Language Modeling

Attention Mechanisms for Extending Context-Awareness of LSTM

JURIK JURASKA

WITH: SARANGARAJAN PARTHASARATHY AND WILLIAM GALE

SUMMER 2018 INTERNSHIP (MICROSOFT, SUNNYVALE)

Outline

- Dataset
  - Properties
- Baselines
  - N-gram, RNN
- Self-attention
  - Vanilla
  - Multi-head
  - Gated
- Optimization
  - Gumbel softmax
- Evaluation
  - Quantitative
  - Visual

Motivation

- Traditional language models for ASR are sentence-level only
- Potential of the context beyond single sentences
  - Paragraphs, documents, meeting transcriptions, etc.
- Limited reach of LSTMs back in time
  - Mechanism to make use of additional context (e.g. memory, attention)
- Interpretability of the neural model

Motivation

- *"Dotted lines that appear in an **Excel worksheet** usually represent page breaks, which are displayed so you can see how adjustments to your document affect the printed copy. If you find the lines annoying, you can turn them off in the **Excel** options. Dotted lines also may be used as cell borders to separate groups of data."*

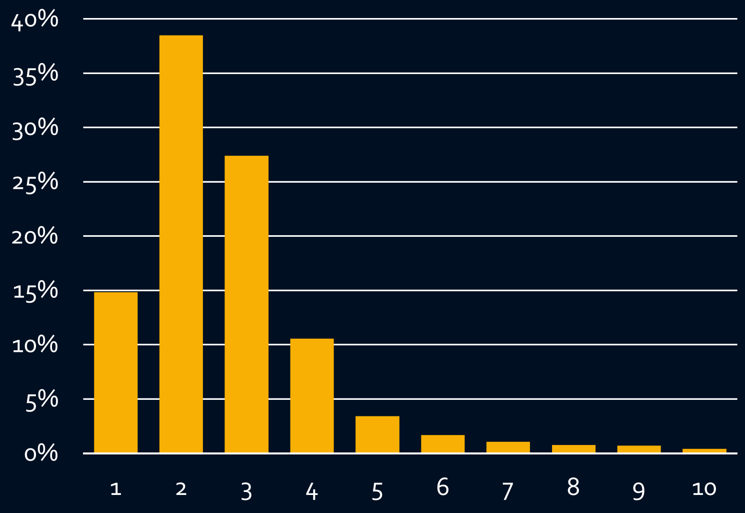
Outline

- Dataset
  - Properties
- Baselines
  - N-gram, RNN
- Self-attention
  - Vanilla
  - Multi-head
  - Gated
- Optimization
  - Gumbel softmax
- Evaluation
  - Quantitative
  - Visual

Dataset: *Properties*

- Web search questions and answers
  - 6.5M questions and answers
  - 17.5M sentences in answers (~ 278M words)
- Answers simulate context
  - All sentences related to the same question
- Vocabulary: 240K (out of original 1.5M)
  - Removed words with frequency < 10

Number of sentences in answers



Dataset: *Example*

- Question:

  "how far does light go in a day"

- Answer:

  "A light year does not travel, because it is a distance (= how far light travels in a vacuum in a year). In a day, light will travel a distance of 1 light-year divided by about 365.25 . Light travels about 300 thousand km per sec, and there are 86400 seconds in a day."

Outline

- Dataset
  - Properties
- **Baselines**
  - N-gram, RNN
- Self-attention
  - Vanilla
  - Multi-head
  - Gated
- Optimization
  - Gumbel softmax
- Evaluation
  - Quantitative
  - Visual

Baselines

- N-gram model:
  - Knesser-Ney interpolated, n = 5
- RNN model:
  - Single-layer LSTM
  - Embedding dimension: 512
  - Hidden state dimension: 2,048

Outline

- Dataset
  - Properties
- Baselines
  - N-gram, RNN
- **Self-attention**
  - Vanilla
  - Multi-head
  - Gated
- Optimization
  - Gumbel softmax
- Evaluation
  - Quantitative
  - Visual

Standard Attention (Luong et al., 2015)

- Typically in seq-to-seq models
- Calculate query-context alignment
  - Alignment vector a_t
- Transform to probabilities (softmax)
- Weighted average of the context
  - Context vector c_t
- $h_t{}^* = tanh(W_c[c_t; h_t])$

Self-attention (Mei et al., 2017)

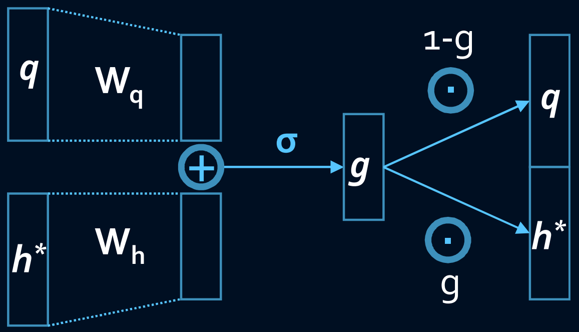
- Single-sequence scenario
- Instead of attending to encoder outputs, attends to all previous time steps
- On top of the LSTM
- Requires masking of the future hidden states

Multi-head Self-attention (Vaswani et al., 2017)

- V, K, Q are the same
  - Hidden states of LSTM

Gated Self-attention

- Motivated by the self-attention being skewed towards the first token
  - Gating of the attention outputs could counteract the skewness
- Gate computed from two vectors:
  1. query + context
  2. query + h_{t-1}
  3. **query + attention output**



Outline

- Dataset
  - Properties
- Baselines
  - N-gram, RNN
- Self-attention
  - Vanilla
  - Multi-head
  - Gated
- **Optimization**
  - Gumbel softmax
- Evaluation
  - Quantitative
  - Visual

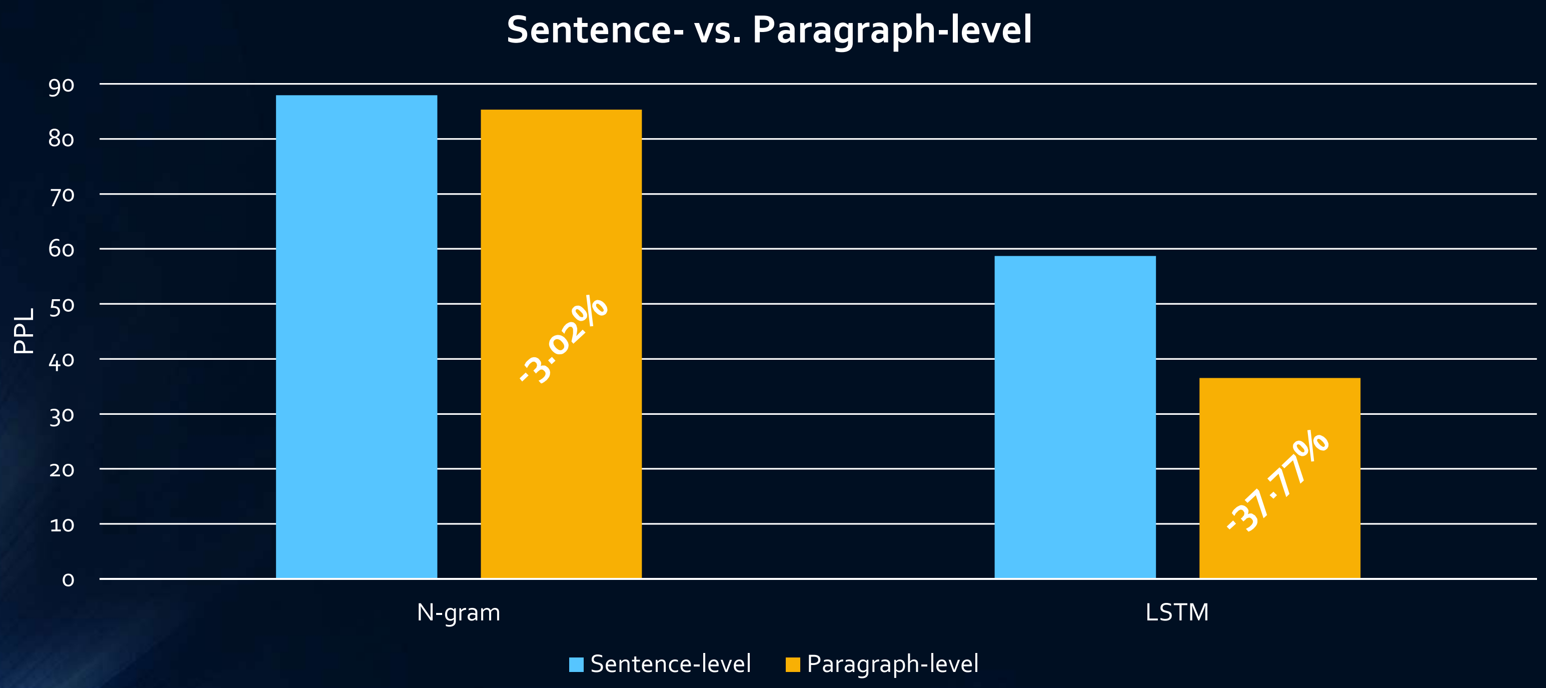
Sparse Attention: *Gumbel-Softmax*

- Issues with soft self-attention
  - Probability mass spread over number of words proportional to the index of the target word in the sentence
  - Leads to noisy attention which is less interpretable
- Sparse attention
  - Force the model to attend to most relevant words in the past
  - One approach to encourage sparsity
    - Gumbel-Softmax distribution

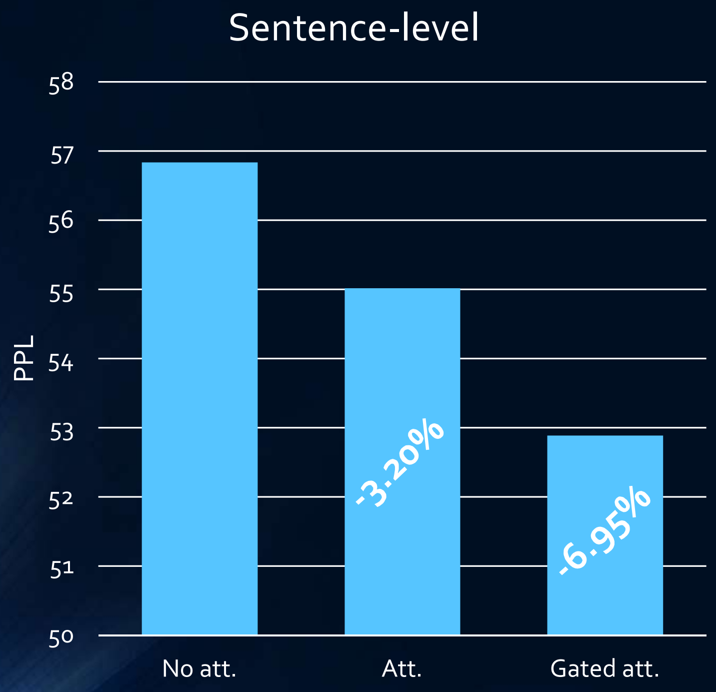
Outline

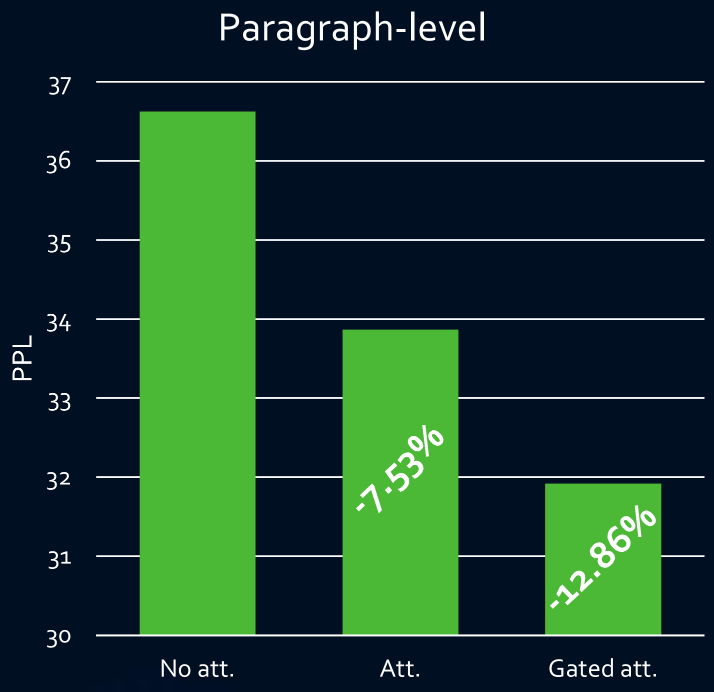
- Dataset
  - Properties
- Baselines
  - N-gram, RNN
- Self-attention
  - Vanilla
  - Multi-head
  - Gated
- Optimization
  - Gumbel softmax
- **Evaluation**
  - Quantitative
  - Visual

Evaluation: *N-gram vs. LSTM*

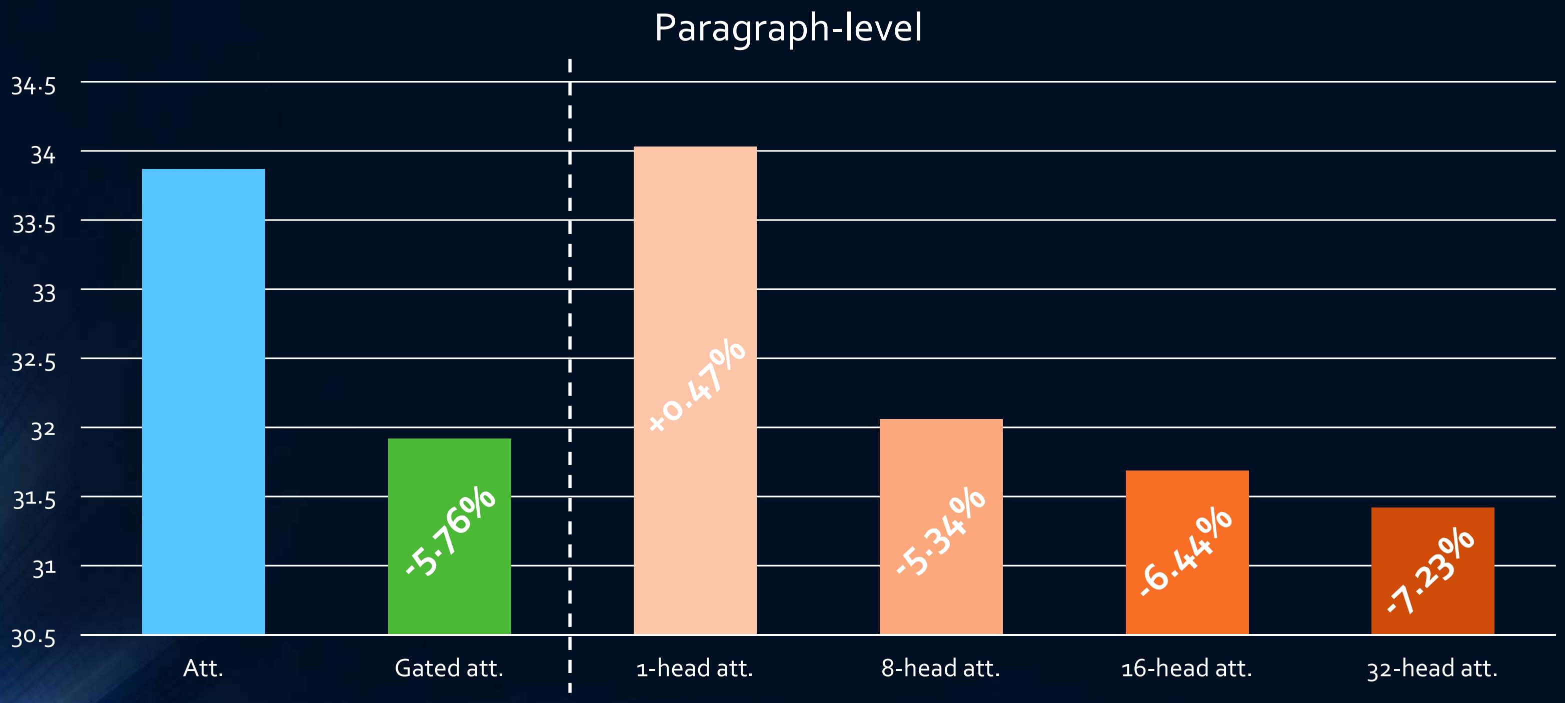


Evaluation: *LSTM with Attention*

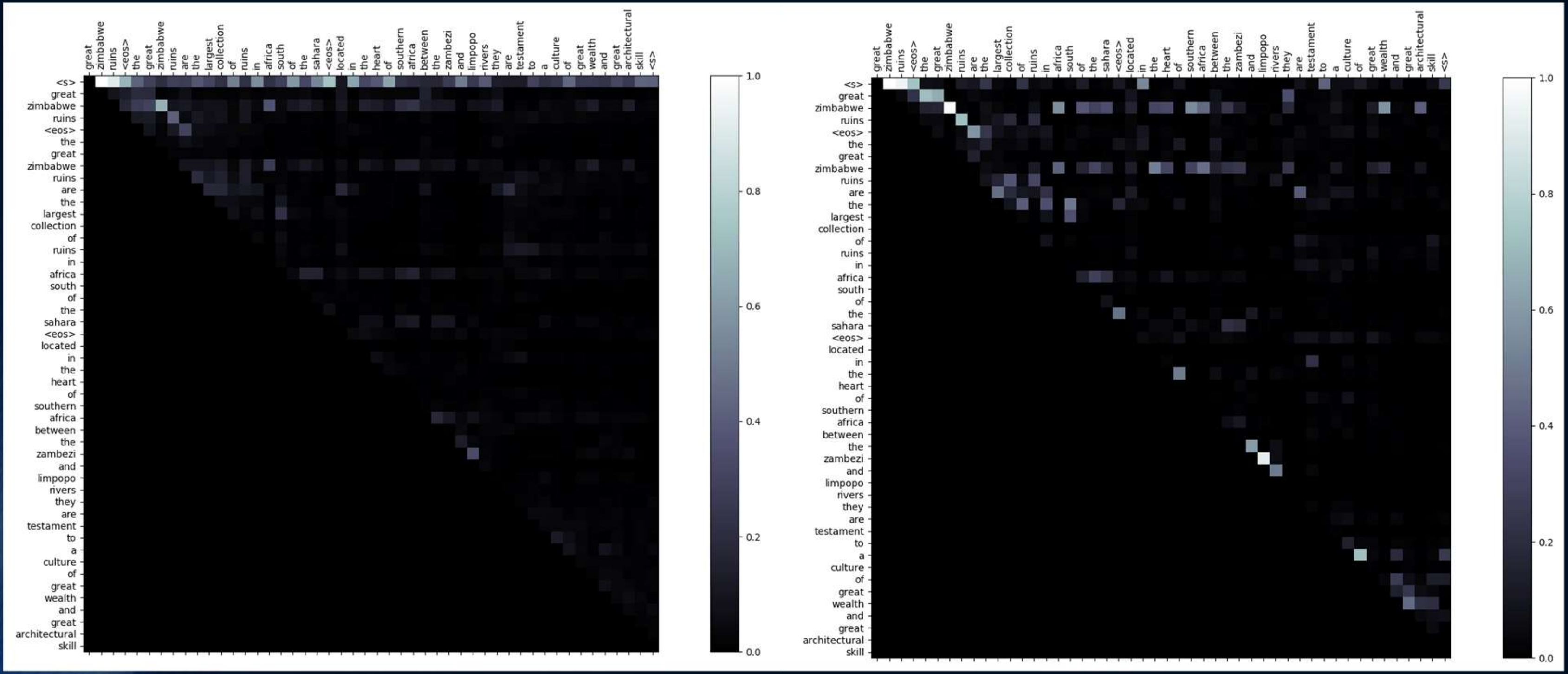

Sentence-level


Paragraph-level

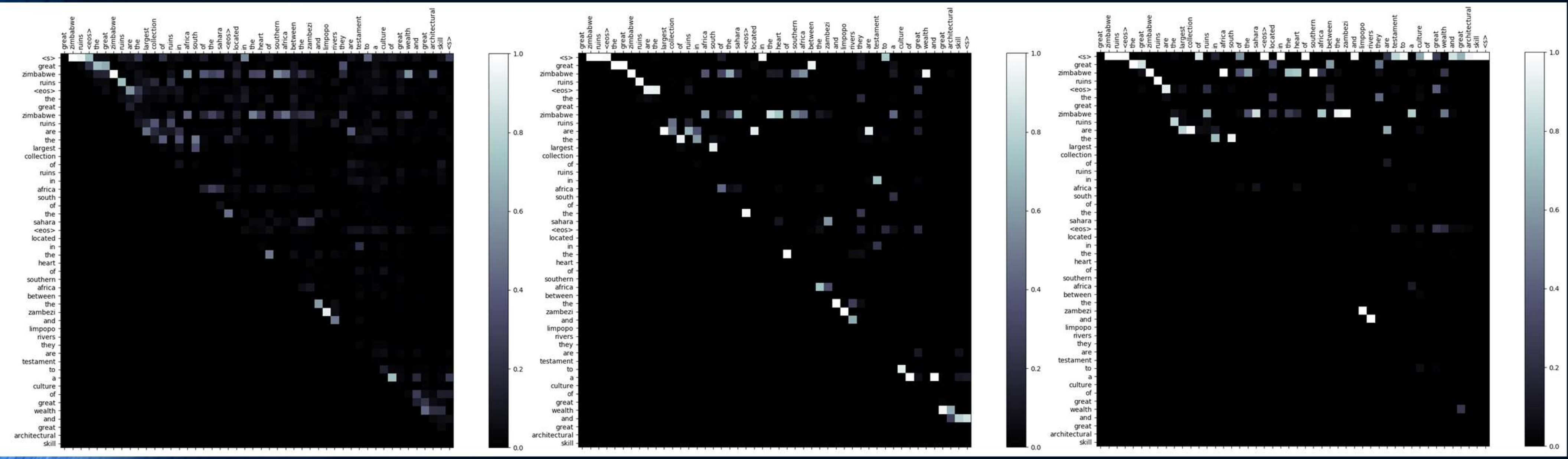
Evaluation: *Multi-head Attention*

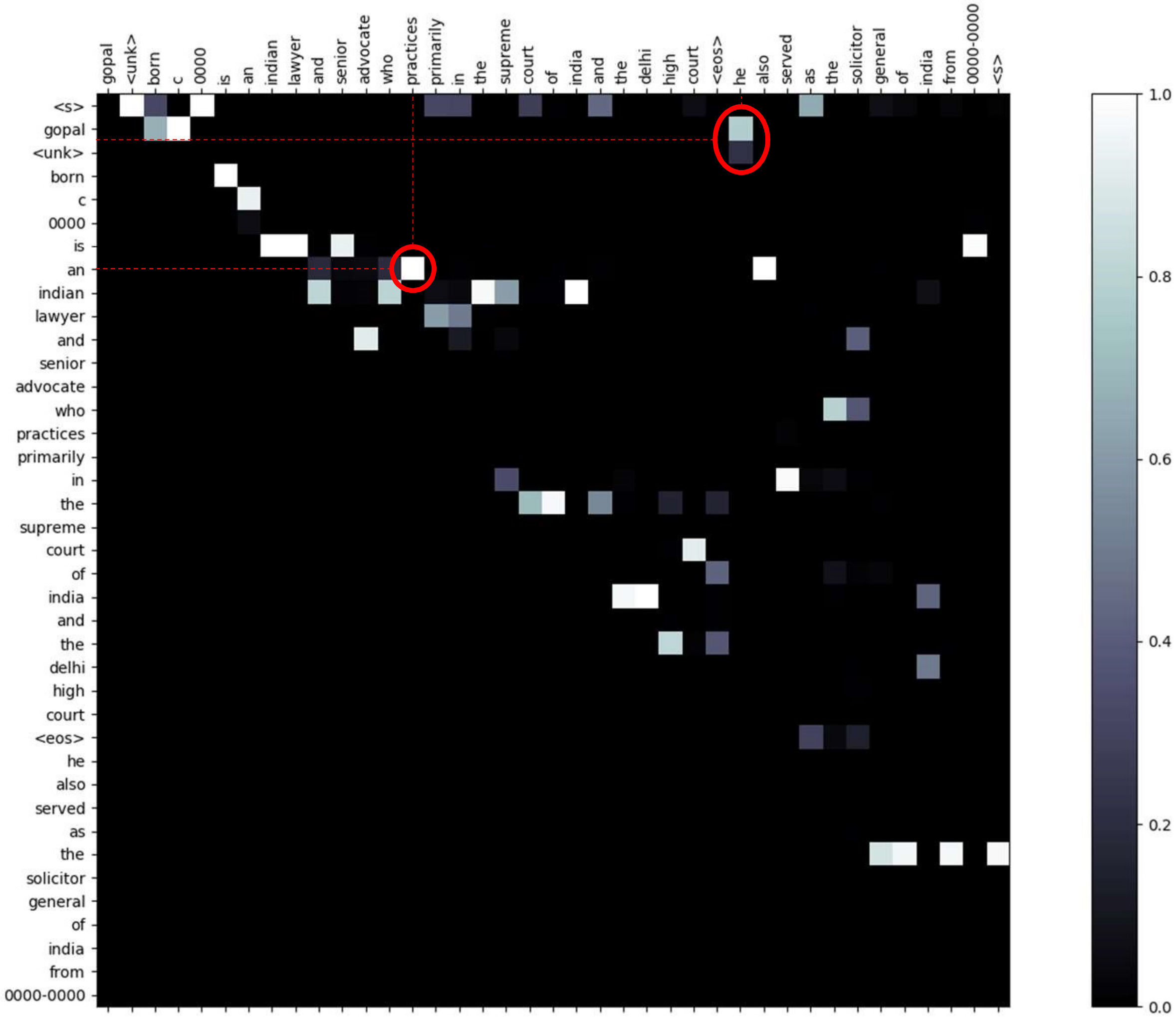


Evaluation: *Att. x Gated Att.*



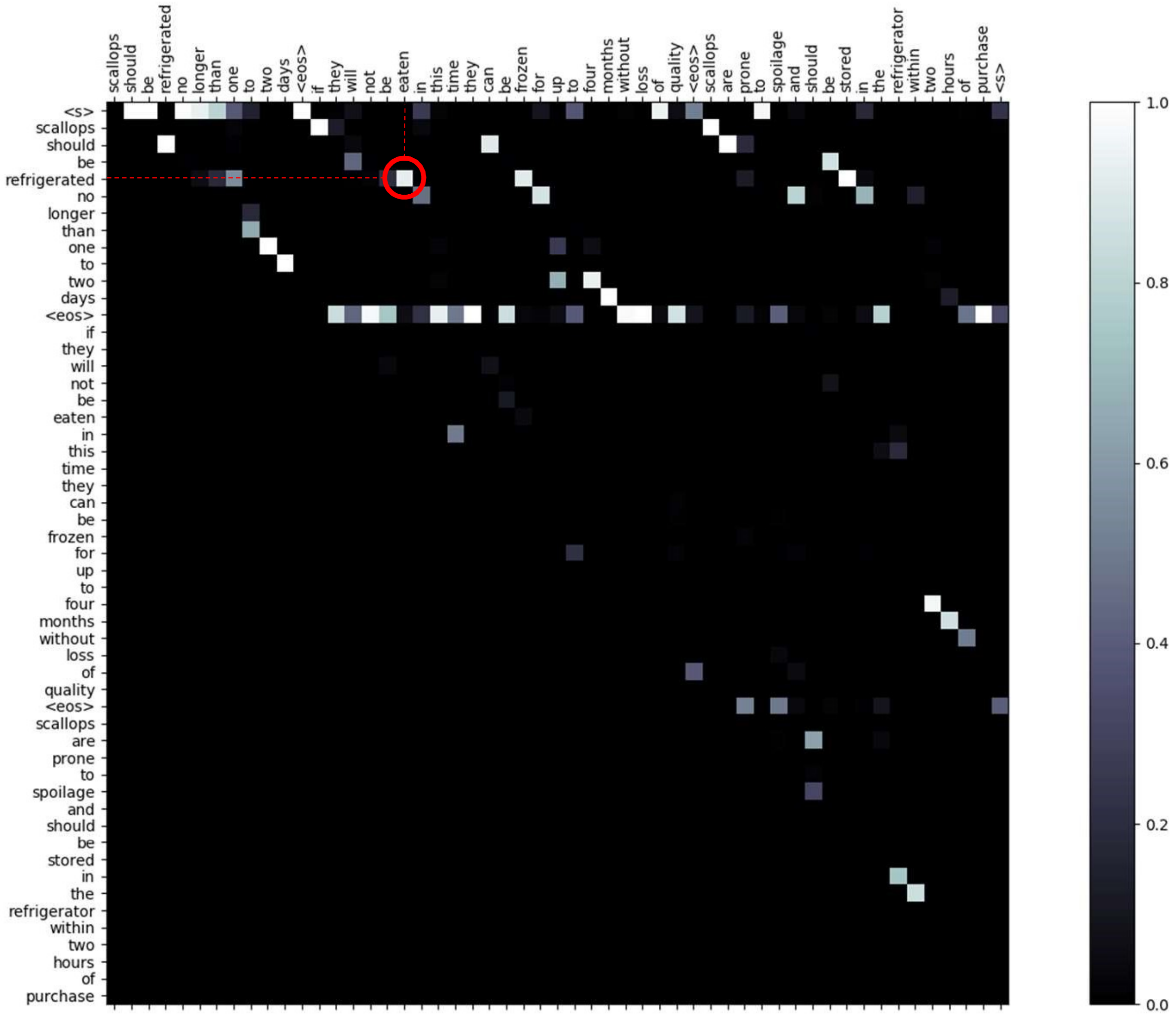
Evaluation: *Gated x Gated+Gumbel x Gumbel*




gopal
<unk>
born
c
0000
is
an
indian
lawyer
and
senior
advocate
who
practices
primarily
in
the
supreme
court
of
india
and
the
delhi
high
court
<eos>
he
also
served
as
the
solicitor
general
of
india
from
0000-0000
<s>
<s>
gopal
<unk>
born
c
0000
is
an
indian
lawyer
and
senior
advocate
who
practices
primarily
in
the
supreme
court
of
india
and
the
delhi
high
court
<eos>
he
also
served
as
the
solicitor
general
of
india
from
0000-0000
1.0
0.8
0.6
0.4
0.2
0.0

eaten <- refrigerated
PMI = 6.04

eaten <- scallops
PMI = 8.74

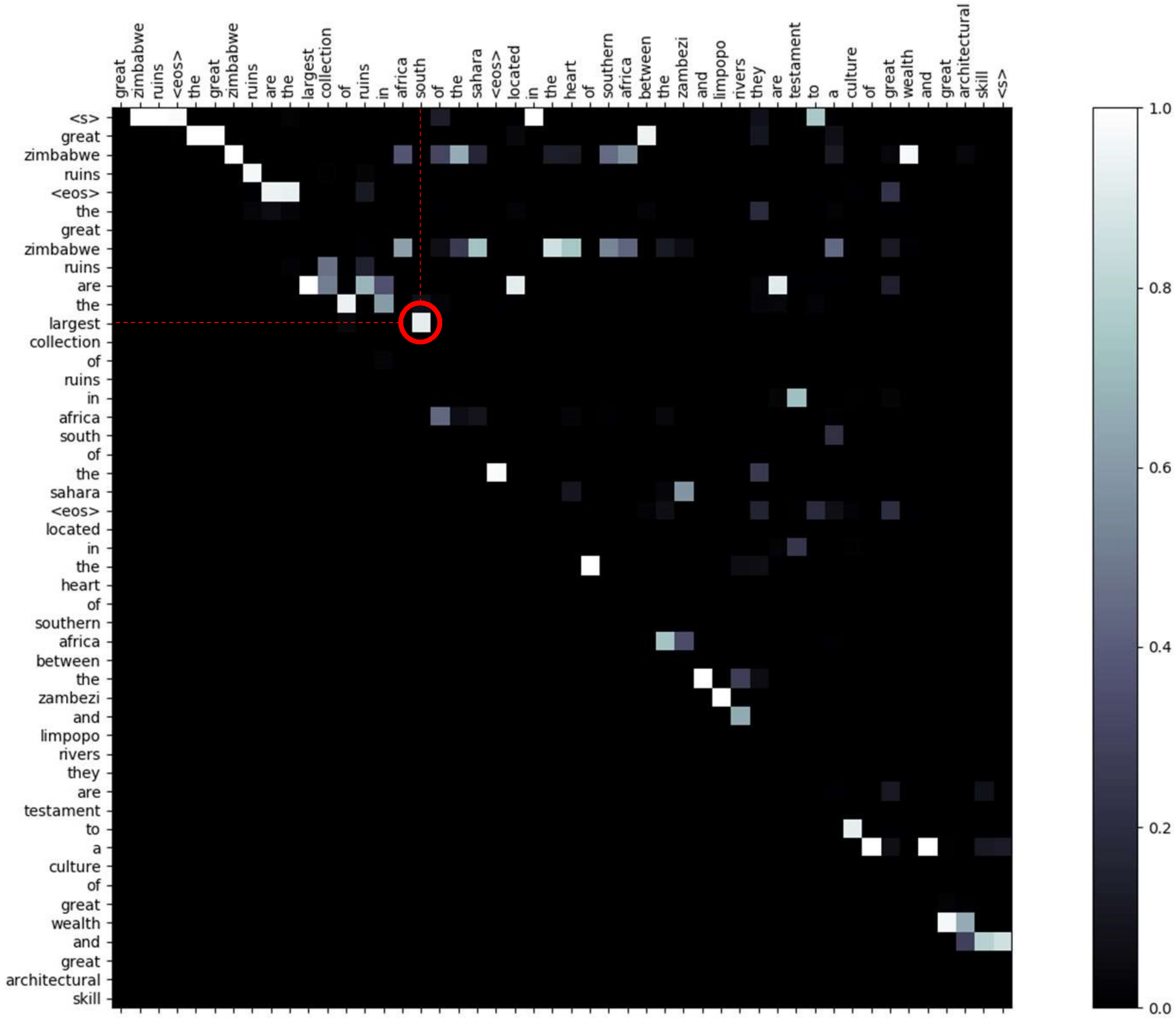


south <- largest
PMI = 1.75

south <- africa
PMI = 5.40

south <- zimbabwe
PMI = 4.17

south <- ruins
PMI = 3.50



Evaluation: *Attention Reach & Focus*

Distance	"plant"	"cell"	"sell"
<= 5	animal, nuclear,…	nucleus, phone, laptops,…	manufacture, purchase,…
<= 10	plants, station,…	lymphoma, rows, workbook,…	seller, pray,…
<= 20	seed, species,…	cells, body, records,…	selling, service,…
<= 50	seeds, tomato,…	cancer, mitochondria,…	offer, compensating,…
> 50	ivy, lettuce,…	cell, formula,…	sell, buy,…

Conclusion

- Paragraph-level language modeling is useful
  - Significant context carryover across sentence boundaries
  - N-gram language models unable to capture much context
  - Baseline LSTM language models provide significant gains
- Attention models enhance (long-term) context-awareness of LSTM
  - Reasonable perplexity gains
  - Interpretable word associations
  - Sparse attention potentially useful