8 | 48.0 | 128.6 |
| +Dynamic input only | 62.8 | 42.4 | 109.5 ↓19.1 |
| +Dynamic output only | 62.5 | 35.9 | 129.0 |
| +Dynamic input & output | **60.7** ↓4.1 | **34.0** ↓14.0 | **106.8** ↓21.8 |

“… <John; guitarist> [???]”

# Result: Language Modeling

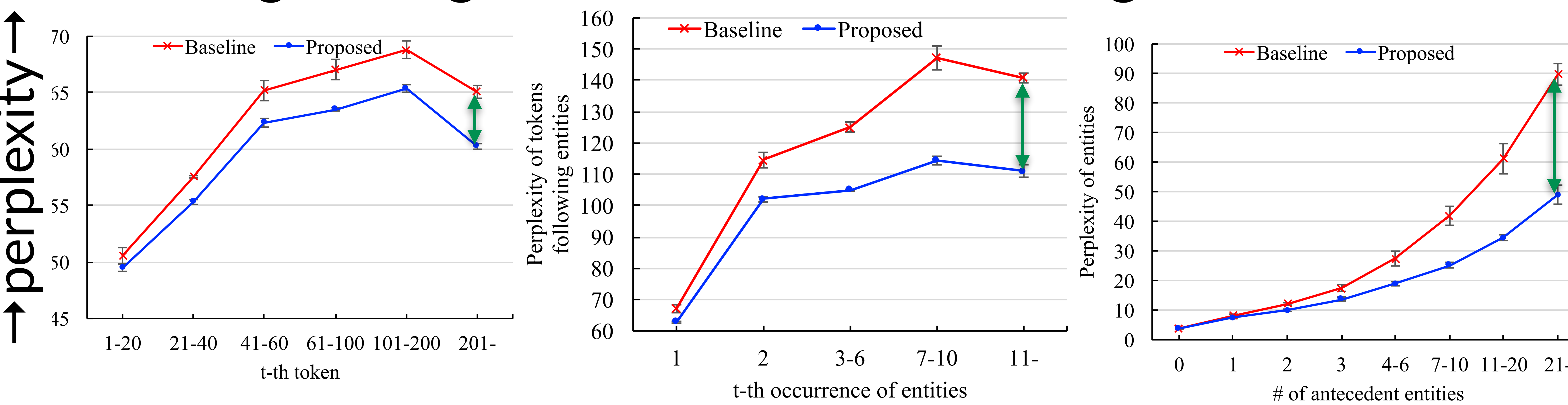
- Dynamic modeling improves perplexity
  - Especially when entities reappear

| | All tokens | Reappearing entities | Tokens following them |
|---|---|---|---|
| **Baseline** | 64.8 | 48.0 | 128.6 |
| +Dynamic **input only** | 62.8 | 42.4 | 109.5 |
| +Dynamic **output only** | 62.5 | 35.9 | 129.0 |
| +Dynamic **input & output** | **60.7** ↓*4.1* | **34.0** ↓*14.0* | **106.8** ↓*21.8* |

# Result: Language Modeling

- Dynamic modeling works well for long documents
  The latter of a document,
  The more often targets occur, the more improved
  The more targets occur,

- Organizing context is useful for long documents




→latter→ →more often→ →more targets→

# Summary

- Dynamic modeling of word vectors improves language models
  - For prediction of the unknowns
  - For prediction of tokens following the unknowns
- Future work
  - Story generation with organizing entities
  - Joint modeling with coreference resolution
  - Joint modeling with character/subword vectors

# Result: Cloze Test

- Pseudo coreference resolution task
- Solve this task by calculating
  the sentence likelihood by filling in with each entity

> ***[UNK1]*** loves guitars.
> ***[UNK2]*** did not prefer music.
> But, many people are big fans of *[???]*. ...

- *Mean Quantile* (mean rank of answers)
  is improved **.525→.642** by dynamic modeling