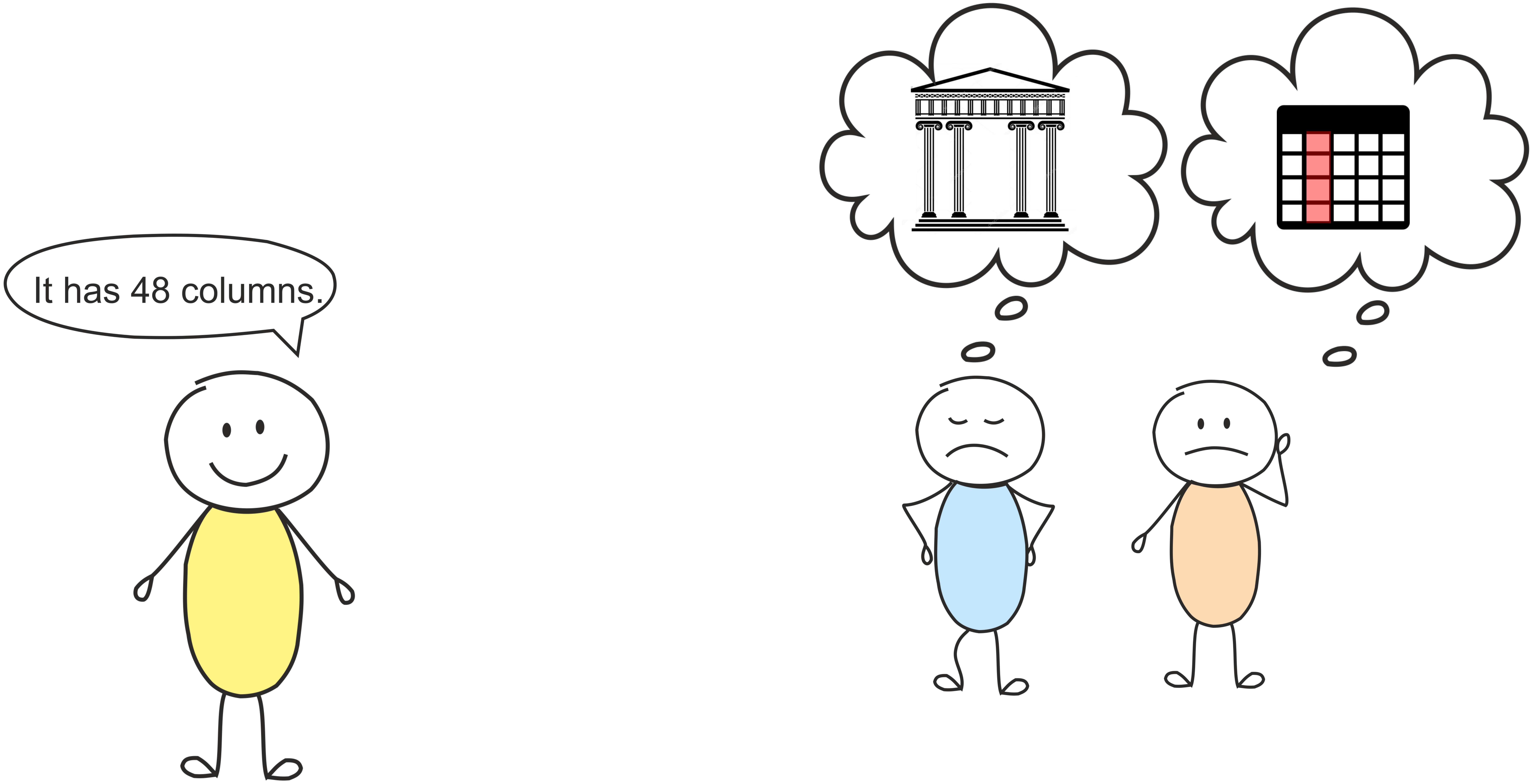
Context-Aware Neural Machine Translation Learns Anaphora Resolution

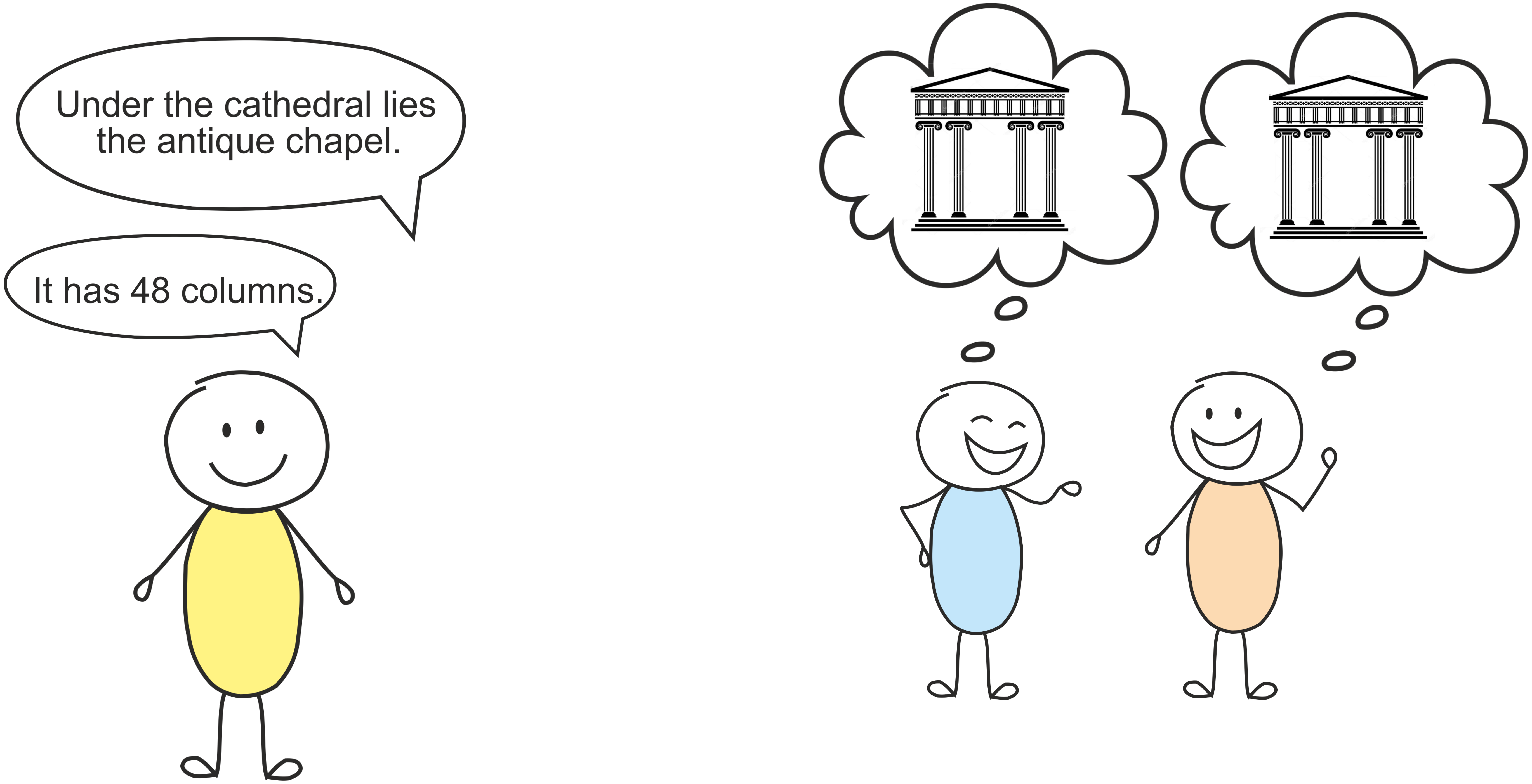
Yandex
Universiteit van Amsterdam
The University of Edinburgh

Elena Voita, Pavel Serdyukov, Rico Sennrich, Ivan Titov

Do we really need context?



Do we really need context?



Do we really need context?

Source:

- It has 48 columns.

Do we really need context?

Source:

- **It** has 48 columns.

What does “it” refer to?

Do we really need context?

Source:

- **It** has 48 columns.

Possible translations into Russian:

- У **него** 48 колонн. (masculine or neuter)
- У **нее** 48 колонн. (feminine)
- У **них** 48 колонн. (plural)

Do we really need context?

Source:

- It has 48 **columns**.

What do “columns” mean?

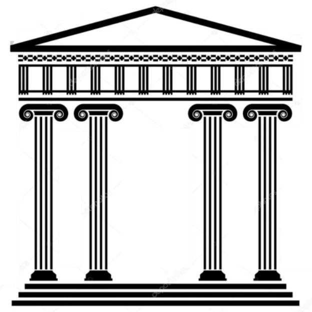
Do we really need context?

Source:

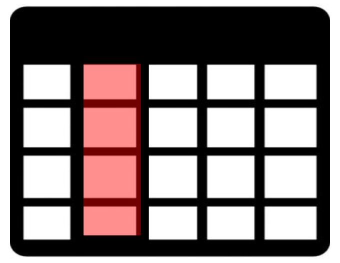
- It has 48 **columns**.

Possible translations into Russian:

- У него/нее/них 48 **колонн**.



- У него/нее/них 48 **колонок**.



Do we really need context?

Context:

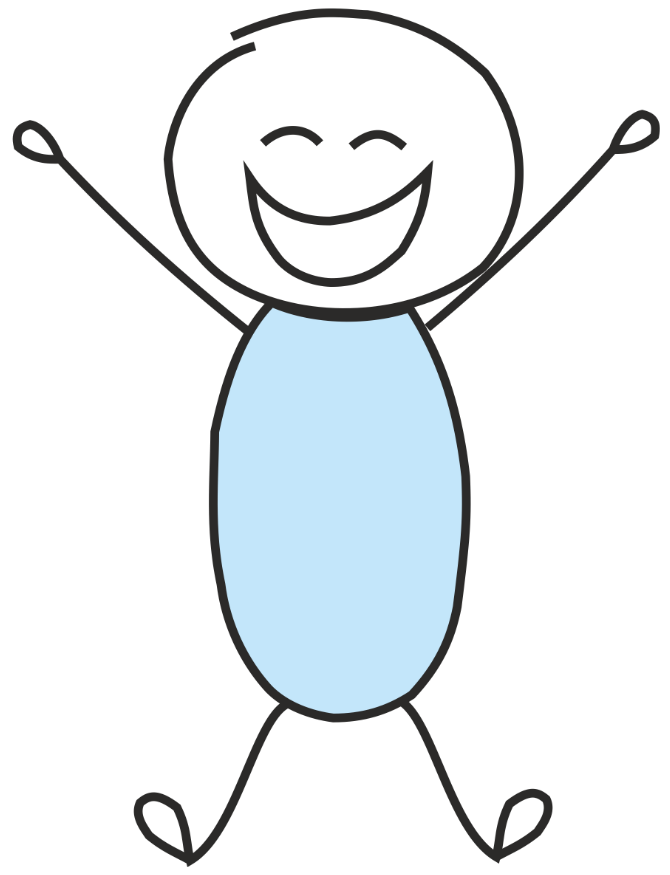
› Under the cathedral lies the antique chapel.

Source:

› It has 48 columns.

Translation:

› У нее 48 колонн.



Recap: antecedent and anaphora resolution

Under the cathedral lies the antique chapel It has 48 columns.

antecedent *anaphoric pronoun*

Wikipedia:

An ***antecedent*** is an expression that gives its meaning to a proform (pronoun, pro-verb, pro-adverb, etc.)

Anaphora resolution is the problem of resolving references to earlier or later items in the discourse.

Context in Machine Translation

SMT

- focused on handling specific phenomena
- used special-purpose features

  ([Le Nagard and Koehn, 2010]; [Hardmeier and Federico, 2010]; [Hardmeier et al., 2015], [Meyer et al., 2012], [Gong et al., 2012], [Carpuat, 2009]; [Tiedemann, 2010]; [Gong et al., 2011])

Context in Machine Translation

SMT

- focused on handling specific phenomena
- used special-purpose features

  ([Le Nagard and Koehn, 2010]; [Hardmeier and Federico, 2010]; [Hardmeier et al., 2015], [Meyer et al., 2012], [Gong et al., 2012], [Carpuat, 2009]; [Tiedemann, 2010]; [Gong et al., 2011])

NMT

- directly provide context to an NMT system at training time

  ([Jean et al., 2017]; [Wang et al., 2017]; [Tiedemann and Scherrer, 2017]; [Bawden et al., 2018])

Context in Machine Translation

SMT

- focused on handling specific phenomena
- used special-purpose features

  ([Le Nagard and Koehn, 2010]; [Hardmeier and Federico, 2010]; [Hardmeier et al., 2015], [Meyer et al., 2012], [Gong et al., 2012], [Carpuat, 2009]; [Tiedemann, 2010]; [Gong et al., 2011])

NMT

- directly provide context to an NMT system at training time

  ([Jean et al., 2017]; [Wang et al., 2017]; [Tiedemann and Scherrer, 2017]; [Bawden et al., 2018])

- not clear:

  what kinds of discourse phenomena are successfully handled

  how they are modeled

Plan

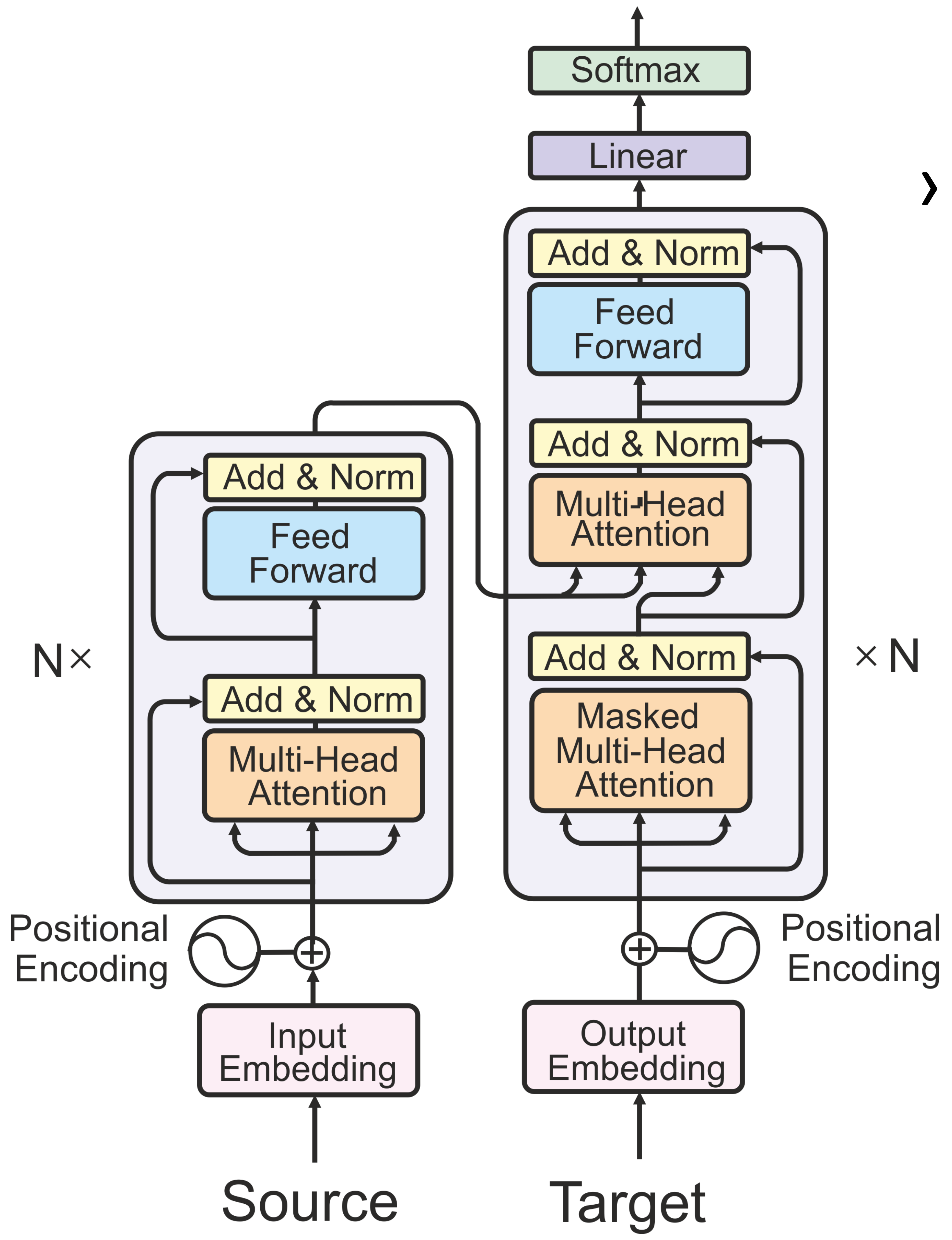
1 | Model Architecture

2 | Overall performance

3 | Analysis

Context-Aware Model Architecture

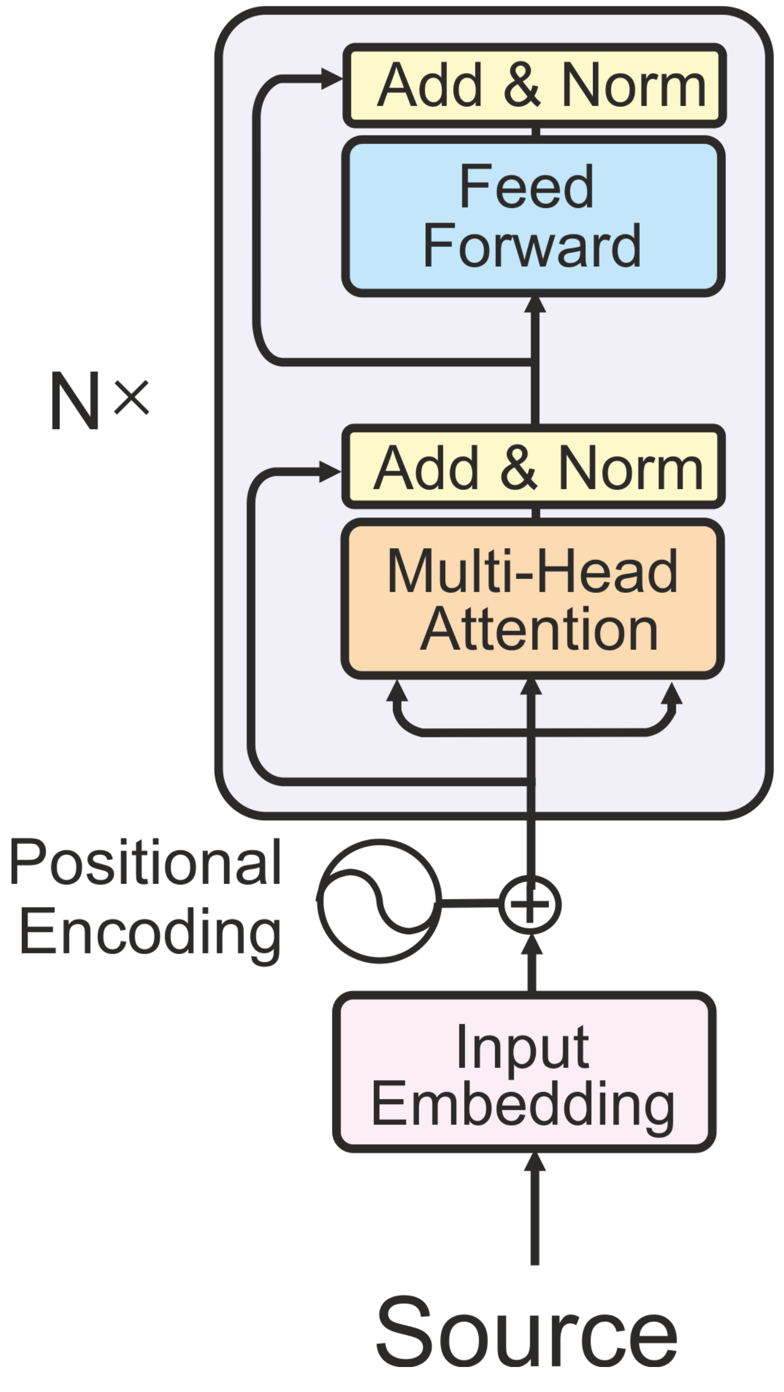
Transformer model architecture



- start with the Transformer

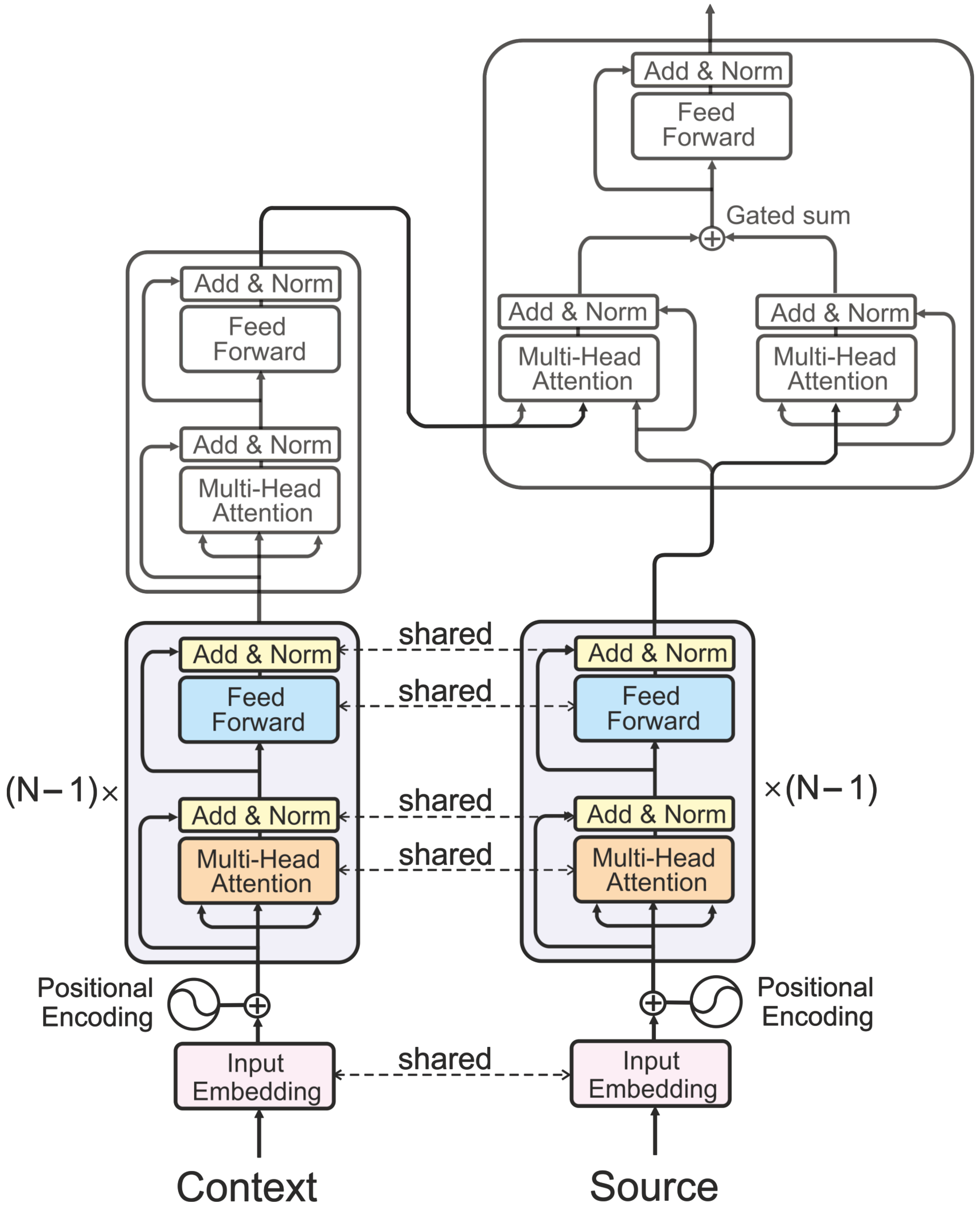
  [Vaswani et al, 2018]

Context-aware model architecture



- start with the Transformer [Vaswani et al, 2018]
- incorporate context information on the encoder side

Context-aware model architecture



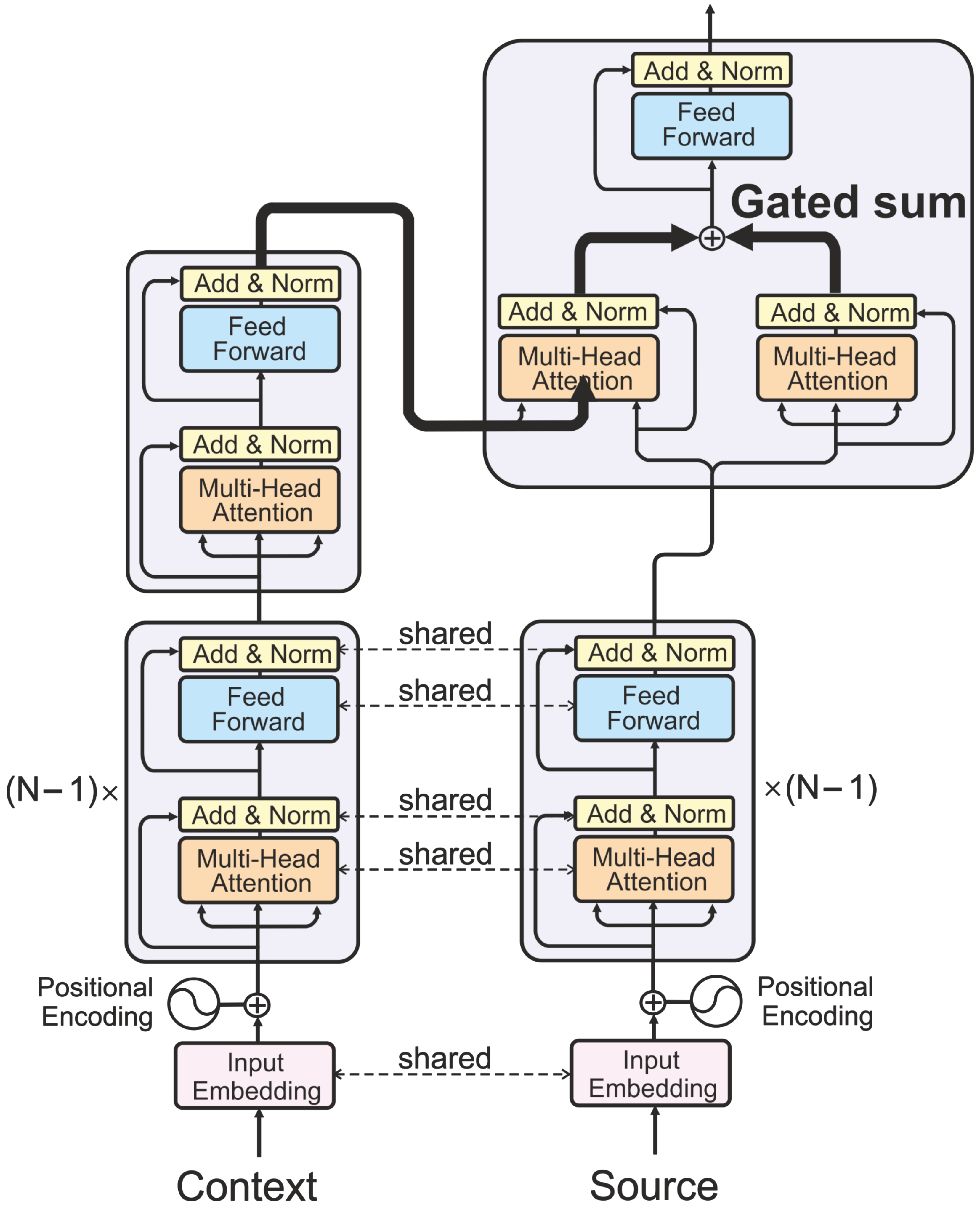
› start with the Transformer [Vaswani et al, 2018]

› incorporate context information on the encoder side

› use a separate encoder for context

› share first N-1 layers of source and context encoders

Context-aware model architecture



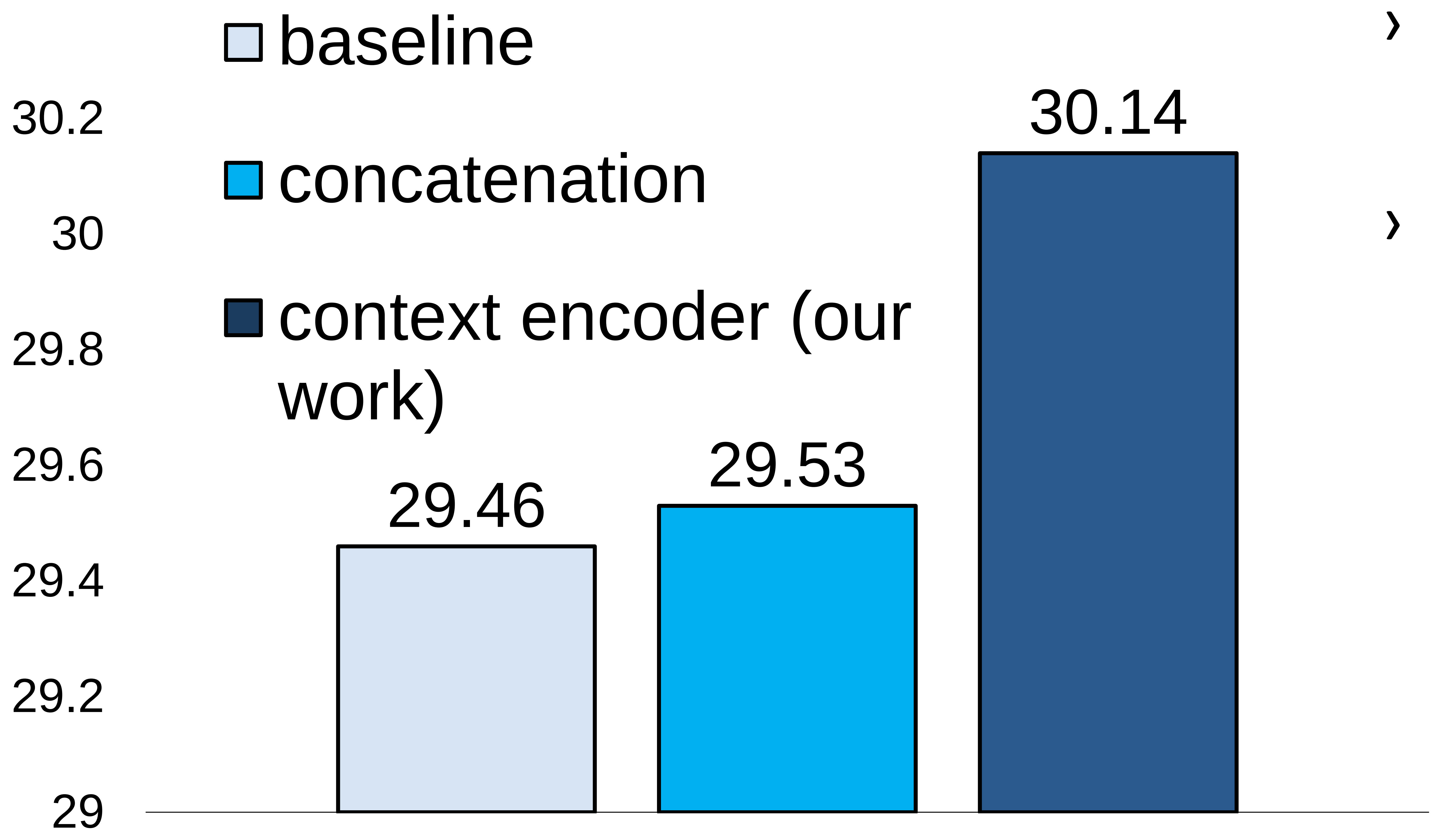
- start with the Transformer [Vaswani et al, 2018]
- incorporate context information on the encoder side
- use a separate encoder for context
- share first N-1 layers of source and context encoders
- the last layer incorporates contextual information

Overall performance

Dataset: OpenSubtitles2018 (Lison et al., 2018) for English and Russian

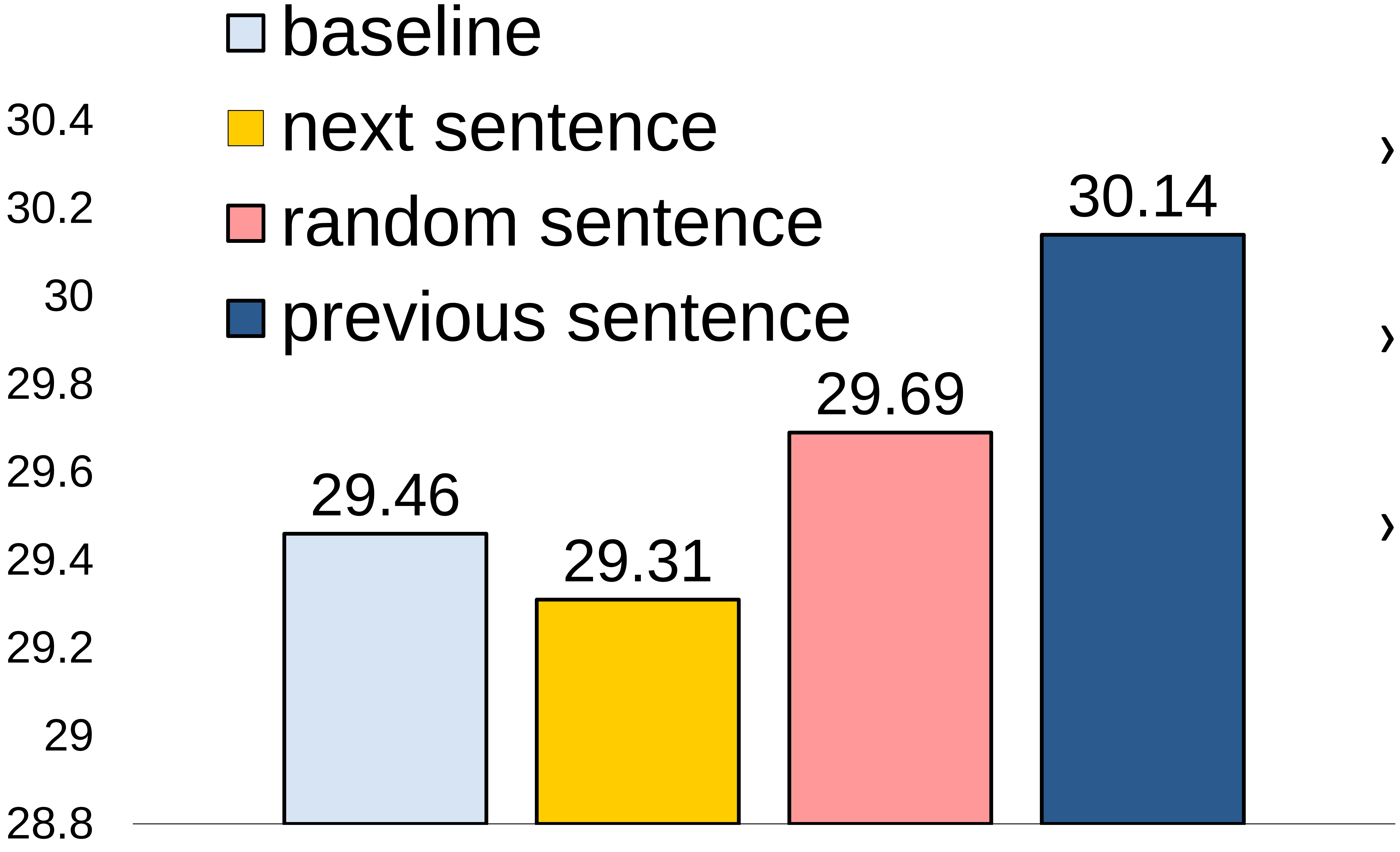
Overall performance: models comparison

(context is the previous sentence)



- baseline: context-agnostic Transformer
- concatenation: modification of the approach by [Tiedemann and Scherrer, 2017]

Our model: different types of context



(the only significant at p<0.01 difference is with the best model; differences between other results are not significant)

- Next sentence does not appear beneficial
- Performance drops for a random context sentence
- Model is robust towards being shown a random context sentence

Analysis

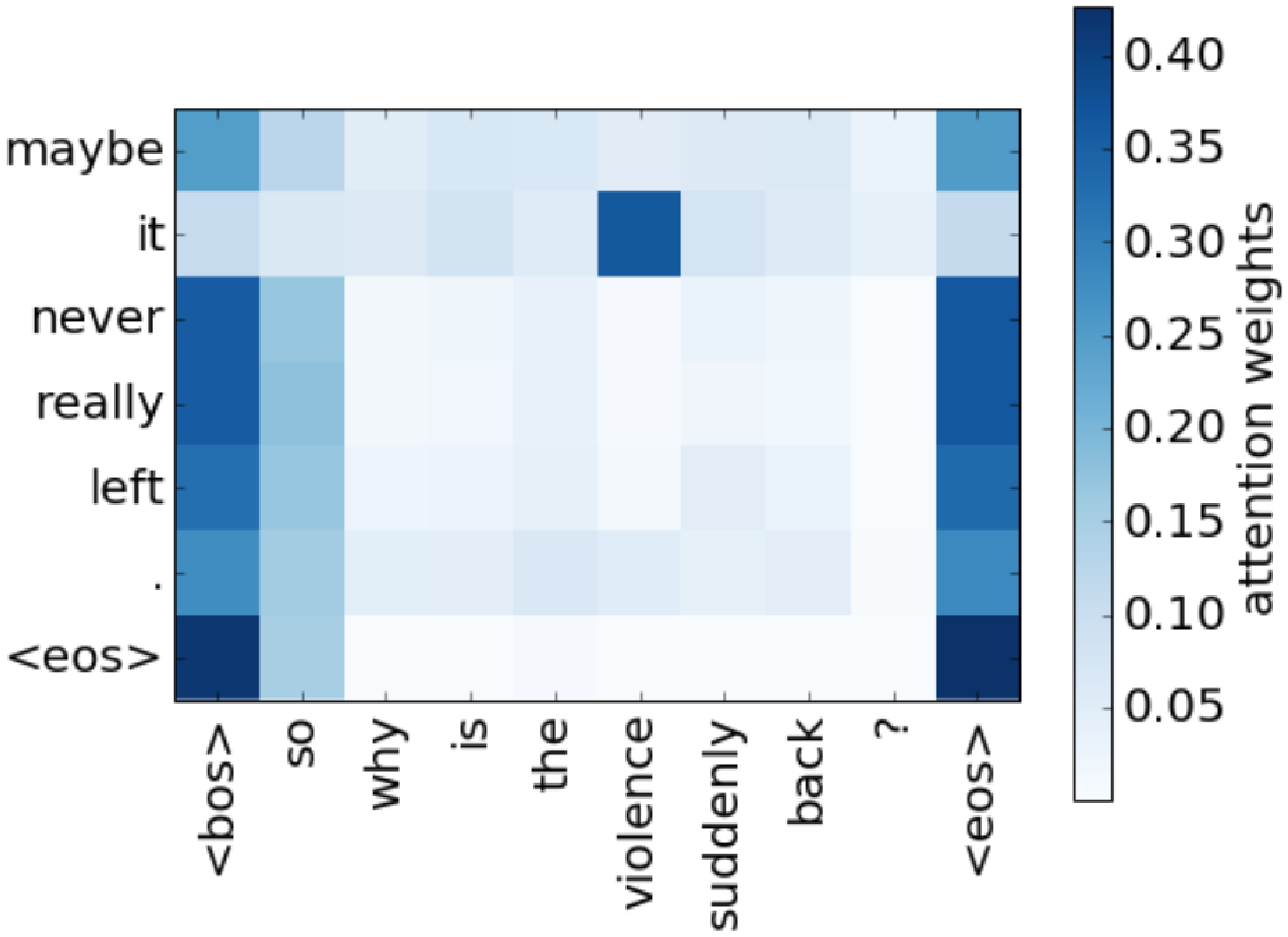
Analysis

1 | Top words influenced by context

2 | Non-lexical patterns affecting attention to context

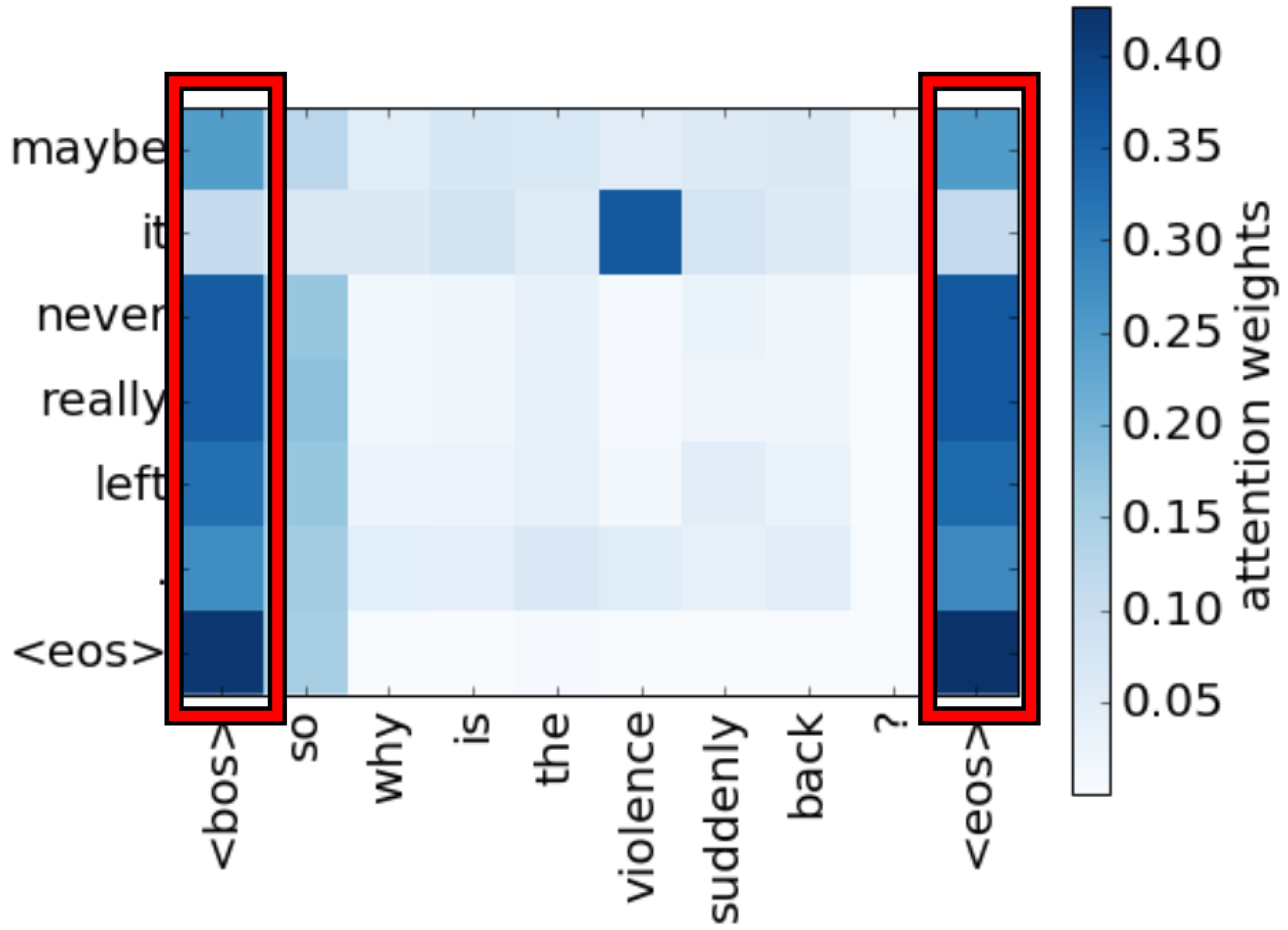
3 | Latent anaphora resolution

What do we mean by “attention to context”?



- attention from source to context
- mean over heads of per-head attention weights

What do we mean by “attention to context”?



- attention from source to context
- mean over heads of per-head attention weights
- take sum over context words (excluding <bos>, <eos> and punctuation)

Top words influenced by context

word	pos
it	5.5
yours	8.4
yes	2.5
i	3.3
yeah	1.4
you	4.8
ones	8.3
‘m	5.1
wait	3.8
well	2.1

Top words influenced by context

word	pos
it	5.5
yours	8.4
yes	2.5
i	3.3
yeah	1.4
you	4.8
ones	8.3
‘m	5.1
wait	3.8
well	2.1

Third person

- singular masculine
- singular feminine
- singular neuter
- plural

Top words influenced by context

word	pos
it	5.5
yours	3.4
yes	2.5
i	3.3
yeah	1.4
you	4.8
ones	8.3
‘m	5.1
wait	3.8
well	2.1

Second person

- singular impolite
- singular polite
- plural

Top words influenced by context

word	pos
it	5.5
yours	8.4
yes	2.5
i	3.3
yeah	1.4
you	4.8
ones	8.3
‘m	5.1
wait	3.8
well	2.1

Need to know gender, because verbs must agree in gender with “I” (in past tense)

Top words influenced by context

word	pos
it	5.5
yours	8.4
yes	2.5
i	3.3
yeah	1.4
you	4.8
ones	8.3
‘m	5.1
wait	3.8
well	2.1

Many of these words appear at sentence initial position.

Maybe this is all that matters?

Top words influenced by context

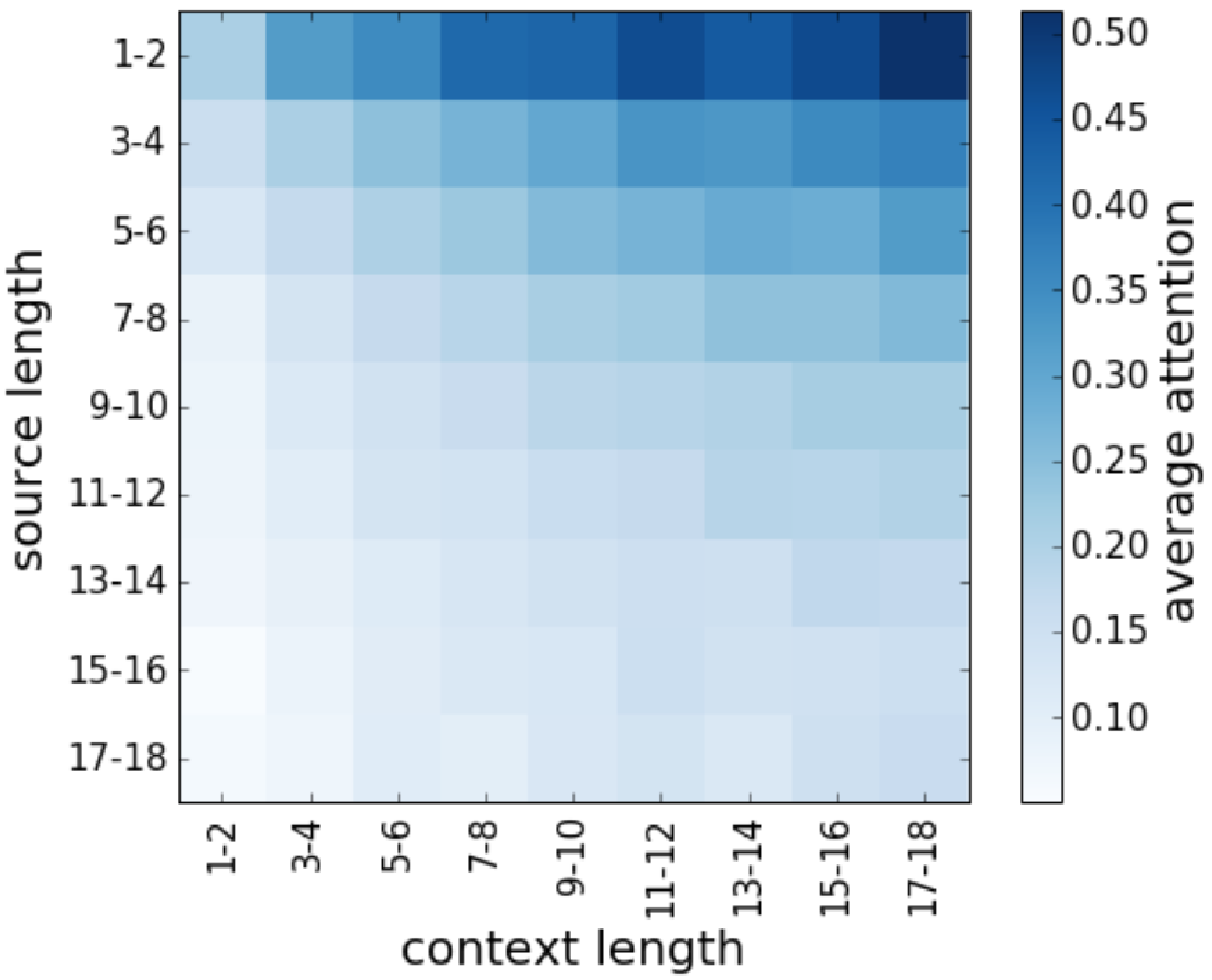
word	pos
it	5.5
yours	8.4
yes	2.5
i	3.3
yeah	1.4
you	4.8
ones	8.3
‘m	5.1
wait	3.8
well	2.1

Only positions after the first

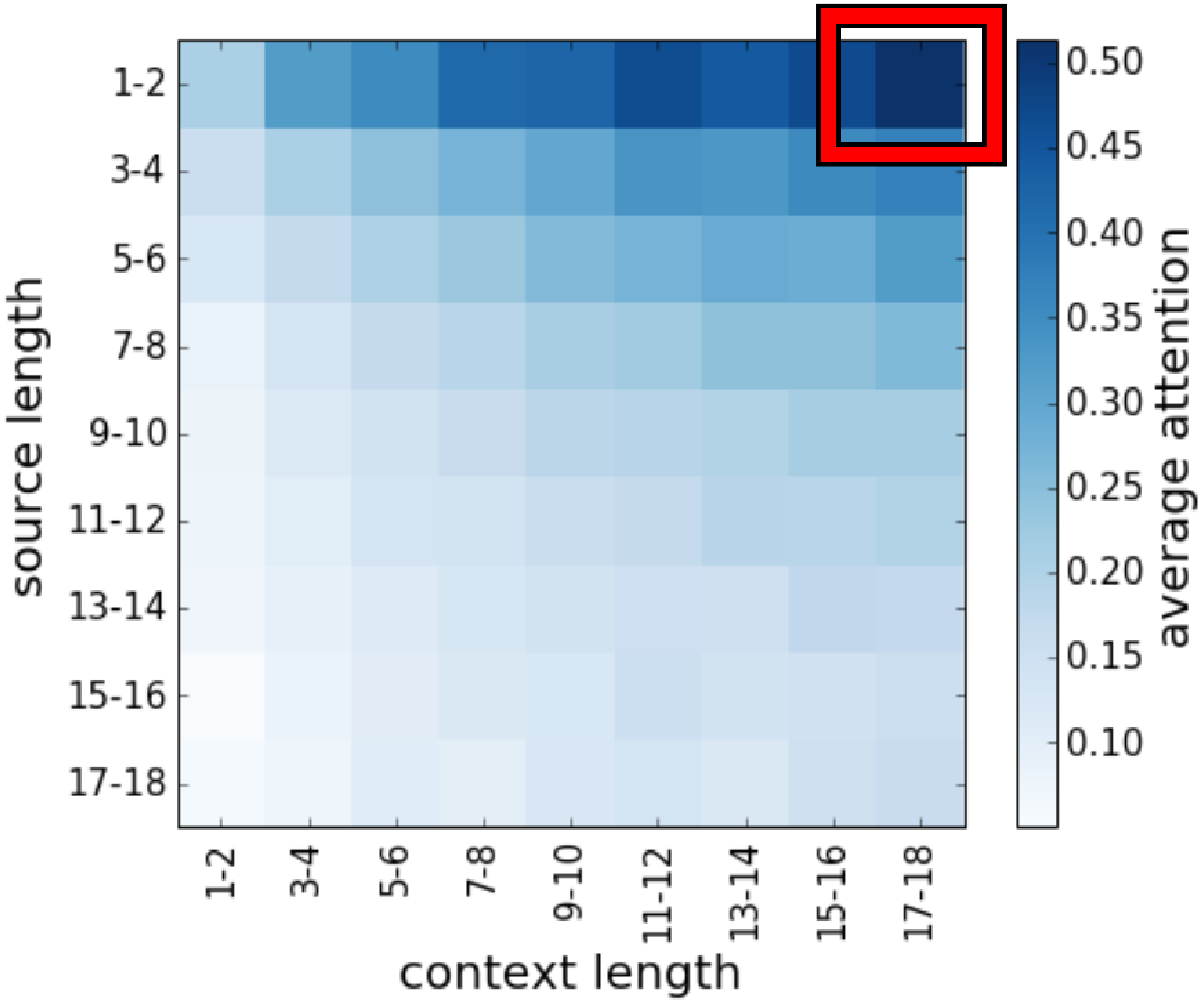
word	pos
it	6.8
yours	8.3
ones	7.5
‘m	4.8
you	5.6
am	4.4
i	5.2
‘s	5.6
one	6.5
won	4.6

Does the amount of attention to context depend on factors such as sentence length and position?

Dependence on sentence length



Dependence on sentence length



short source
long context

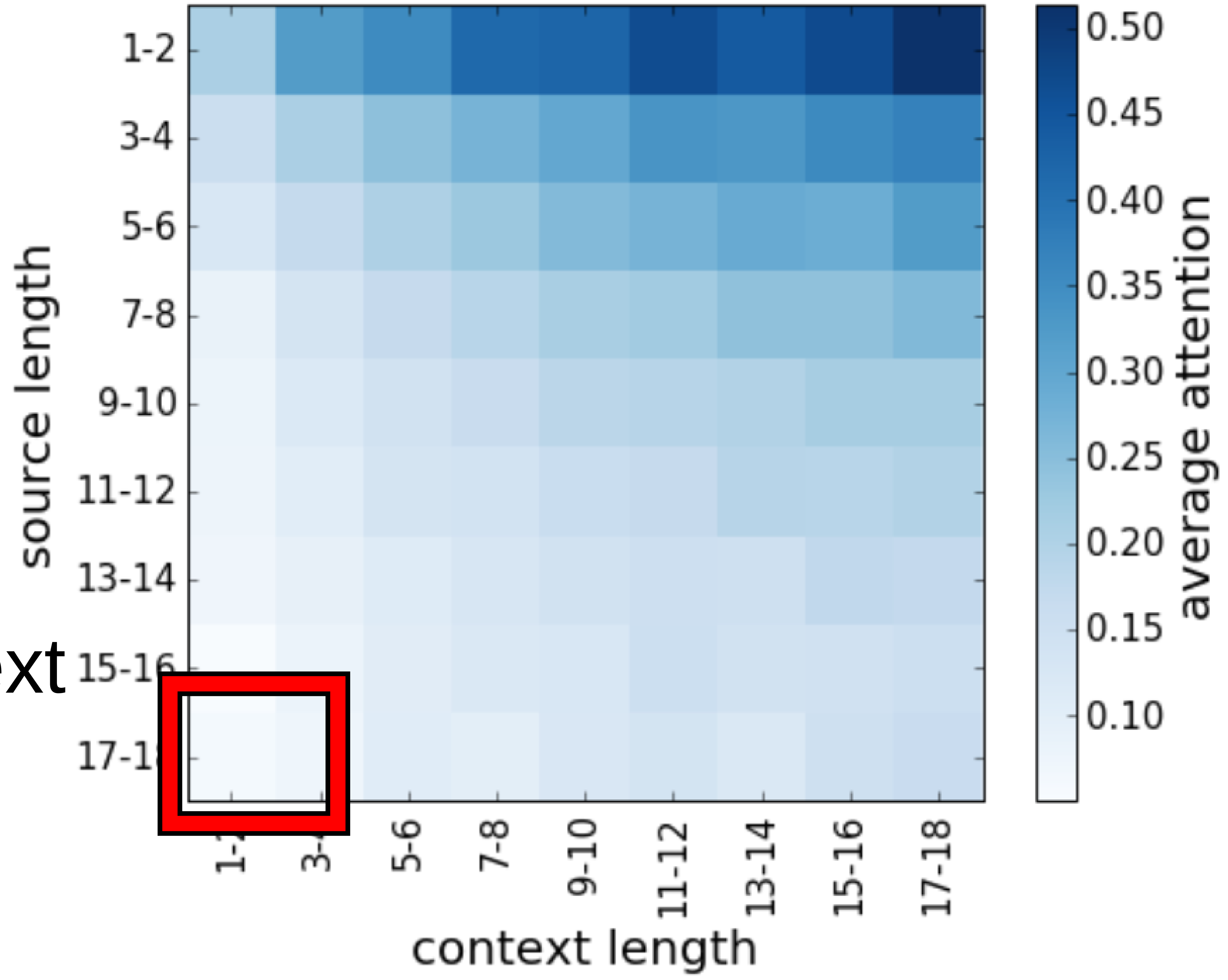
high attention to context

Dependence on sentence length

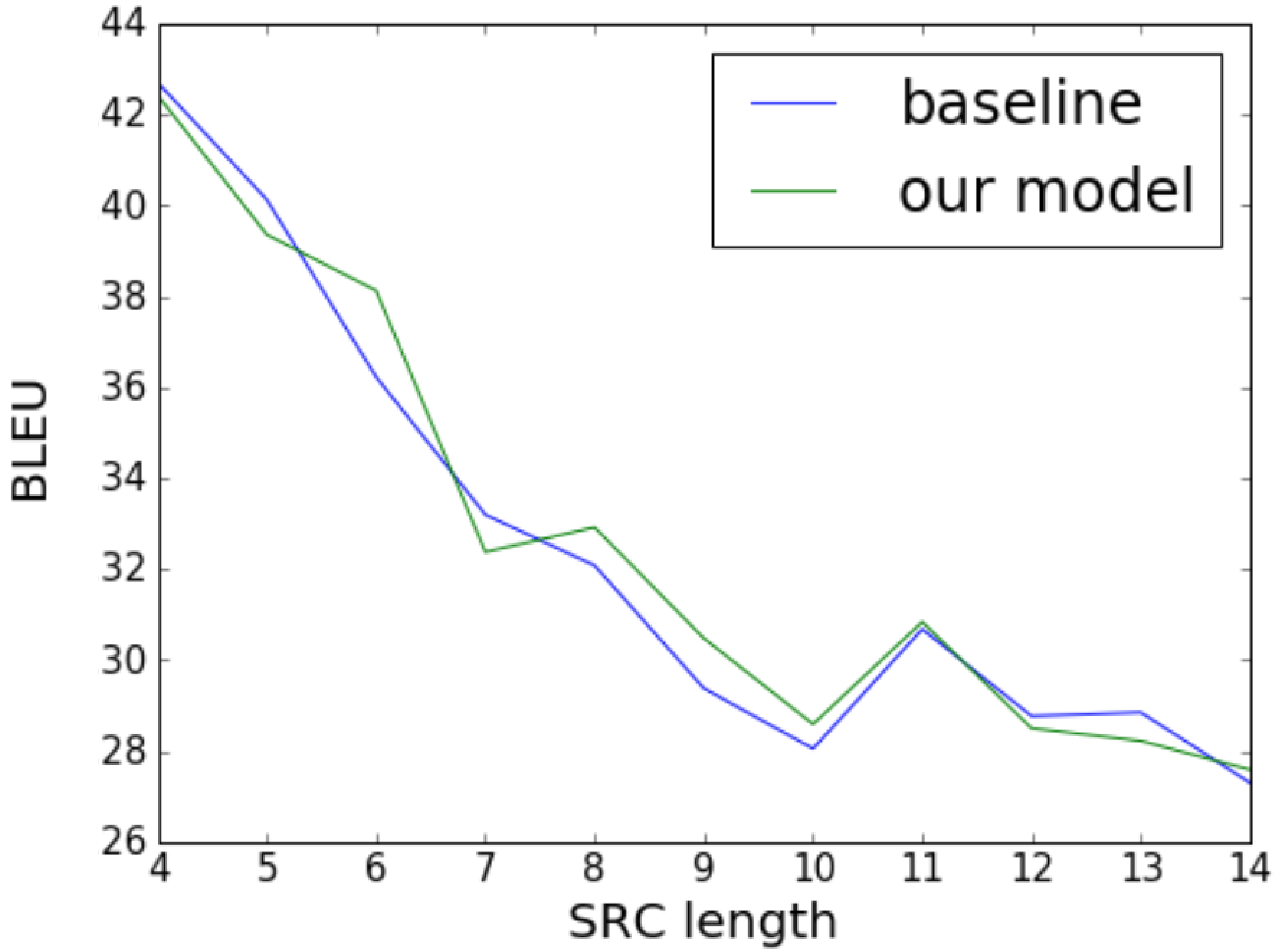
long source

short context

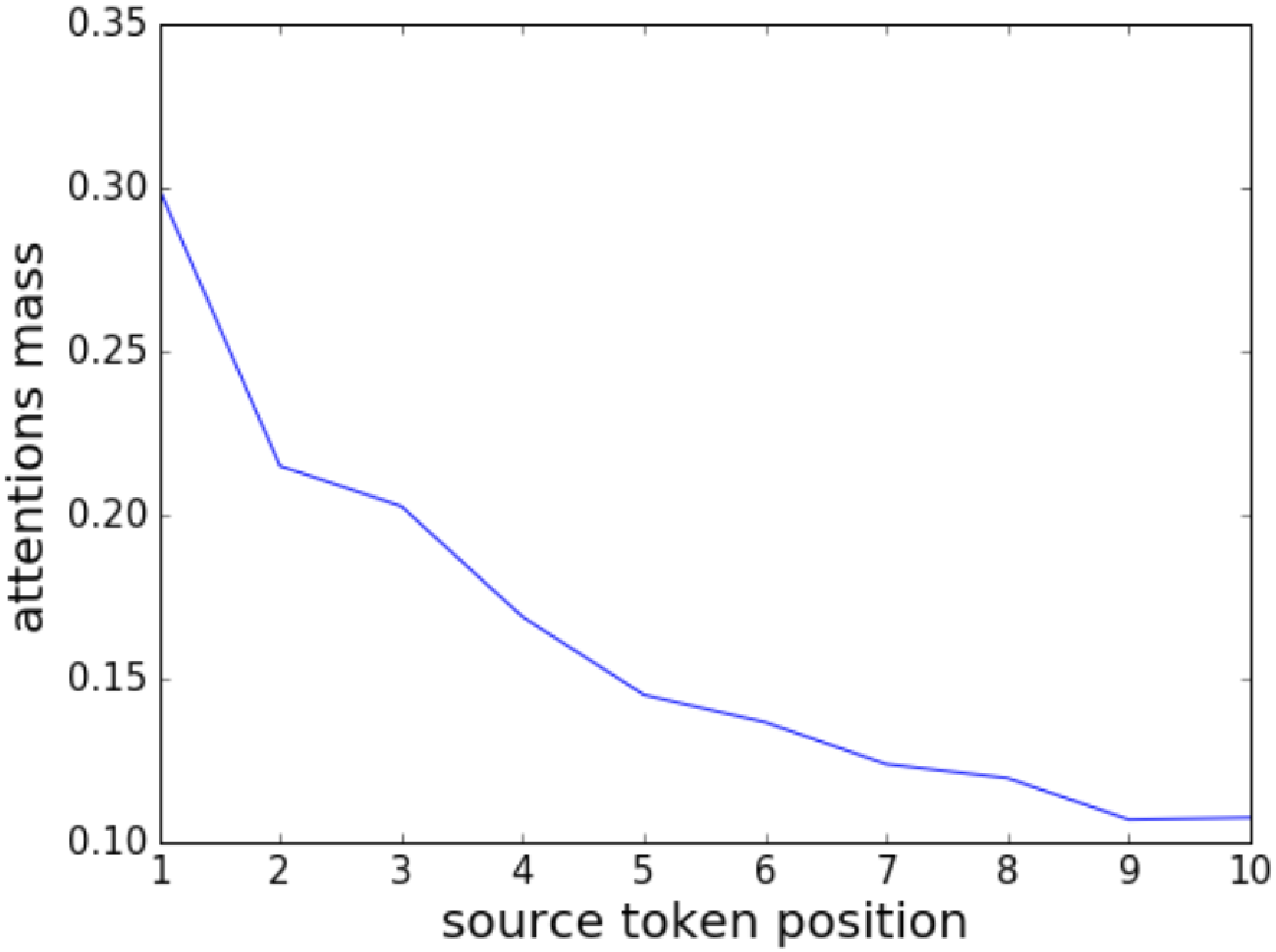
low attention to context



Is context especially helpful for short sentences?



Dependence on token position



Analysis of pronoun translation

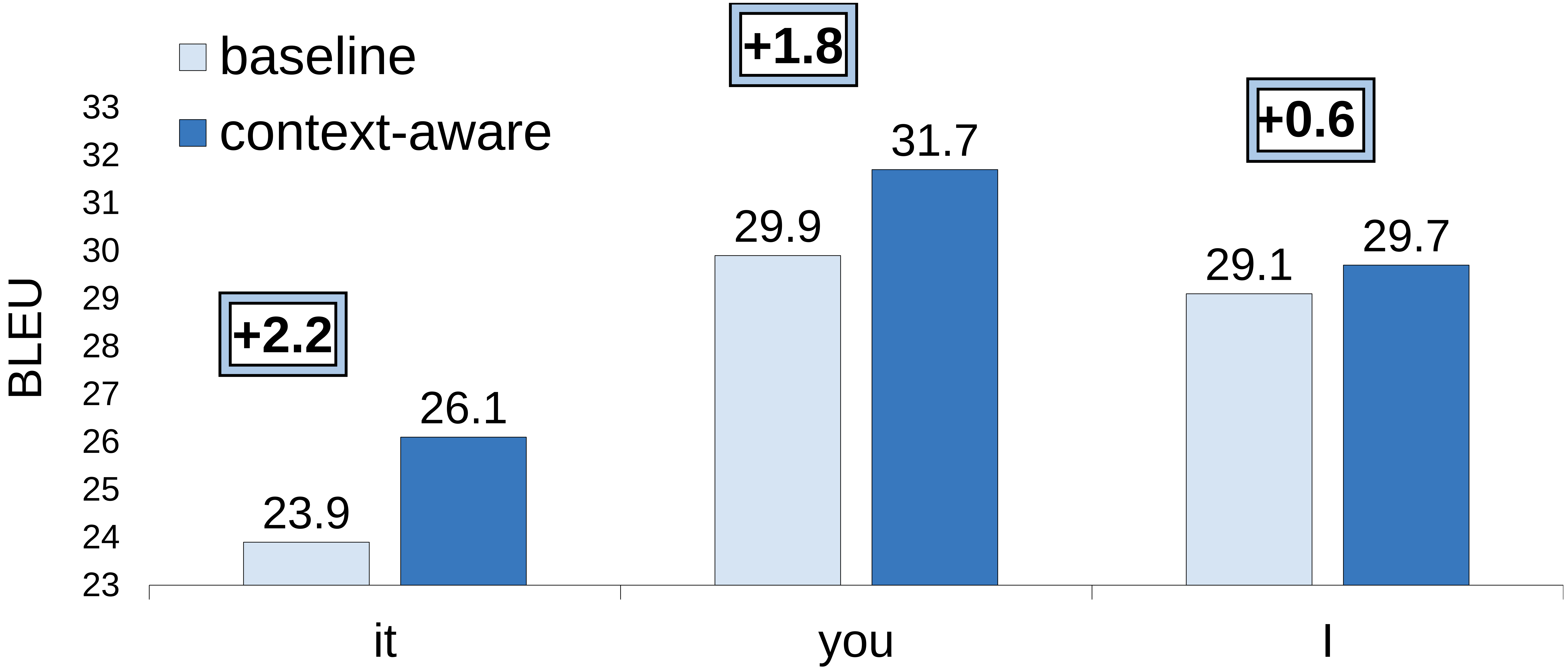
Ambiguous pronouns and translation quality: how to evaluate

Metric: BLEU (standard metric for MT)

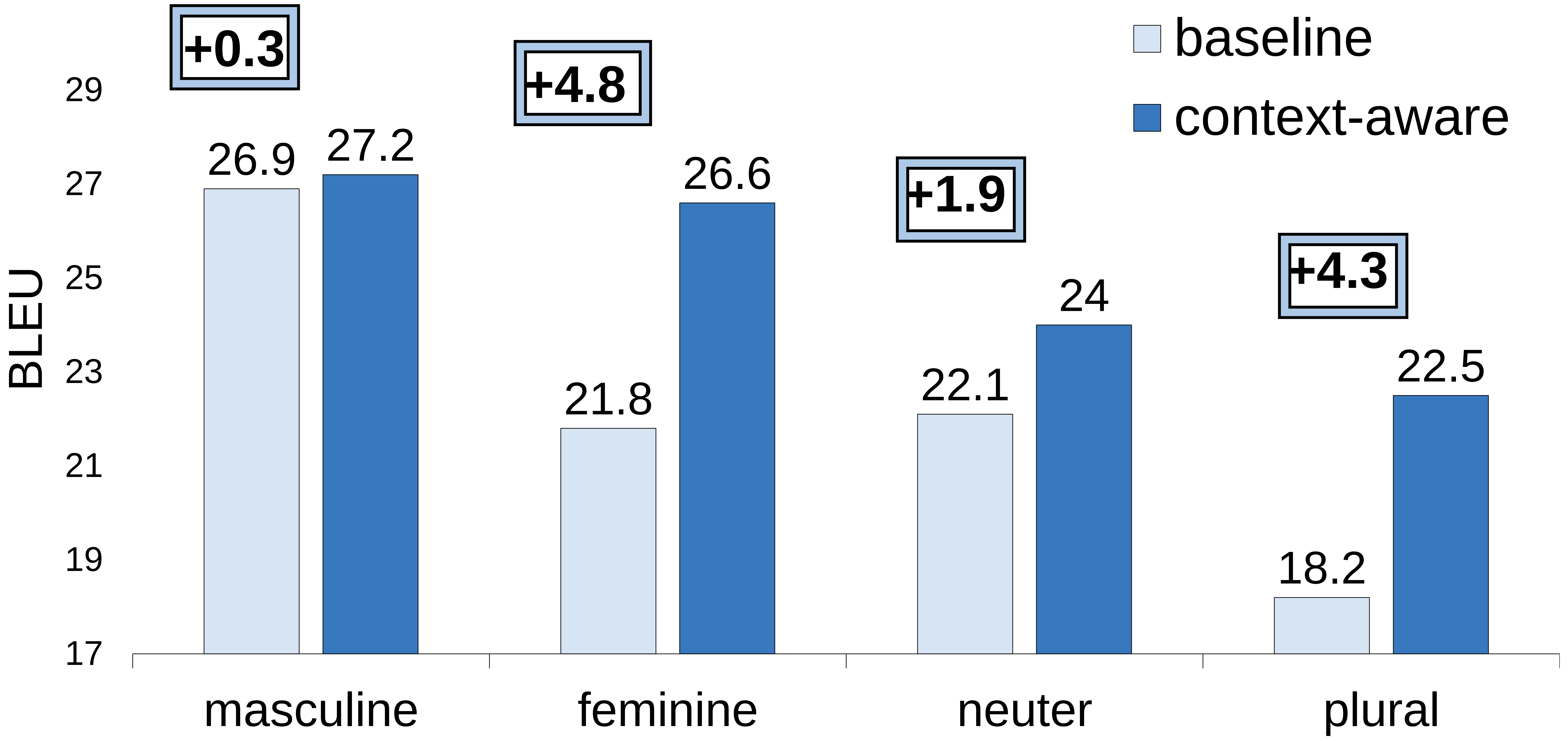
Specific test sets:

- feed CoreNLP (Manning et al., 2014) with pairs of sentences
- pick examples with a link between the pronoun and a noun group in a context
- gather a test set for each pronoun
- use the test sets to evaluate the context-aware NMT system

Ambiguous pronouns and translation quality: noun antecedent



Ambiguous “it”: noun antecedent



“It” with noun antecedent: example

Source:

- **It** was locked up in the hold with 20 other boxes of supplies.

Possible translations into Russian:

- **Он** был заперт в трюме с 20 другими ящиками с припасами. (masculine)
- **Оно** был**о** заперт**о** в трюме с 20 другими ящиками с припасами. (neuter)
- **Она** был**а** заперт**а** в трюме с 20 другими ящиками с припасами. (feminine)
- **Они** был**и** заперт**ы** в трюме с 20 другими ящиками с припасами. (plural)

“It” with noun antecedent: example

Context:

- You left **money** unattended?

Source:

- **It** was locked up in the hold with 20 other boxes of supplies.

Possible translations into Russian:

- **Они** был**и** заперт**ы** в трюме с 20 другими ящиками с припасами. (plural)

Latent anaphora resolution

Hypothesis

Observation:

- Large improvements in BLEU on test sets with pronouns co-referent with an expression in context

?

Attention mechanism = Latent anaphora resolution

How to test the hypothesis: agreement with CoreNLP

Test set:

- Find an antecedent noun phrase (using CoreNLP)
- Pick examples where the noun phrase contains a single noun
- Pick examples with several nouns in context

How to test the hypothesis: agreement with CoreNLP

Test set:

- Find an antecedent noun phrase (using CoreNLP)
- Pick examples where the noun phrase contains a single noun
- Pick examples with several nouns in context

Calculate an agreement:

- Identify the token with the largest attention weight (excluding punctuation, <bos> and <eos>)
- If the token falls within the antecedent span, then it's an agreement

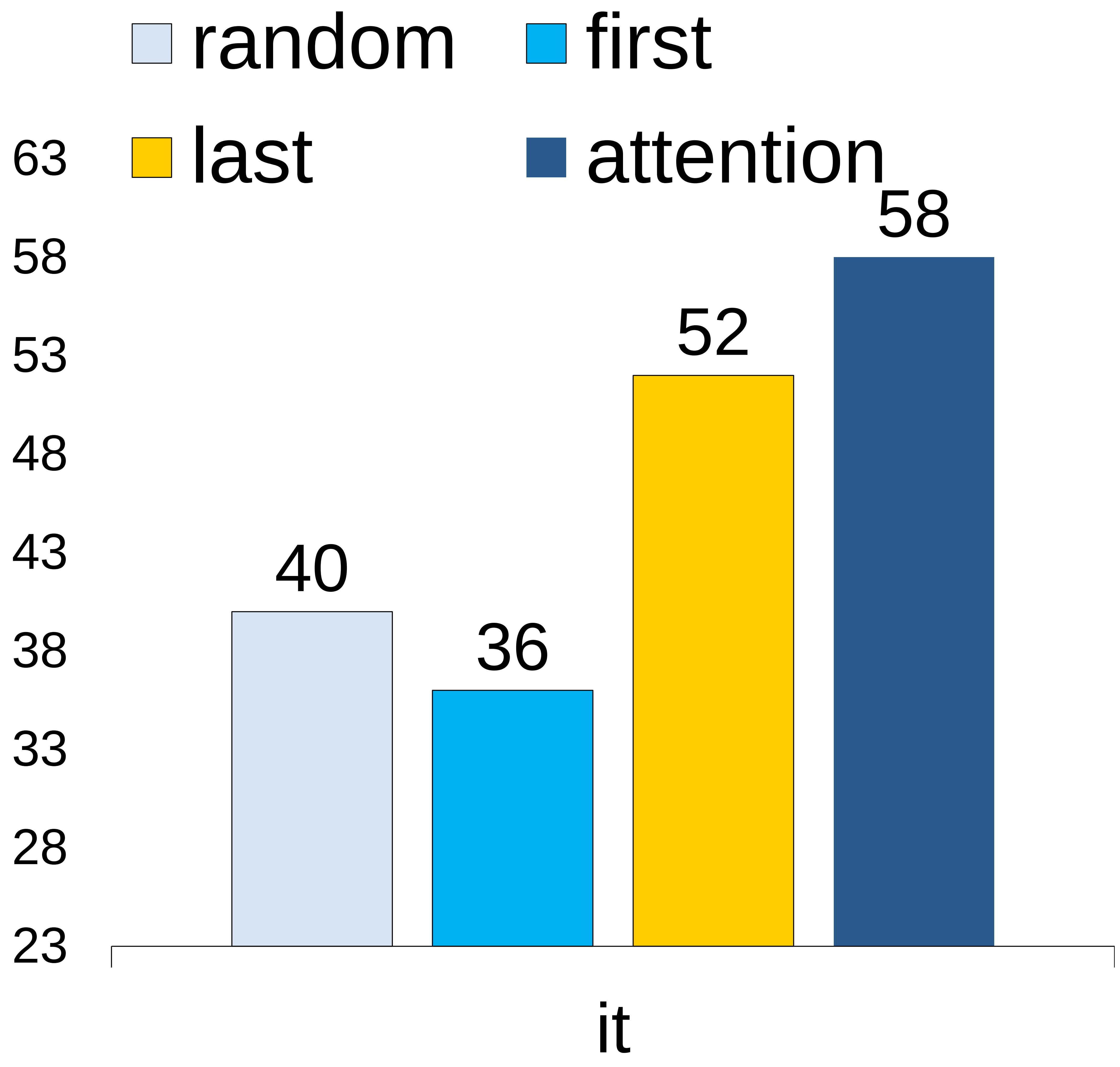
Does the model learn anaphora,

or just some simple heuristic?

Use several baselines:

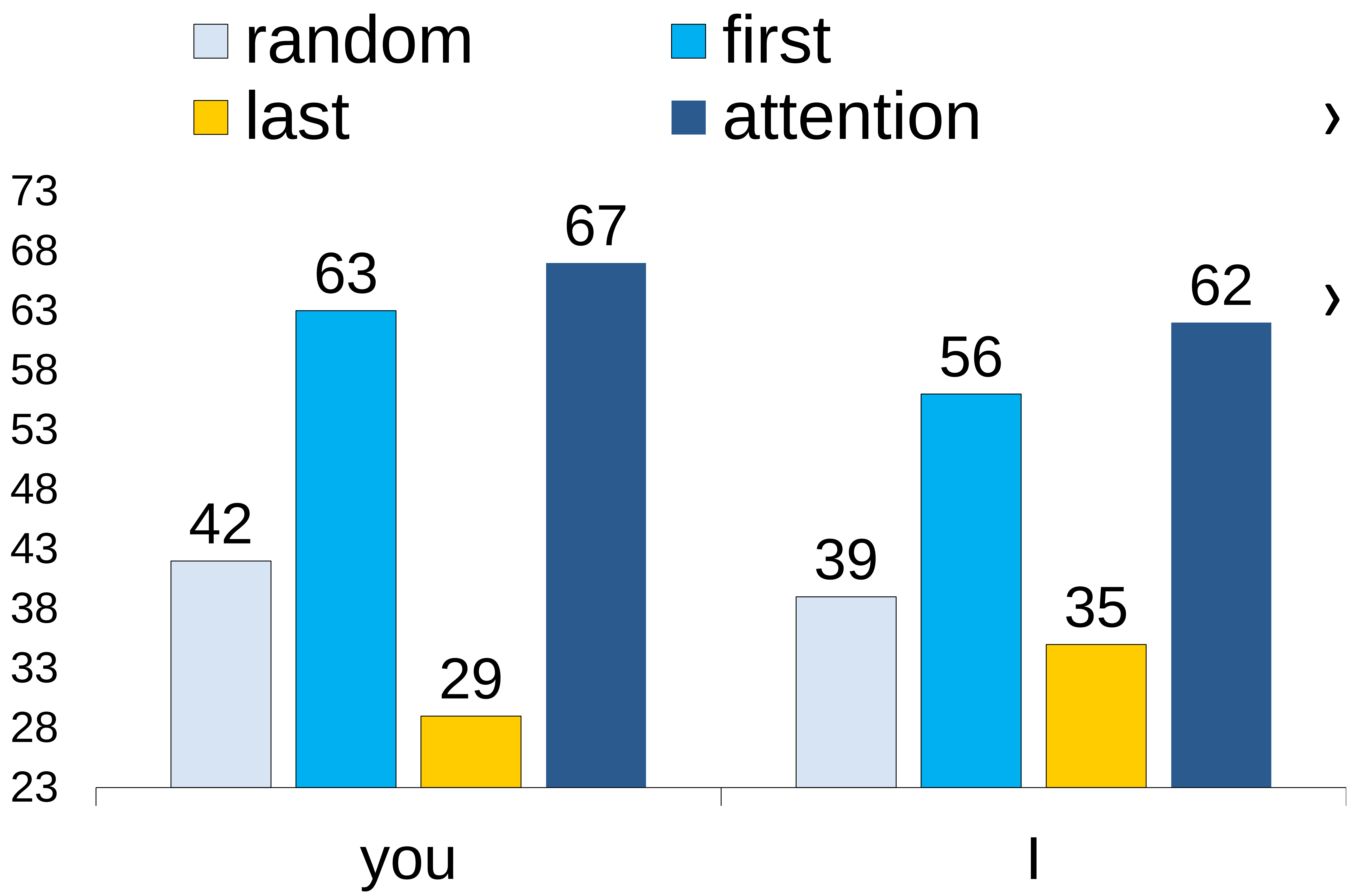
- random noun
- first noun
- last noun

Agreement with CoreNLP predictions



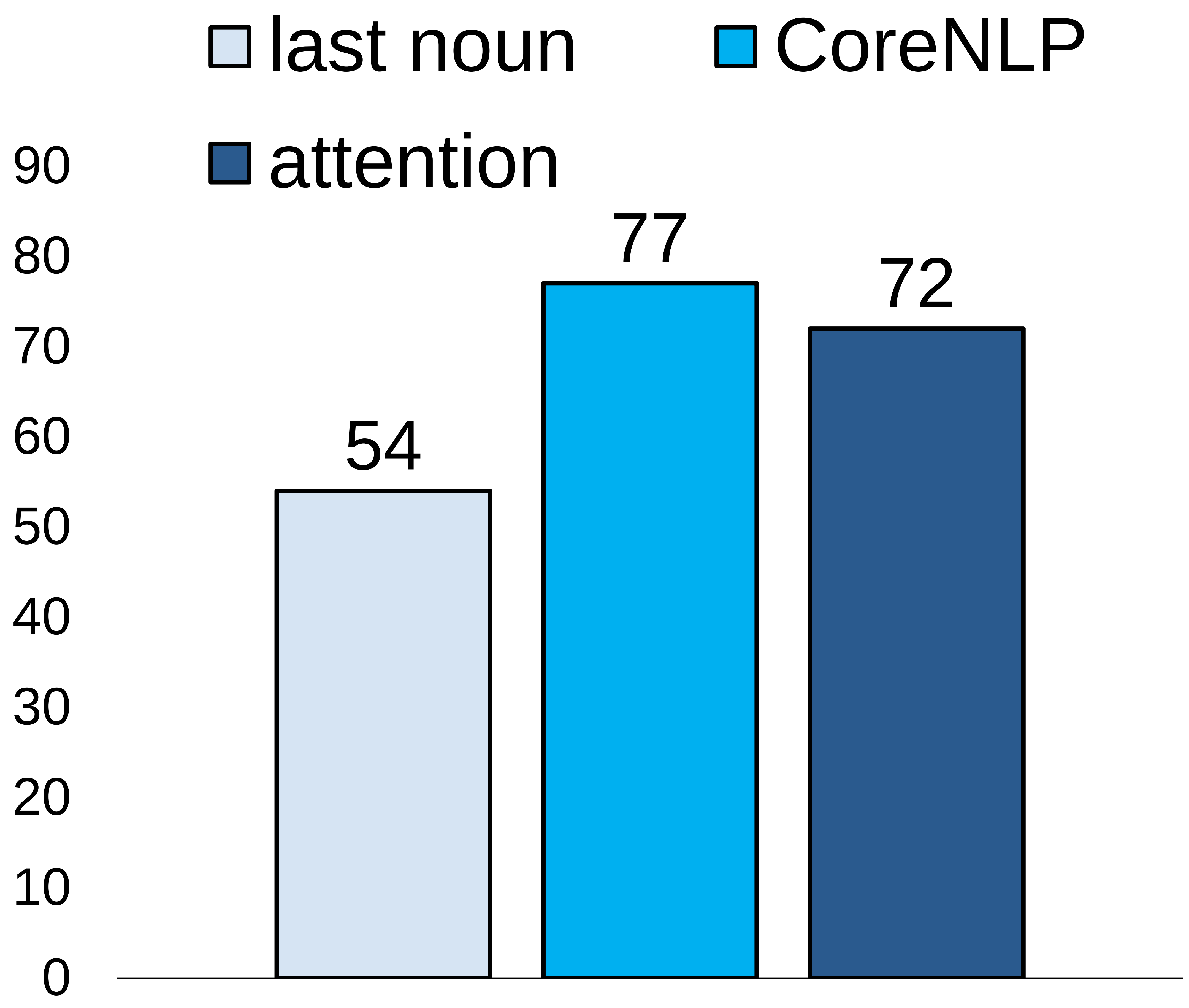
- agreement of attention is the highest
- last noun is the best heuristic

Agreement with CoreNLP predictions



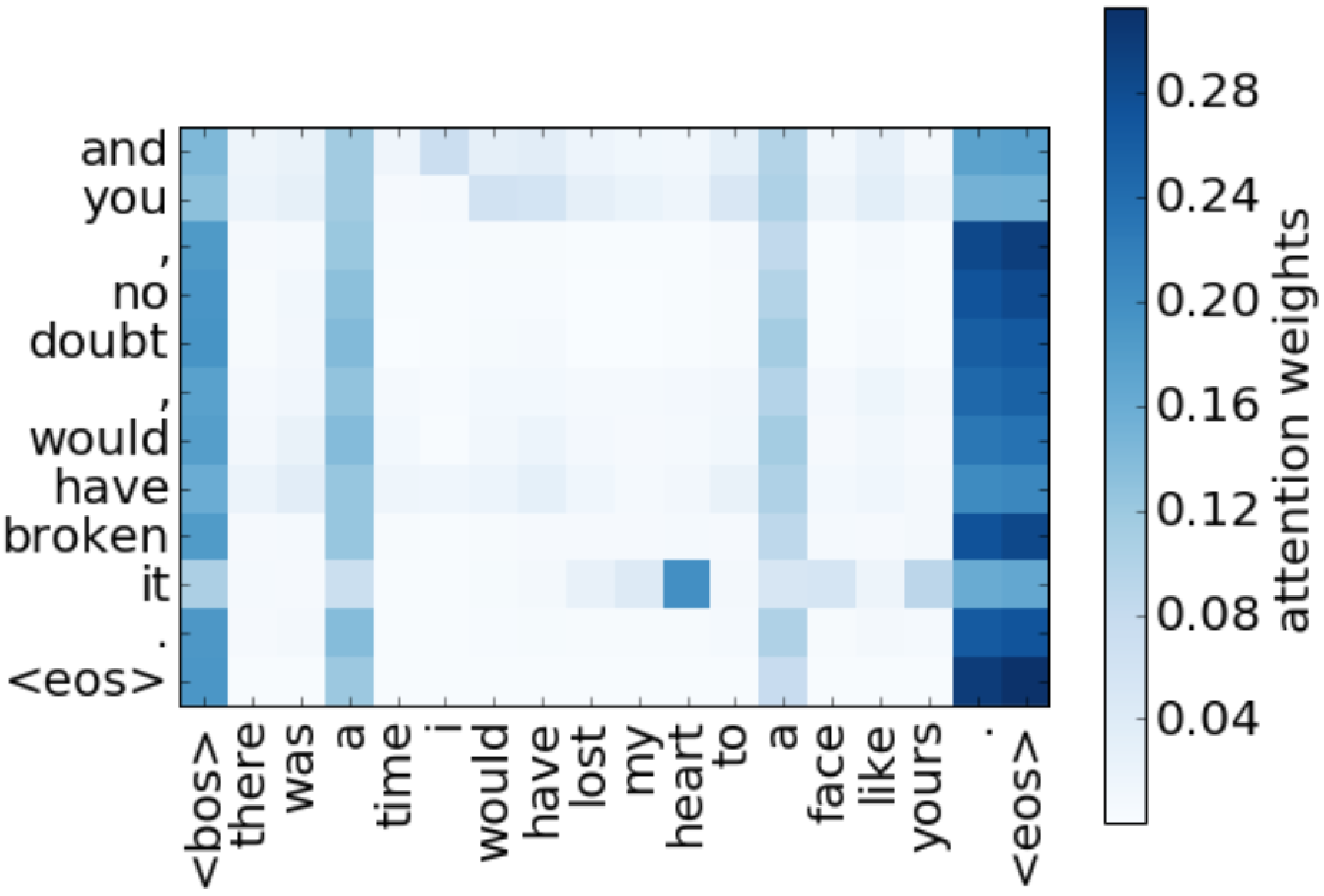
- agreement of attention is the highest
- first noun is the best heuristic

Compared to human annotations for “it”



- pick 500 examples from the previous experiment
- ask human annotators to mark an antecedent
- pick examples where an antecedent is a noun phrase
- calculate the agreement with human antecedents

Attention map examples



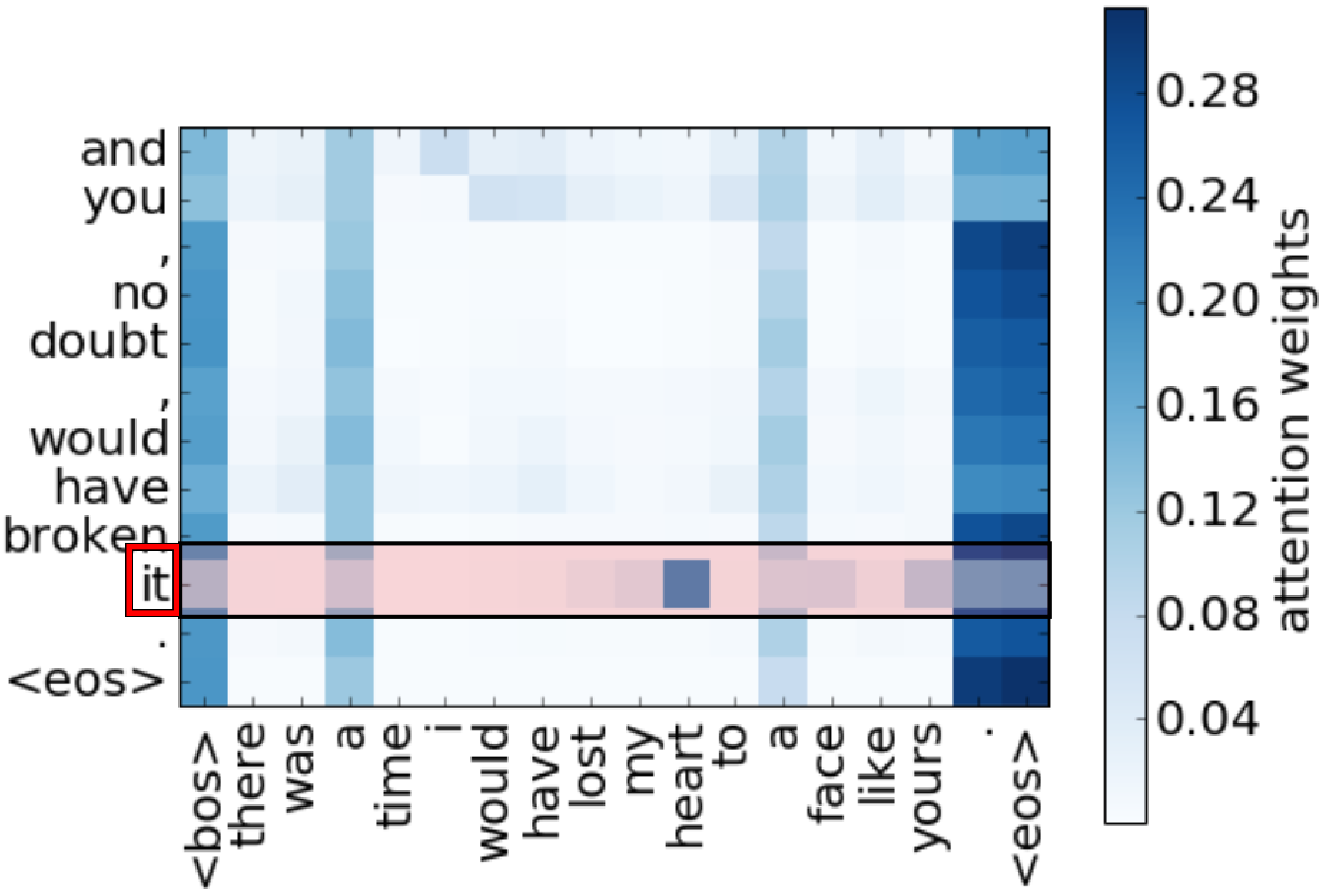
Context:

- There was a time I would have lost my heart to a face like yours.

Source:

- And you, no doubt, would have broken **it**.

Attention map examples



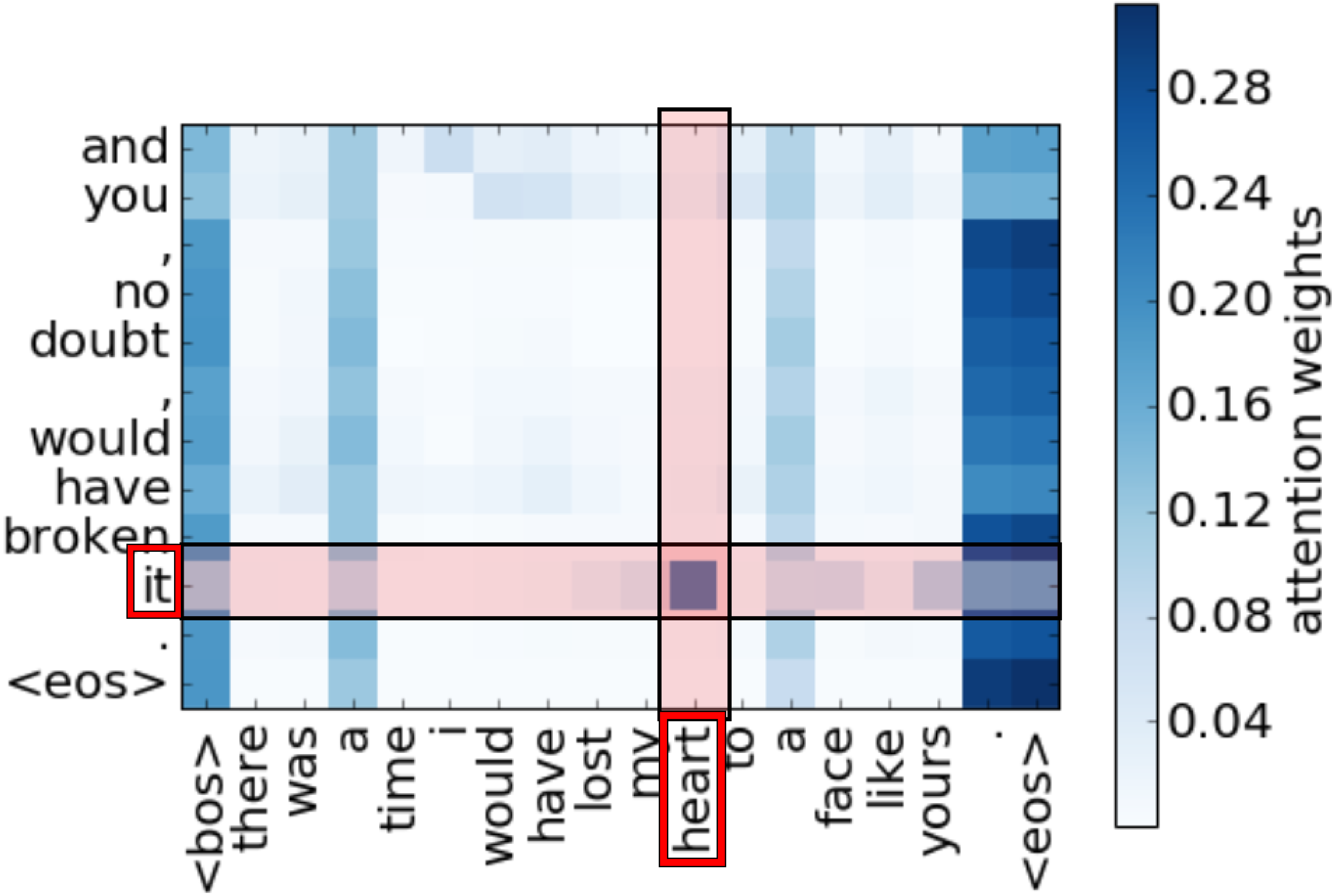
Context:

- There was a time I would have lost my heart to a face like yours.

Source:

- And you, no doubt, would have broken **it**.

Attention map examples



Context:

- There was a time I would have lost my **heart** to a face like yours.

Source:

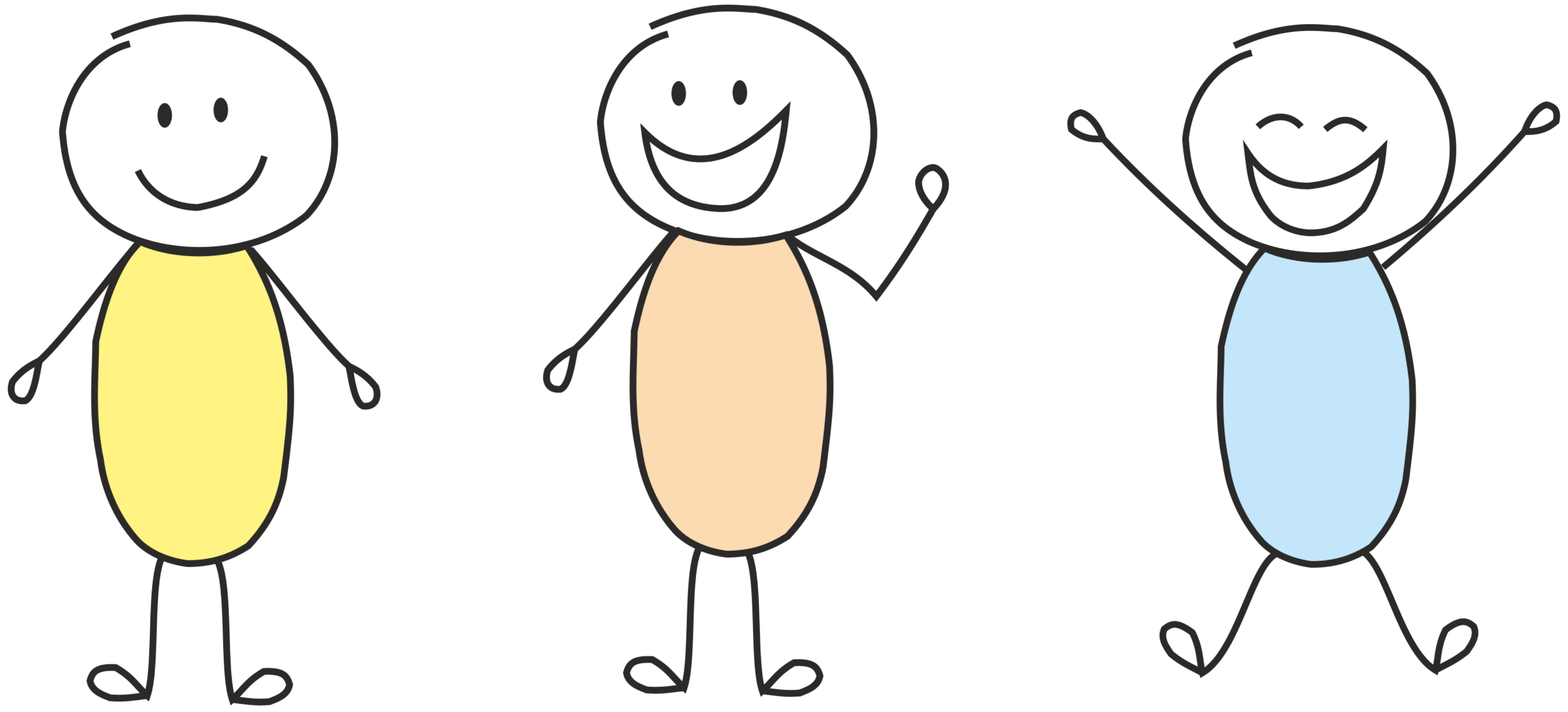
- And you, no doubt, would have broken **it**.

Conclusions

- introduce a context-aware NMT system based on the Transformer
- the model outperforms both the context-agnostic baseline and a simple context-aware baseline (on an En-Ru corpus)
- pronoun translation is the key phenomenon captured by the model
- the model induces anaphora relations

Thank you!

Questions?



References

- Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, USA.
- Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does Neural Machine Translation Benefit from Larger Context? In *arXiv:1704.05135*. ArXiv: 1704.05135.
- Pierre Lison, Jo ̈rg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan.

References

- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014b. The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Baltimore, Maryland, pages 55–60. https://doi.org/10.3115/v1/P14-5010.

- Jo ̈rg Tiedemann and Yves Scherrer. 2017. Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Copenhagen, Denmark, DISCOMT’17, pages 82–92. https://doi.org/10.18653/v1/W17- 4811.

- Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting Cross-Sentence Con- text for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Den- mark, Copenhagen, EMNLP’17, pages 2816–2821. https://doi.org/10.18653/v1/D17-1301.

References

- Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with coreference resolution. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. Association for Computational Linguistics, Uppsala, Sweden, pages 252–261. http://www.aclweb.org/anthology/W10-1737.
- Christian Hardmeier and Marcello Federico. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT). pages 283–289.
- Christian Hardmeier, Preslav Nakov, Sara Stymne, Jo¨rg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation. In Proceedings of the Second Workshop on Discourse in Machine Translation. Association for Computational Linguistics, Lisbon, Portugal, pages 1–16. https://doi.org/10.18653/v1/W15-2501.

References

- Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. 2012. Machine Translation of Labeled Discourse Connectives. In Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA). http://www.mt-archive.info/AMTA-2012- Meyer.pdf.
- Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based Document-level Statistical Machine Translation. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Edinburgh, Scotland, UK., pages 909–919. http://www.aclweb.org/anthology/D11-1084.
- Marine Carpuat. 2009. One Translation Per Discourse. In Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. Association for Computational Linguistics, Boulder, Colorado, pages 19–27. http://www.aclweb.org/anthology/W09-2404.

References

- Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based Document-level Statistical Machine Translation. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Edinburgh, Scotland, UK., pages 909–919. http://www.aclweb.org/anthology/D11-1084.
- Jo ̈rg Tiedemann. 2010. Context Adaptation in Statistical Machine Translation Using Models with Exponentially Decaying Cache. In Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing. Association for Computational Linguistics, Uppsala, Sweden, pages 8–15. http://www.aclweb.org/anthology/W10-2602.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In NIPS. Los Angeles. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.