Split and Rephrase: Better Evaluation and a Stronger Baseline

Roee Aharoni and Yoav Goldberg
NLP Lab, Bar Ilan University, Israel
ACL 2018

Motivation

Motivation

- Processing long, complex sentences is hard!

B I U
N L P

Motivation

- Processing long, complex sentences is hard!
- Children, people with reading disabilities, L2 learners…

Simple English
WIKIPEDIA

When writing articles here:

- **Use Basic English vocabulary and shorter sentences.** This allows people to understand normally complex terms or phrases.

Motivation

- Processing long, complex sentences is hard!
- Children, people with reading disabilities, L2 learners…
- Sentence level NLP systems:

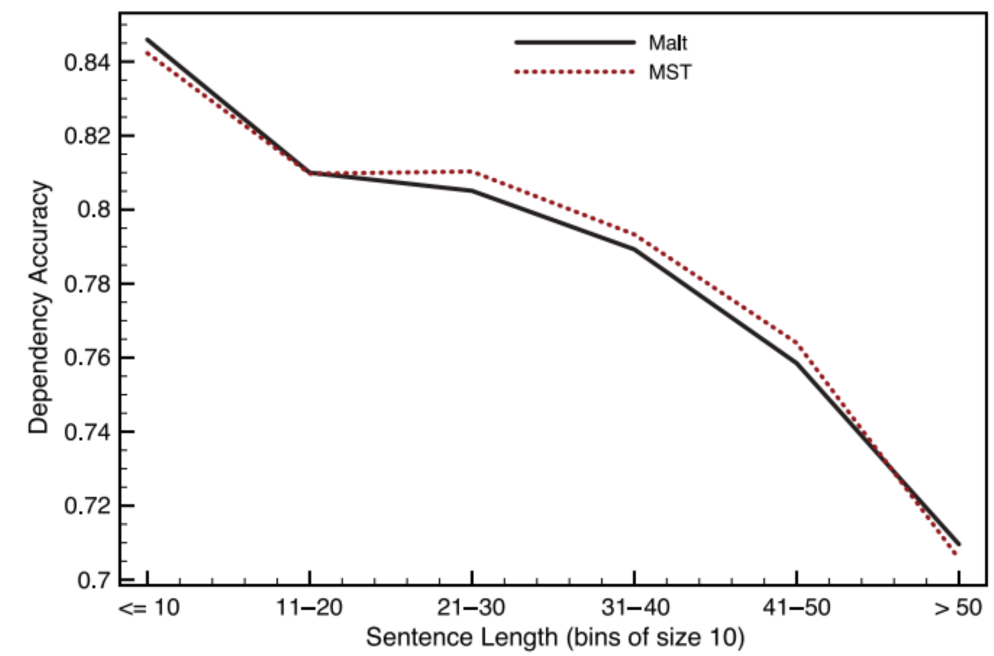
Simple English
WIKIPEDIA

When writing articles here:

- **Use** Basic English **vocabulary and shorter sentences.** This allows people to understand normally complex terms or phrases.

Motivation

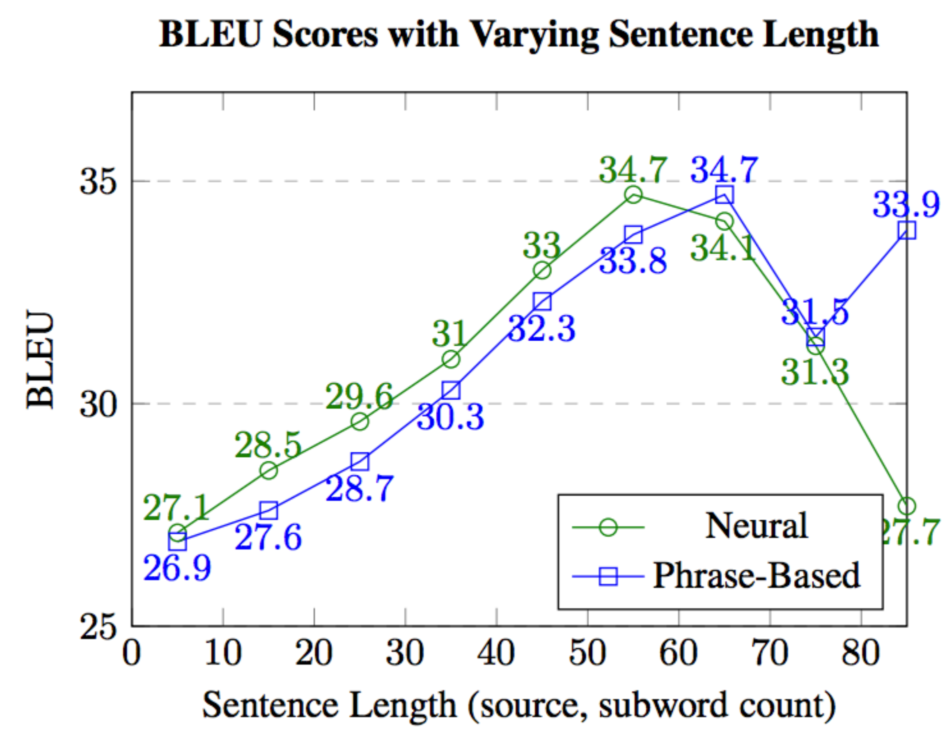
- Processing long, complex sentences is hard!
- Children, people with reading disabilities, L2 learners…
- Sentence level NLP systems:
  - Dependency Parsers



McDonald & Nivre, 2011

Motivation

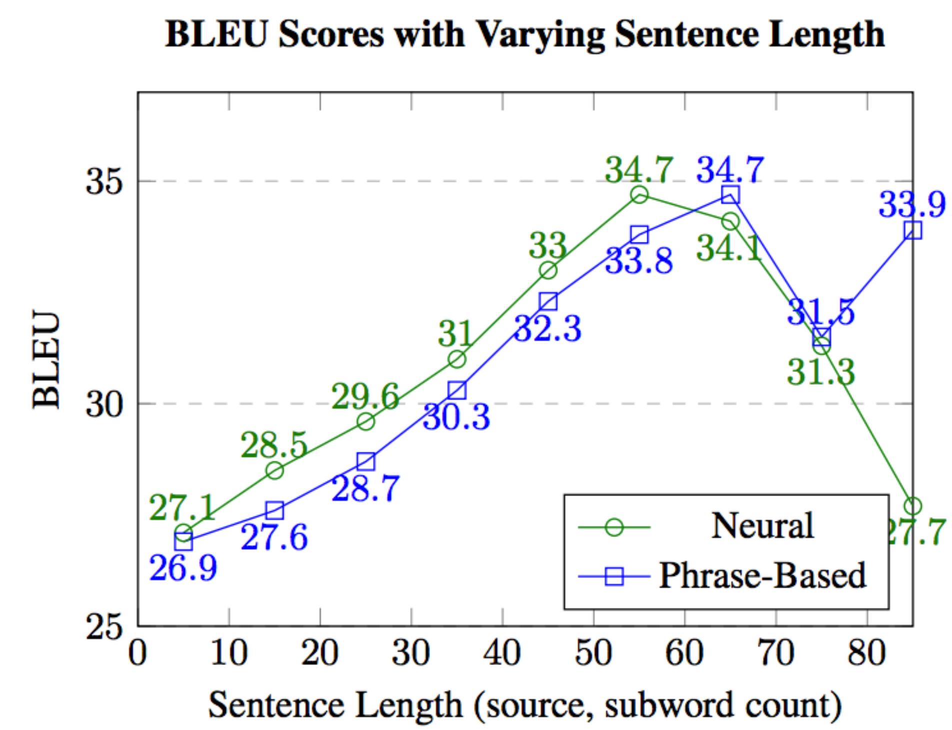
- Processing long, complex sentences is hard!
- Children, people with reading disabilities, L2 learners…
- Sentence level NLP systems:
  - Dependency Parsers
  - Neural Machine Translation



Koehn & Knowles, 2017

Motivation

- Processing long, complex sentences is hard!
- Children, people with reading disabilities, L2 learners…
- Sentence level NLP systems:
  - Dependency Parsers
  - Neural Machine Translation
- **Can we automatically break a complex sentence into several simple ones while preserving its meaning?**



Koehn & Knowles, 2017

The Split and Rephrase Task

The Split and Rephrase Task

- Narayan, Gardent, Cohen & Shimorina, EMNLP 2017

The Split and Rephrase Task

- Narayan, Gardent, Cohen & Shimorina, EMNLP 2017
- Dataset, evaluation method, baseline models

The Split and Rephrase Task

- Narayan, Gardent, Cohen & Shimorina, EMNLP 2017
- Dataset, evaluation method, baseline models
- Task definition: complex sentence -> several simple sentences with the same meaning

The Split and Rephrase Task

- Narayan, Gardent, Cohen & Shimorina, EMNLP 2017
- Dataset, evaluation method, baseline models
- Task definition: complex sentence -> several simple sentences with the same meaning

Alan Bean joined NASA in 1963 where he became a member of the Apollo 12 mission along with Alfred Worden as back up pilot and David Scott as commander .

The Split and Rephrase Task

- Narayan, Gardent, Cohen & Shimorina, EMNLP 2017
- Dataset, evaluation method, baseline models
- Task definition: complex sentence -> several simple sentences with the same meaning

Alan Bean joined NASA in 1963 where he became a member of the Apollo 12 mission along with Alfred Worden as back up pilot and David Scott as commander .

The Split and Rephrase Task

- Narayan, Gardent, Cohen & Shimorina, EMNLP 2017
- Dataset, evaluation method, baseline models
- Task definition: complex sentence -> several simple sentences with the same meaning

Alan Bean joined NASA in 1963 where he became a member of the Apollo 12 mission along with Alfred Worden as back up pilot and David Scott as commander .

Alan Bean served as a crew member of Apollo 12 .
Alfred Worden was the backup pilot of Apollo 12 .
Apollo 12 was commanded by David Scott .
Alan Bean was selected by Nasa in 1963 .

The Split and Rephrase Task

- Narayan, Gardent, Cohen & Shimorina, EMNLP 2017
- Dataset, evaluation method, baseline models
- Task definition: complex sentence -> several simple sentences with the same meaning
- Requires (a) **identifying** independent semantic units (b) **rephrasing** those units to single sentences

Alan Bean joined NASA in 1963 where he became a member of the Apollo 12 mission along with Alfred Worden as back up pilot and David Scott as commander .

↓

Alan Bean served as a crew member of Apollo 12 .
Alfred Worden was the backup pilot of Apollo 12 .
Apollo 12 was commanded by David Scott .
Alan Bean was selected by Nasa in 1963 .

This Work

This Work

- We show that simple neural models seem to perform very on the original benchmark due to **memorization** of the training set

This Work

- We show that simple neural models seem to perform very on the original benchmark due to **memorization** of the training set
- We propose a **more challenging data split** for the task to discourage memorization

This Work

- We show that simple neural models seem to perform very on the original benchmark due to **memorization** of the training set
- We propose a **more challenging data split** for the task to discourage memorization
- We perform automatic evaluation and error analysis on the new benchmark, showing that the task is **still far from being solved**

WebSplit Dataset Construction
(Narayan et al. 2017)

WebSplit Dataset Construction (Narayan et al. 2017)

Simple RDF Triples
(facts from DBpedia)

<Alan_Bean | nationality | United_States>

<Alan_Bean | mission | Apollo_12>

<Alan_Bean | NASA selection | 1963>

WebSplit Dataset Construction (Narayan et al. 2017)

Simple RDF Triples (facts from DBpedia)

Simple Sentences

<Alan_Bean | nationality | United_States> → Alan Bean is a US national.

<Alan_Bean | mission | Apollo_12> → Alan Bean was on the crew of Apollo 12.

<Alan_Bean | NASA selection | 1963> → Alan Bean was hired by NASA in 1963.

WebSplit Dataset Construction (Narayan et al. 2017)

Simple RDF Triples (facts from DBpedia)

Simple Sentences

<Alan_Bean | nationality | United_States> → Alan Bean is a US national.

<Alan_Bean | mission | Apollo_12> → Alan Bean was on the crew of Apollo 12.

<Alan_Bean | NASA selection | 1963> → Alan Bean was hired by NASA in 1963.

Sets of RDF triples

<Alan_Bean | nationality | United_States,
Alan_Bean | mission | Apollo_12,
Alan_Bean | NASA selection | 1963>

WebSplit Dataset Construction (Narayan et al. 2017)

Simple RDF Triples (facts from DBpedia)

Simple Sentences

<Alan_Bean | nationality | United_States> → Alan Bean is a US national.

<Alan_Bean | mission | Apollo_12> → Alan Bean was on the crew of Apollo 12.

<Alan_Bean | NASA selection | 1963> → Alan Bean was hired by NASA in 1963.

Sets of RDF triples

Complex Sentences

<Alan_Bean | nationality | United_States,
Alan_Bean | mission | Apollo_12,
Alan_Bean | NASA selection | 1963>

→

Alan Bean, born in the United States, was selected by NASA in 1963 and served as a crew member of Apollo 12.

WebSplit Dataset Construction (Narayan et al. 2017)

Simple RDF Triples (facts from DBpedia)

<Alan_Bean | nationality | United_States>

<Alan_Bean | mission | Apollo_12>

<Alan_Bean | NASA selection | 1963>

Simple Sentences

Alan Bean is a US national.

Alan Bean was on the crew of Apollo 12.

Alan Bean was hired by NASA in 1963.

Matching via RDFs

Sets of RDF triples

<Alan_Bean | nationality | United_States,
Alan_Bean | mission | Apollo_12,
Alan_Bean | NASA selection | 1963>

Complex Sentences

Alan Bean, born in the United States, was selected by NASA in 1963 and served as a crew member of Apollo 12.

WebSplit Dataset Construction (Narayan et al. 2017)

Simple RDF Triples (facts from DBpedia)

<Alan_Bean | nationality | United_States>

<Alan_Bean | mission | Apollo_12>

<Alan_Bean | NASA selection | 1963>

Simple Sentences

Alan Bean is a US national.

Alan Bean was on the crew of Apollo 12.

Alan Bean was hired by NASA in 1963.

Matching via RDFs

~1M examples

Sets of RDF triples

<Alan_Bean | nationality | United_States,
Alan_Bean | mission | Apollo_12,
Alan_Bean | NASA selection | 1963>

Complex Sentences

Alan Bean, born in the United States, was selected by NASA in 1963 and served as a crew member of Apollo 12.

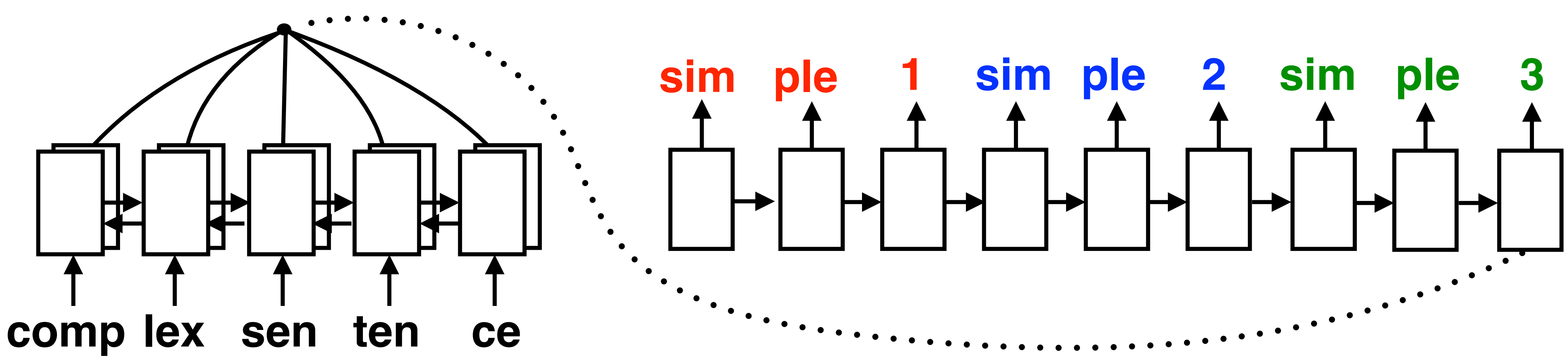
Preliminary Experiments

Preliminary Experiments

- ~1M training examples

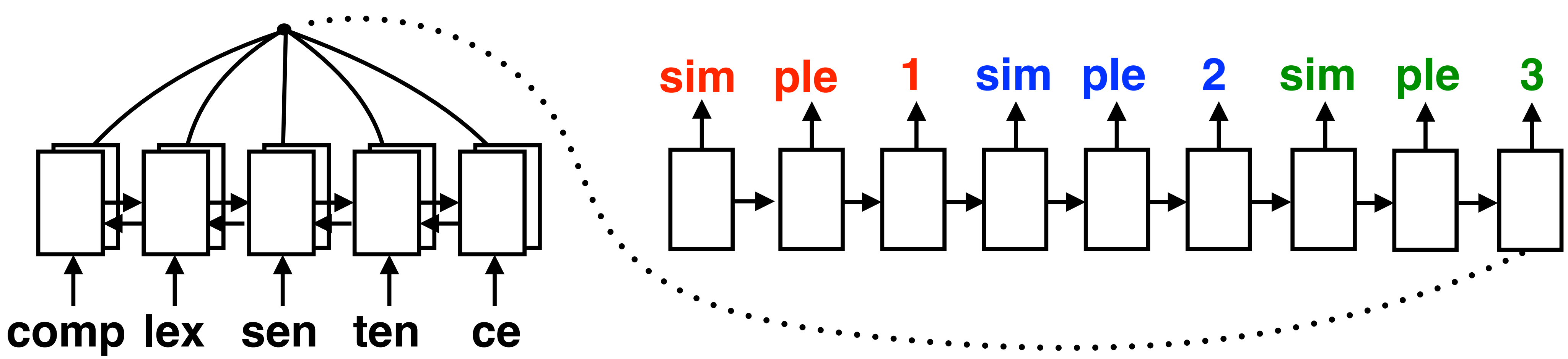
Preliminary Experiments

- ~1M training examples
- “Vanilla” LSTM seq2seq with attention



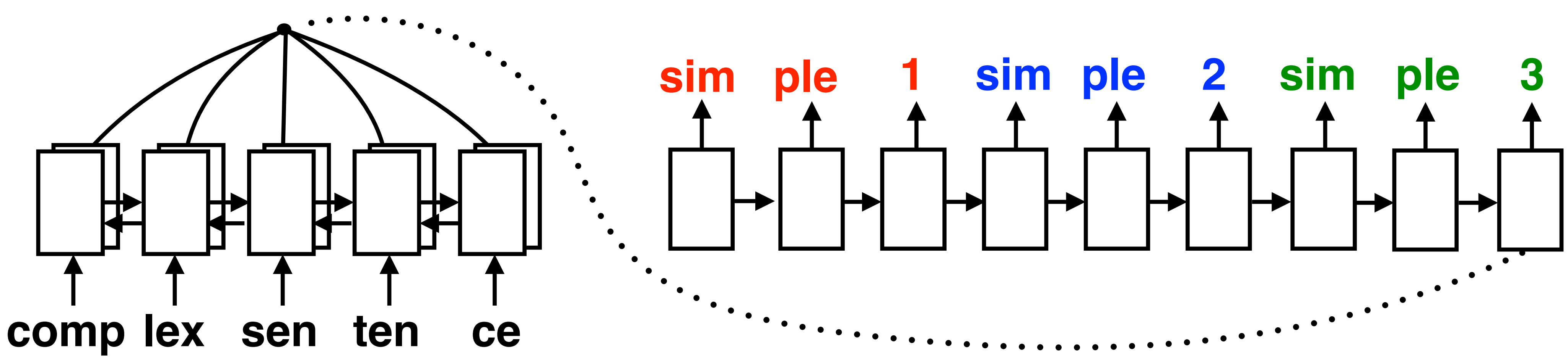
Preliminary Experiments

- ~1M training examples
- “Vanilla” LSTM seq2seq with attention
- Shared vocabulary between the encoder and the decoder



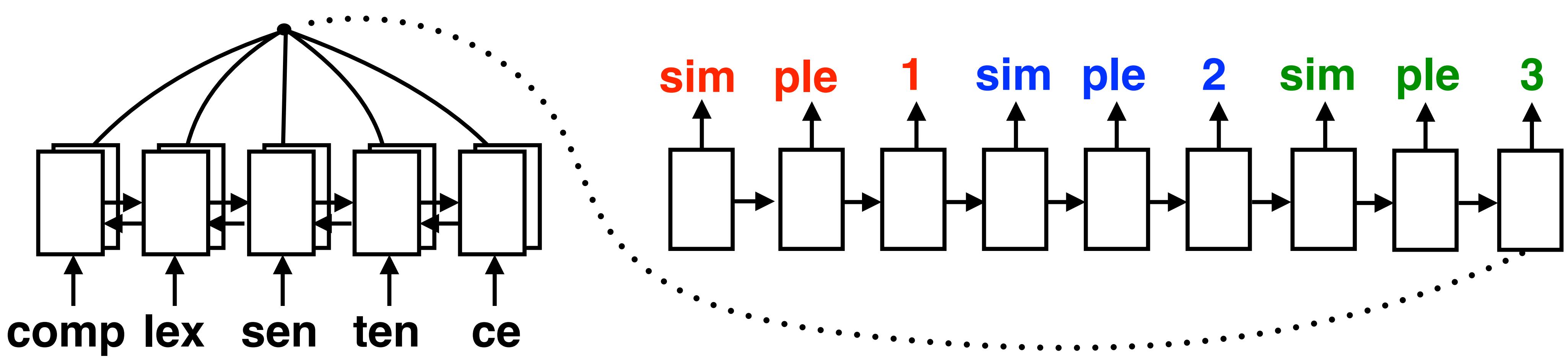
Preliminary Experiments

- ~1M training examples
- “Vanilla” LSTM seq2seq with attention
- Shared vocabulary between the encoder and the decoder
- Simple sentences predicted as a single sequence



Preliminary Experiments

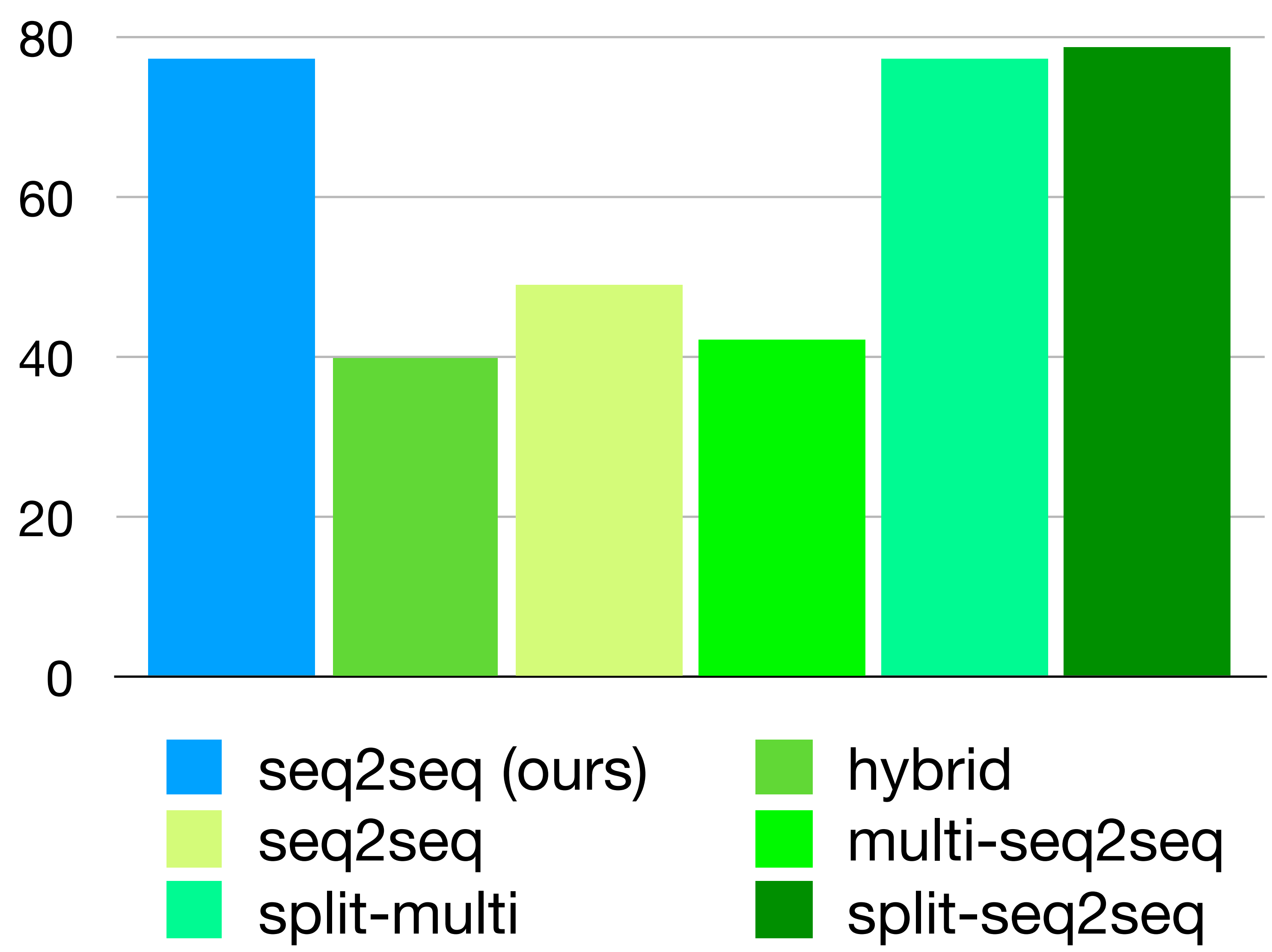
- ~1M training examples
- “Vanilla” LSTM seq2seq with attention
- Shared vocabulary between the encoder and the decoder
- Simple sentences predicted as a single sequence
- Evaluated using single-sentence, multi-reference BLEU as in Narayan et al. 2017



Preliminary Results

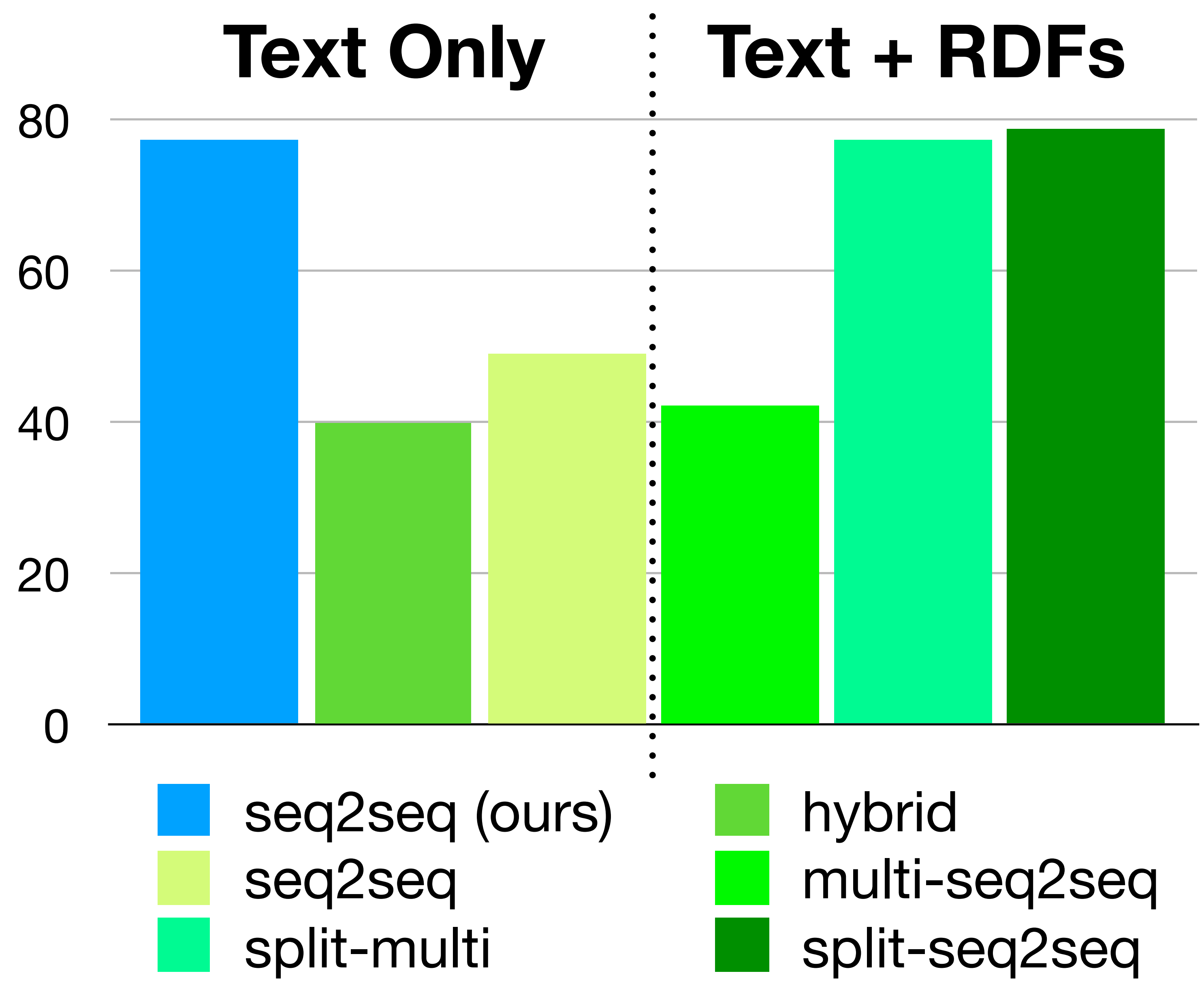
Preliminary Results

- Our simple seq2seq baseline outperform **all but one** of the baselines from Narayan et al. 2017



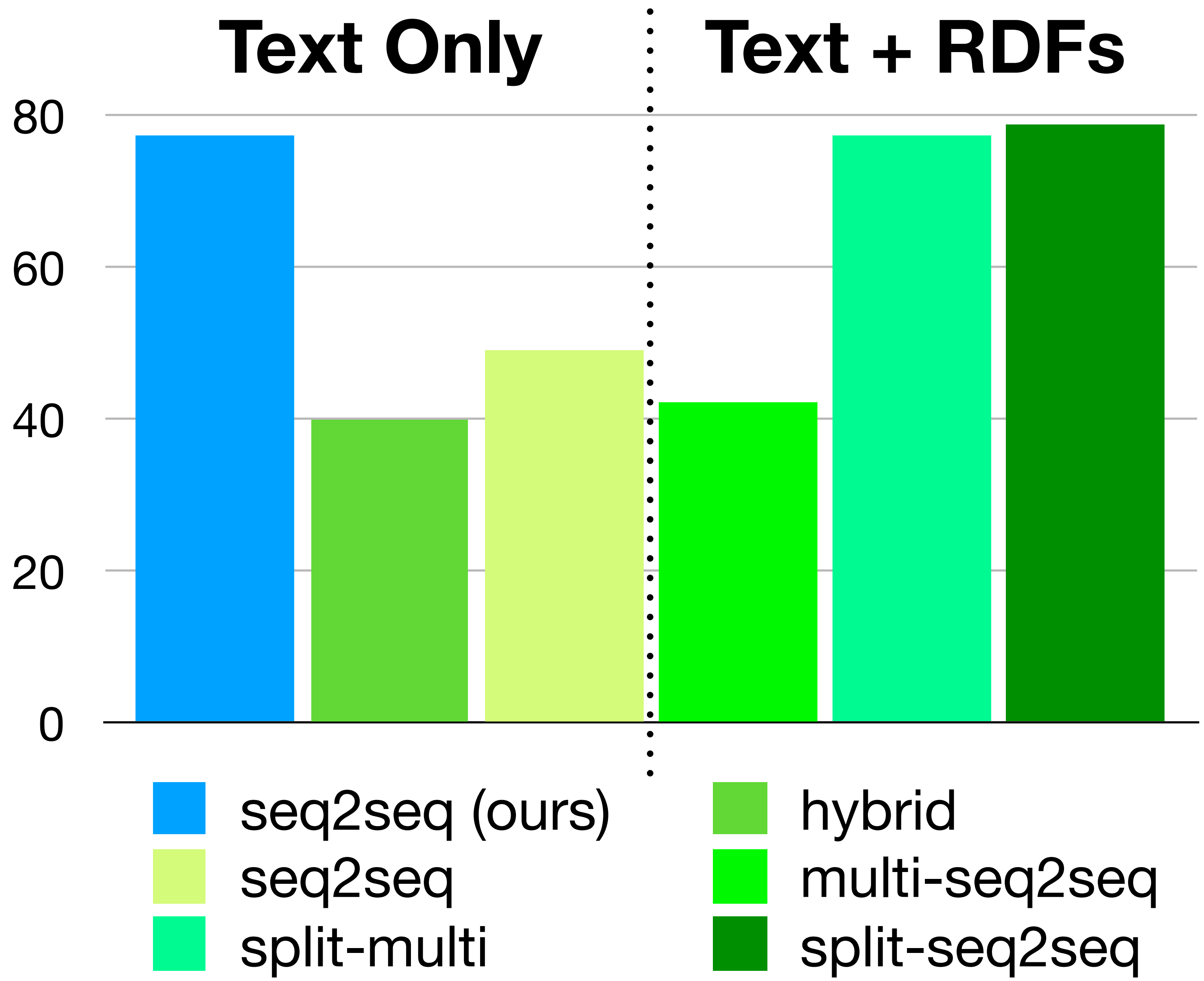
Preliminary Results

- Our simple seq2seq baseline outperform **all but one** of the baselines from Narayan et al. 2017
- Their best baselines were using the RDF structures as additional information



Preliminary Results

- Our simple seq2seq baseline outperform **all but one** of the baselines from Narayan et al. 2017
- Their best baselines were using the RDF structures as additional information
- **Do the simple seq2seq model really performs so well?**



BLEU can be Misleading

BLEU can be Misleading

- In spite of the high BLEU scores, our neural models suffer from:

BLEU can be Misleading

- In spite of the high BLEU scores, our neural models suffer from:
  - Missing facts - appeared in the input but not in the output

Input	Prediction
A Fortress of Grey Ice with ISBM 0-7653-0633-6 has 672 pages .	J.V. Jones authored A Fortress of Grey Ice . A Fortress of Grey Ice has 672 pages .
The address , 11 Diagonal Street is located in South Africa where the leader is Cyril Ramaphosa and some Asian South Africans live .	The address , 11 Diagonal Street is located in South Africa . The leader of South Africa is called Cyril Ramaphosa . The leader of South Africa is called Cyril Ramaphosa . The leader of South Africa is called Cyril Ramaphosa .

BLEU can be Misleading

- In spite of the high BLEU scores, our neural models suffer from:
  - Missing facts - appeared in the input but not in the output
  - Unsupported facts - appeared in the output but not in the input

Input	Prediction
A Fortress of Grey Ice with ISBM 0-7653-0633-6 has 672 pages .	J.V. Jones authored A Fortress of Grey Ice . A Fortress of Grey Ice has 672 pages .
The address , 11 Diagonal Street is located in South Africa where the leader is Cyril Ramaphosa and some Asian South Africans live .	The address , 11 Diagonal Street is located in South Africa . The leader of South Africa is called Cyril Ramaphosa . The leader of South Africa is called Cyril Ramaphosa . The leader of South Africa is called Cyril Ramaphosa .

BLEU can be Misleading

- In spite of the high BLEU scores, our neural models suffer from:
  - Missing facts - appeared in the input but not in the output
  - Unsupported facts - appeared in the output but not in the input
  - Repeated facts - appeared several times in the output

Input	Prediction
A Fortress of Grey Ice with ISBM 0-7653-0633-6 has 672 pages .	J.V. Jones authored A Fortress of Grey Ice . A Fortress of Grey Ice has 672 pages .
The address , 11 Diagonal Street is located in South Africa where the leader is Cyril Ramaphosa and some Asian South Africans live .	The address , 11 Diagonal Street is located in South Africa . The leader of South Africa is called Cyril Ramaphosa . The leader of South Africa is called Cyril Ramaphosa . The leader of South Africa is called Cyril Ramaphosa .

A Closer Look

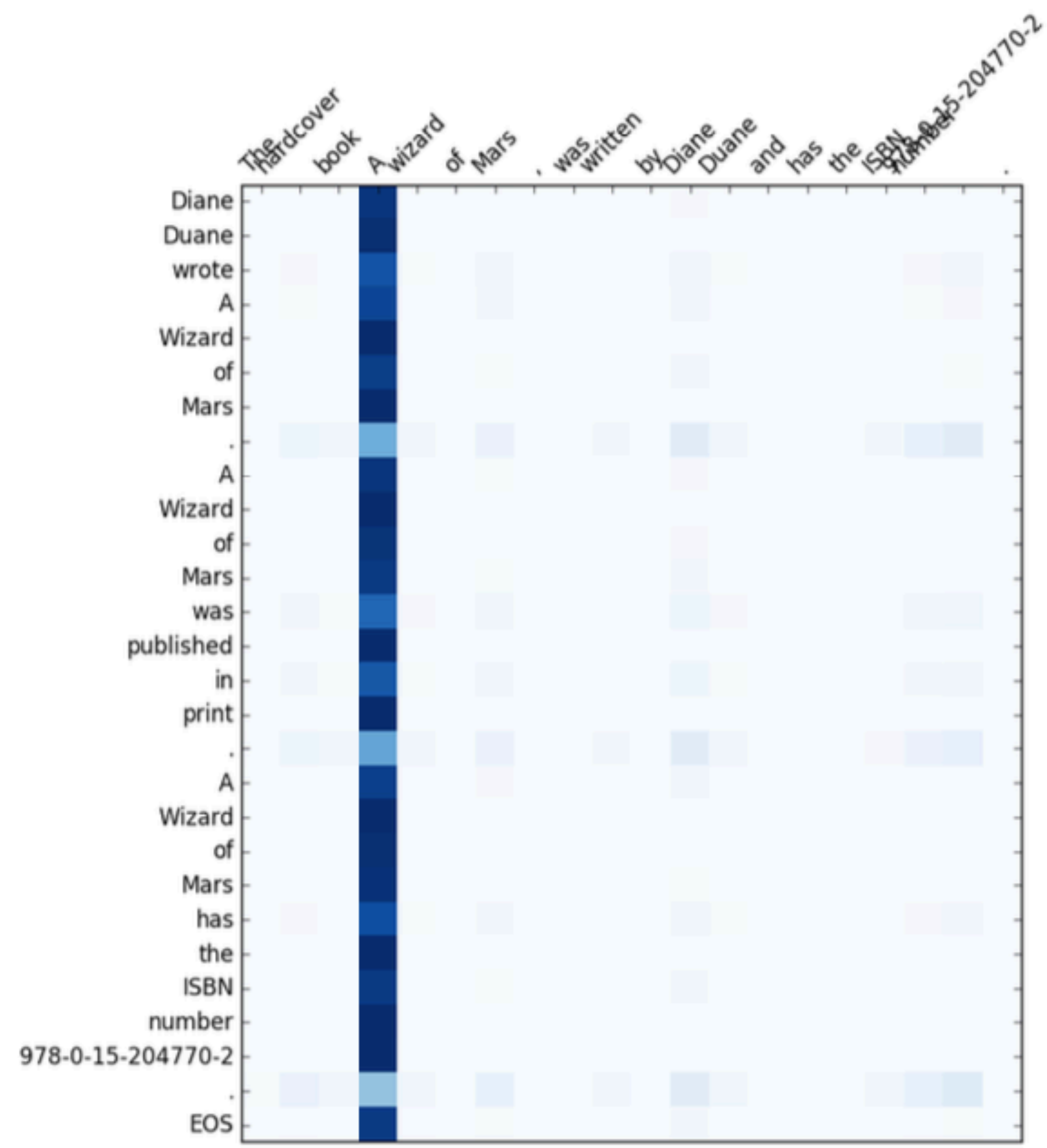
A Closer Look

- Visualizing the attention weights we find an unexpected pattern

A Closer Look

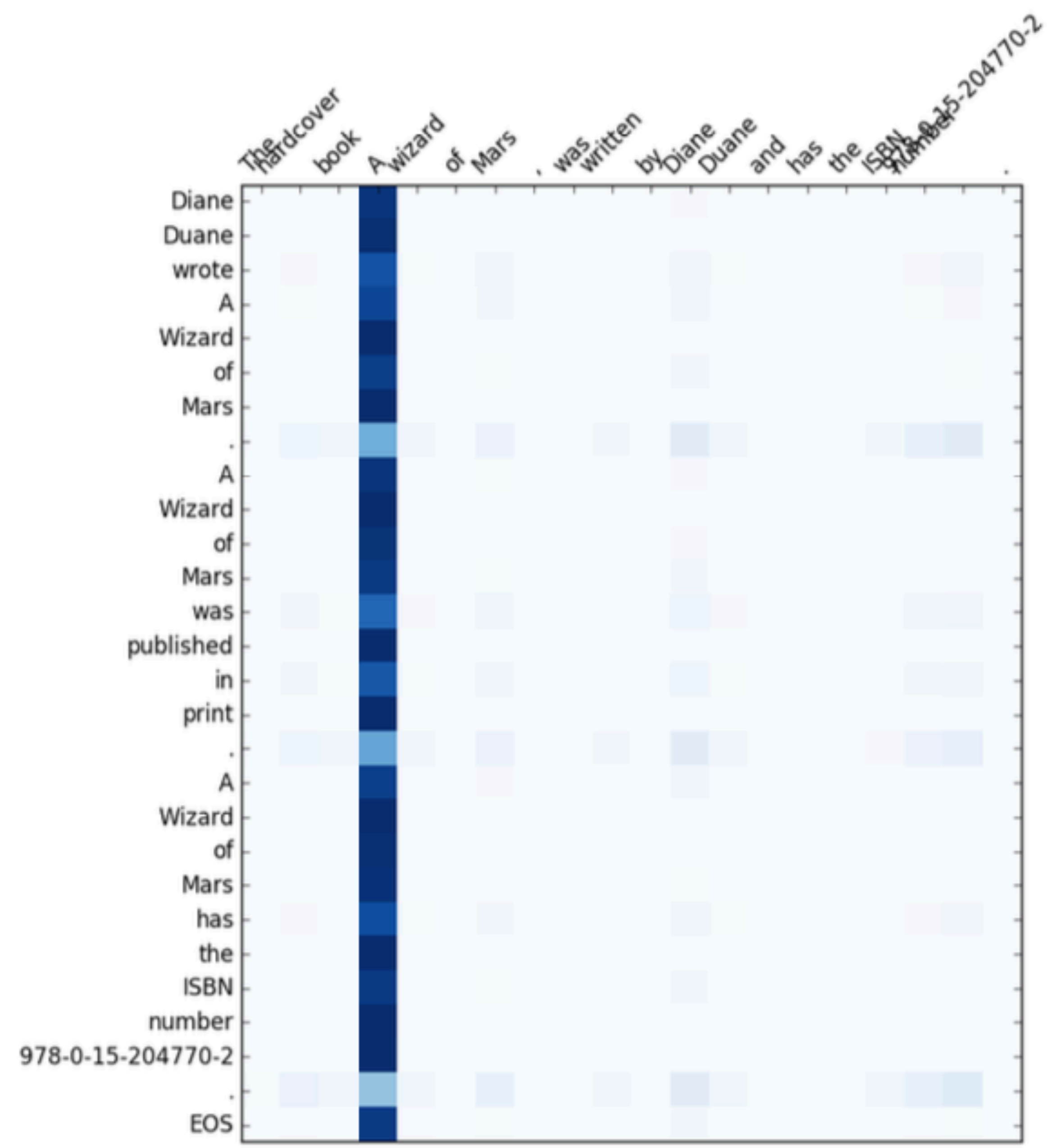
- Visualizing the attention weights we find an unexpected pattern

- The network mainly **attends to a single token** instead of spreading the attention



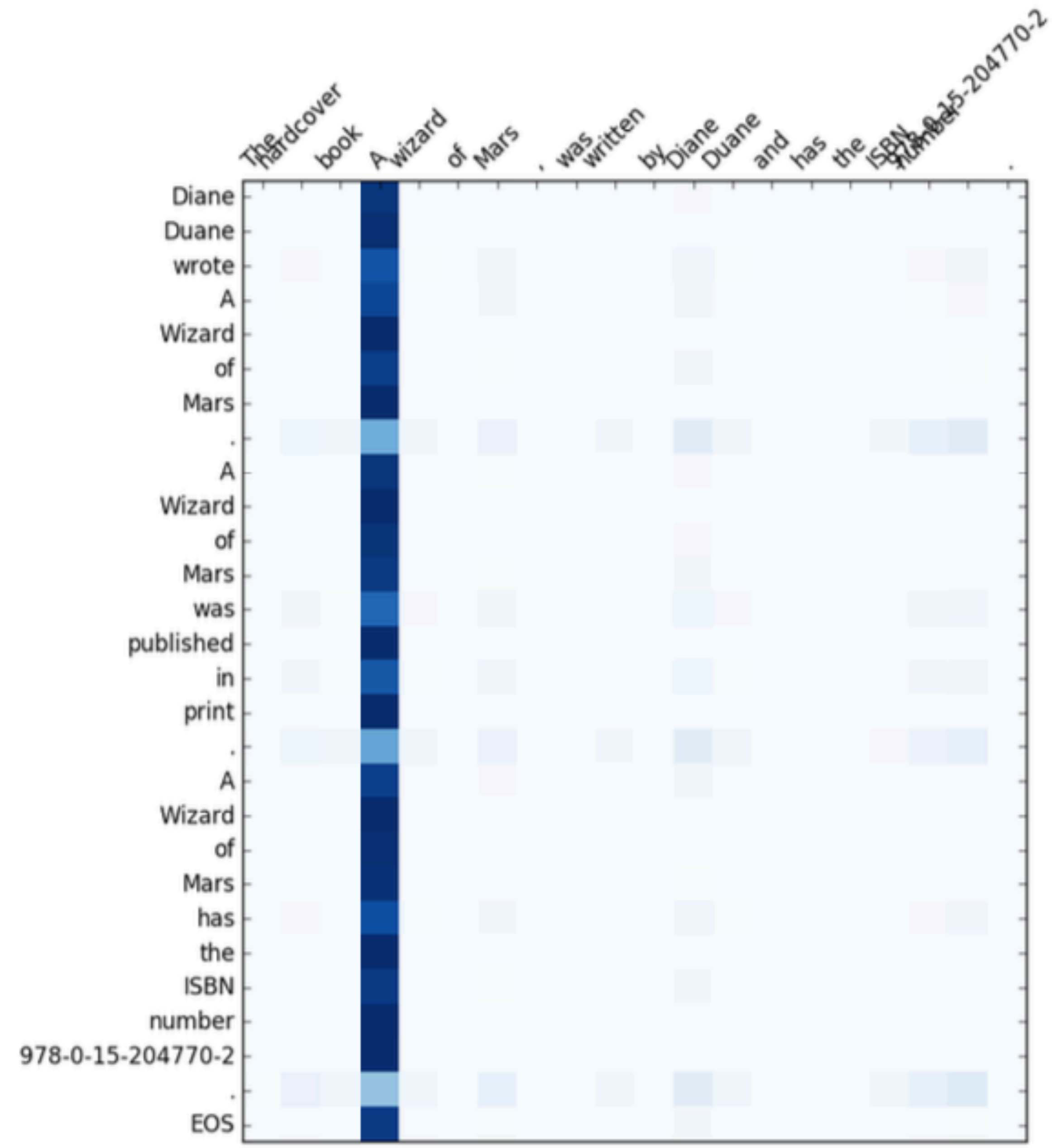
A Closer Look

- Visualizing the attention weights we find an unexpected pattern
- The network mainly **attends to a single token** instead of spreading the attention
- This token was usually a part of the **first mentioned entity**



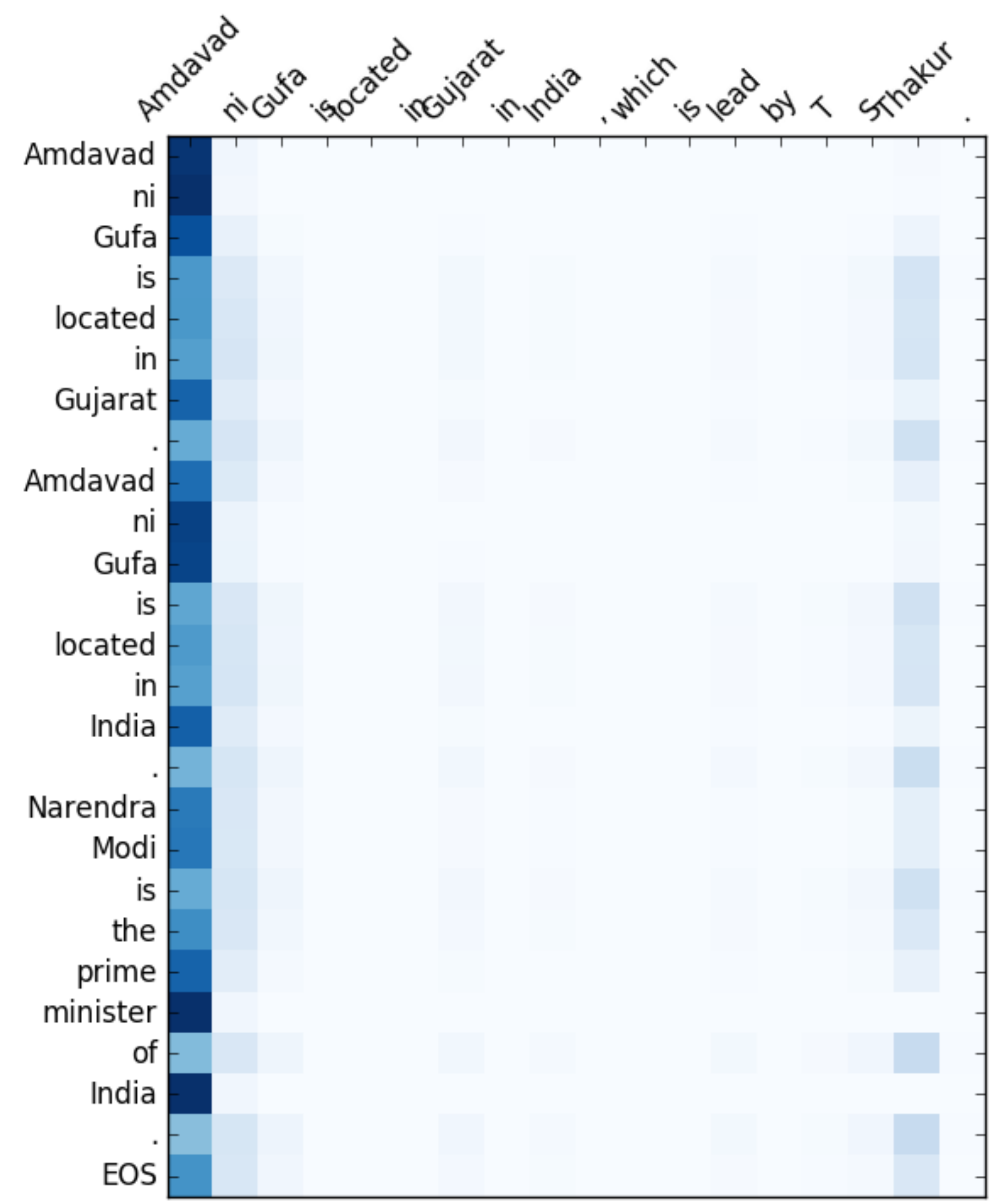
A Closer Look

- Visualizing the attention weights we find an unexpected pattern
- The network mainly **attends to a single token** instead of spreading the attention
- This token was usually a part of the **first mentioned entity**
- Consistent among different input examples



A Closer Look

- Visualizing the attention weights we find an unexpected pattern
- The network mainly **attends to a single token** instead of spreading the attention
- This token was usually a part of the **first mentioned entity**
- Consistent among different input examples



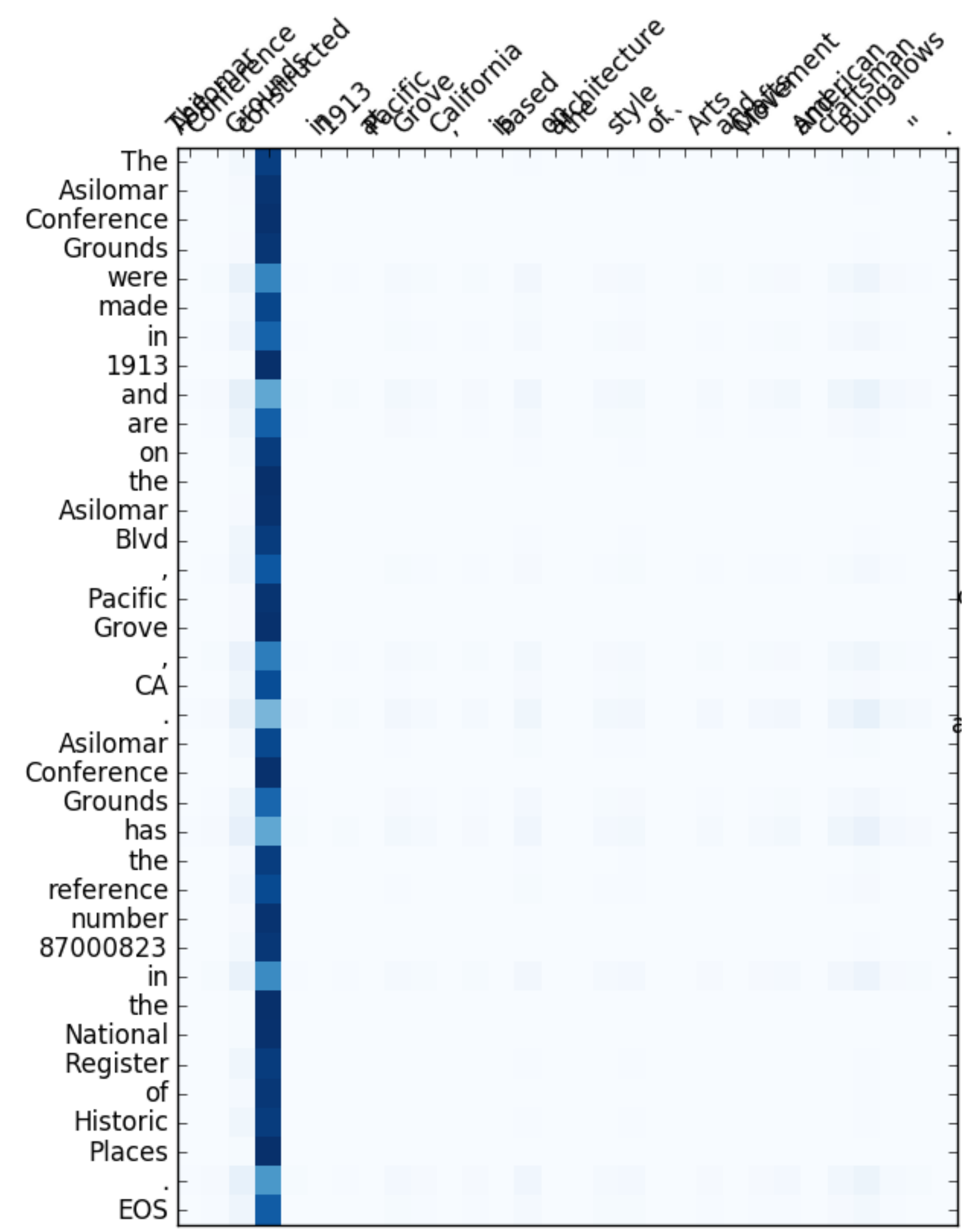
A Closer Look

- Visualizing the attention weights we find an unexpected pattern

- The network mainly **attends to a single token** instead of spreading the attention

- This token was usually a part of the **first mentioned entity**

- Consistent among different input examples



Testing for Over-Memorization

Testing for Over-Memorization

- In this stage we suspect that the network heavily **memorizes** entity-fact pairs

Testing for Over-Memorization

- In this stage we suspect that the network heavily **memorizes** entity-fact pairs
- We test this by introducing it with inputs consisting of repeated entities alone

Input	Prediction
Alan Shepard Alan Shepard Alan Shepard	

Testing for Over-Memorization

- In this stage we suspect that the network heavily **memorizes** entity-fact pairs
- We test this by introducing it with inputs consisting of repeated entities alone
- The network indeed generates facts it memorized about those specific entities

Input	Prediction
Alan Shepard Alan Shepard Alan Shepard	Alan Shepard is dead . Alan Shepard was a test pilot .

Testing for Over-Memorization

- In this stage we suspect that the network heavily **memorizes** entity-fact pairs
- We test this by introducing it with inputs consisting of repeated entities alone
- The network indeed generates facts it memorized about those specific entities

Input	Prediction
Alan Shepard Alan Shepard Alan Shepard	Alan Shepard is dead . Alan Shepard was a test pilot .
AFC Ajax AFC Ajax AFC Ajax	AFC Ajax 's manager is Jong Ajax . AFC Ajax N.V. own Sportpark De Toekomst .

Searching for the Cause: Dataset Artifacts

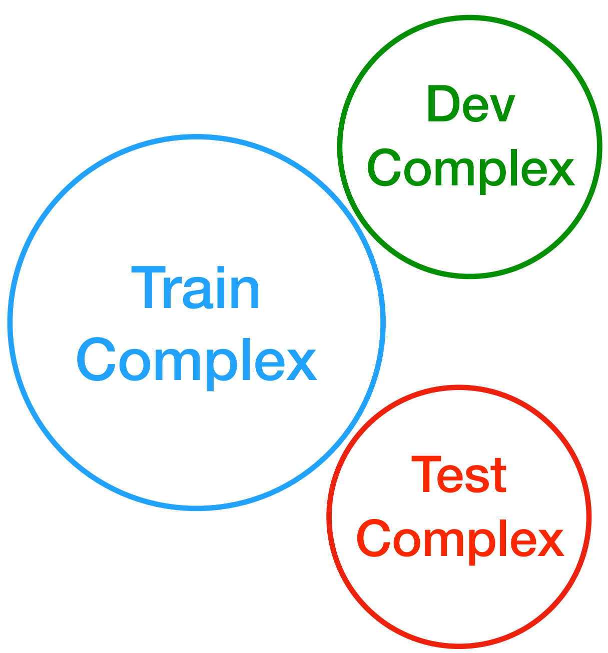
Searching for the Cause: Dataset Artifacts

- The original dataset included overlap between the training/development/test sets

Searching for the Cause: Dataset Artifacts

- The original dataset included overlap between the training/development/test sets
- When looking at the complex sentences side, there is no overlap

source



Searching for the Cause: Dataset Artifacts

- The original dataset included overlap between the training/development/test sets
- When looking at the complex sentences side, there is no overlap
- On the other hand, **most of the simple sentences** did overlap (~90%)

Searching for the Cause: Dataset Artifacts

- The original dataset included overlap between the training/development/test sets
- When looking at the complex sentences side, there is no overlap
- On the other hand, **most of the simple sentences** did overlap (~90%)
- Makes memorization very effective - “leakage” from train on the target side

New Data Split

New Data Split

- To remedy this, we construct a new data split by using the RDF information:

New Data Split

- To remedy this, we construct a new data split by using the RDF information:
    - Ensuring that all RDF **relation types** appear in the training set (enable generalization)

New Data Split

- To remedy this, we construct a new data split by using the RDF information:
  - Ensuring that all RDF **relation types** appear in the training set (enable generalization)
  - Ensuring that no RDF triple (fact) appears in two different sets (reduce memorization)

New Data Split

- To remedy this, we construct a new data split by using the RDF information:
  - Ensuring that all RDF **relation types** appear in the training set (enable generalization)
  - Ensuring that no RDF triple (fact) appears in two different sets (reduce memorization)
- The resulting dataset has no overlapping simple sentences

	Original Split	**New Split**
unique dev simple sentences in train	90.9%	**0.09%**
unique test simple sentences in train	89.8%	**0%**

New Data Split

- To remedy this, we construct a new data split by using the RDF information:
  - Ensuring that all RDF **relation types** appear in the training set (enable generalization)
  - Ensuring that no RDF triple (fact) appears in two different sets (reduce memorization)
- The resulting dataset has no overlapping simple sentences
- Has more unknown symbols in dev/test - **need better models!**

	Original Split	**New Split**
unique dev simple sentences in train	90.9%	**0.09%**
unique test simple sentences in train	89.8%	**0%**
% dev vocabulary in train	97.2%	**63%**
% test vocabulary in train	96.3%	**61.7%**

Copy Mechanism

Copy Mechanism

- To help with the increase in unknown words in the harder split, we incorporate a copy mechanism

Copy Mechanism

- To help with the increase in unknown words in the harder split, we incorporate a copy mechanism
    - Gu et al. 2016, See et al. 2017, Merity et al. 2017

Copy Mechanism

- To help with the increase in unknown words in the harder split, we incorporate a copy mechanism
  - Gu et al. 2016, See et al. 2017, Merity et al. 2017
- Uses a “copy switch” - feed-forward NN component with a sigmoid-activated scalar output

Copy Mechanism

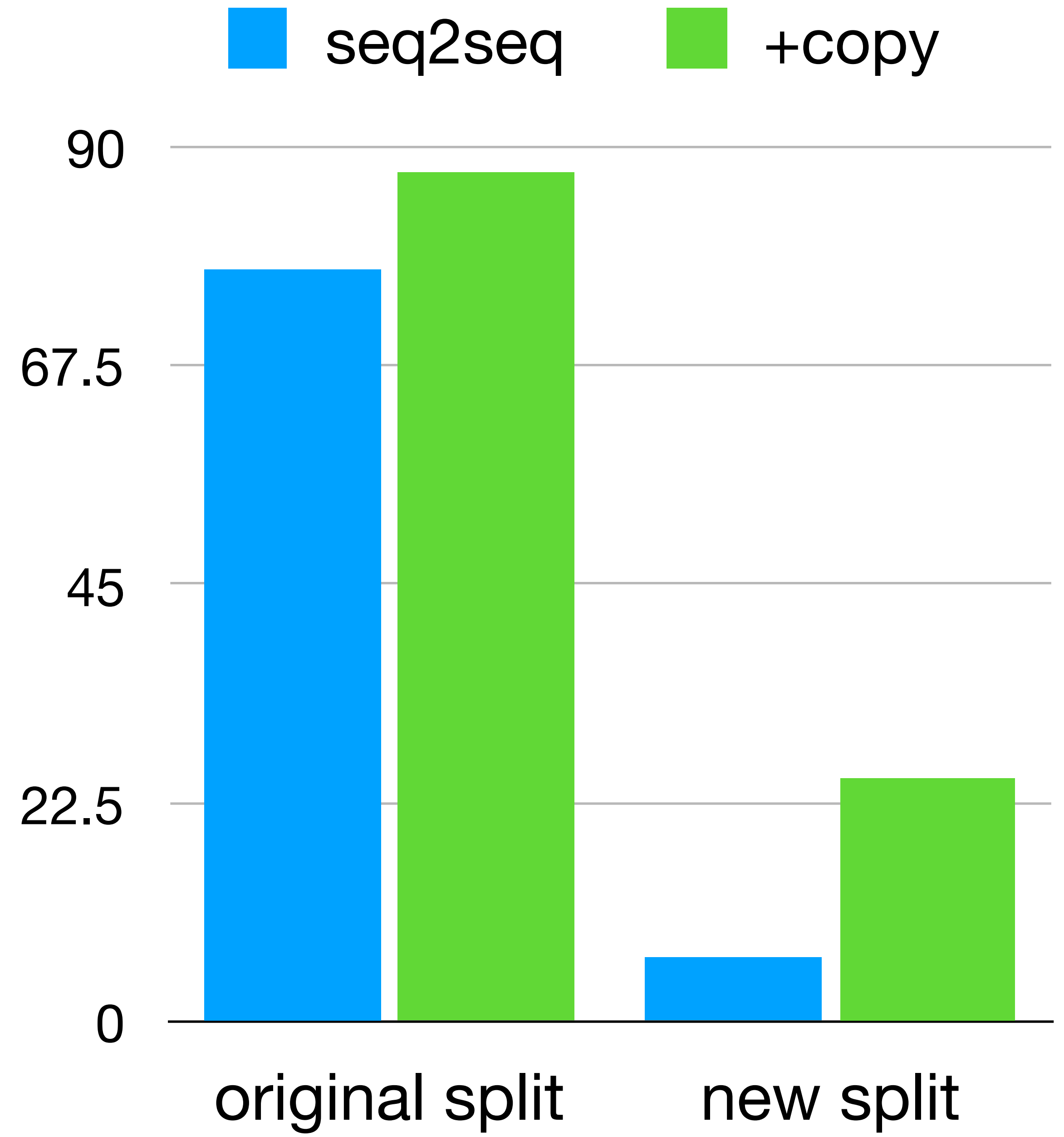
- To help with the increase in unknown words in the harder split, we incorporate a copy mechanism
  - Gu et al. 2016, See et al. 2017, Merity et al. 2017
- Uses a “copy switch” - feed-forward NN component with a sigmoid-activated scalar output
- Controls the **interpolation** of the softmax probabilities and the copy probabilities over the input tokens in each decoder step

$$p(w) = \underbrace{p(z=1)}^{\text{copy switch}}\overbrace{p_{copy}(w)}^{\text{attention weights (copy)}} + \overbrace{p(z=0)}^{\text{1 - copy switch}}\overbrace{p_{softmax}(w)}^{\text{softmax output}}$$

Results - New Split

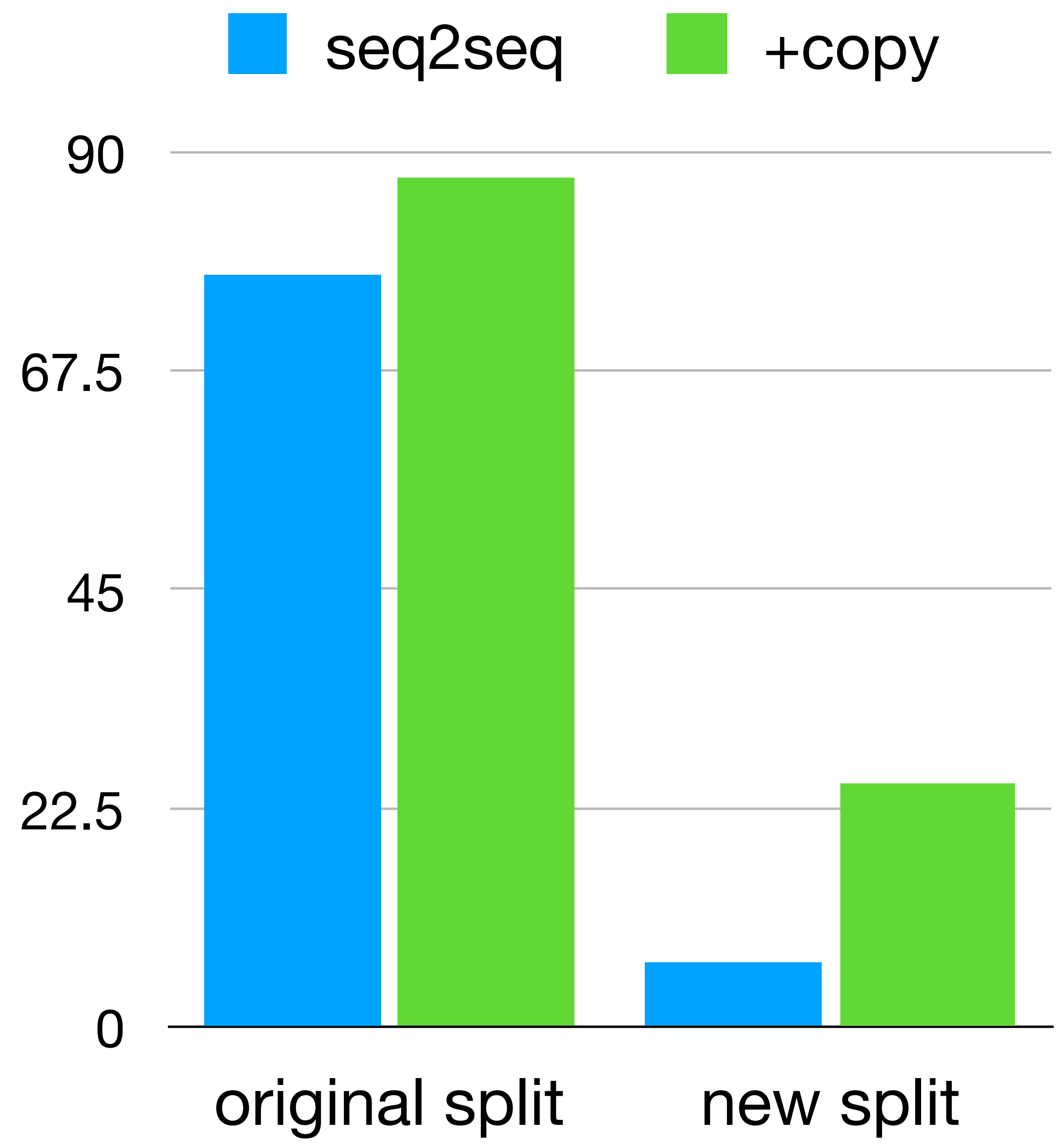
Results - New Split

- Baseline seq2seq models completely break (BLEU < 7) on the new split



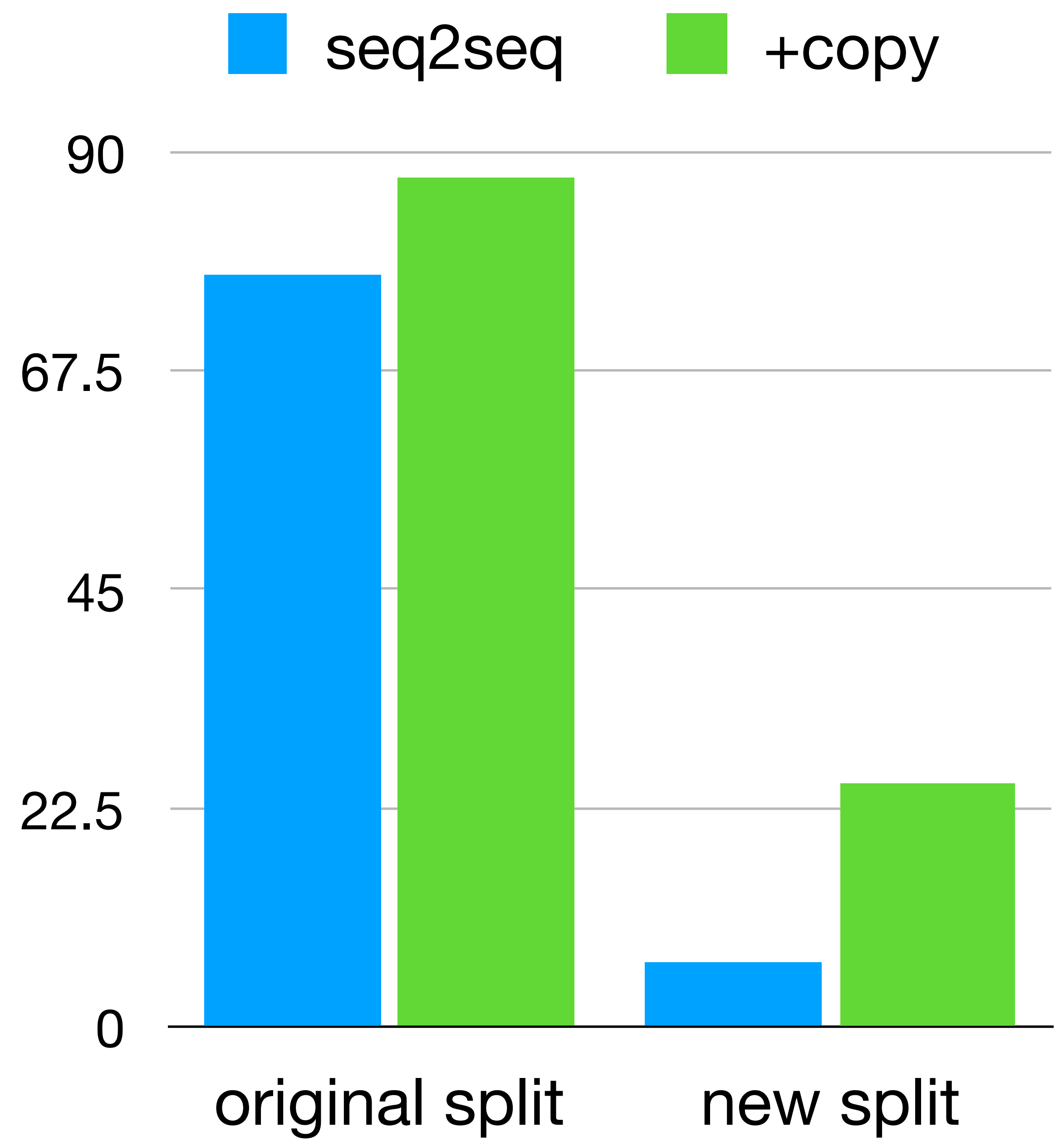
Results - New Split

- Baseline seq2seq models completely break (BLEU < 7) on the new split
- Copy mechanism helps to generalize



Results - New Split

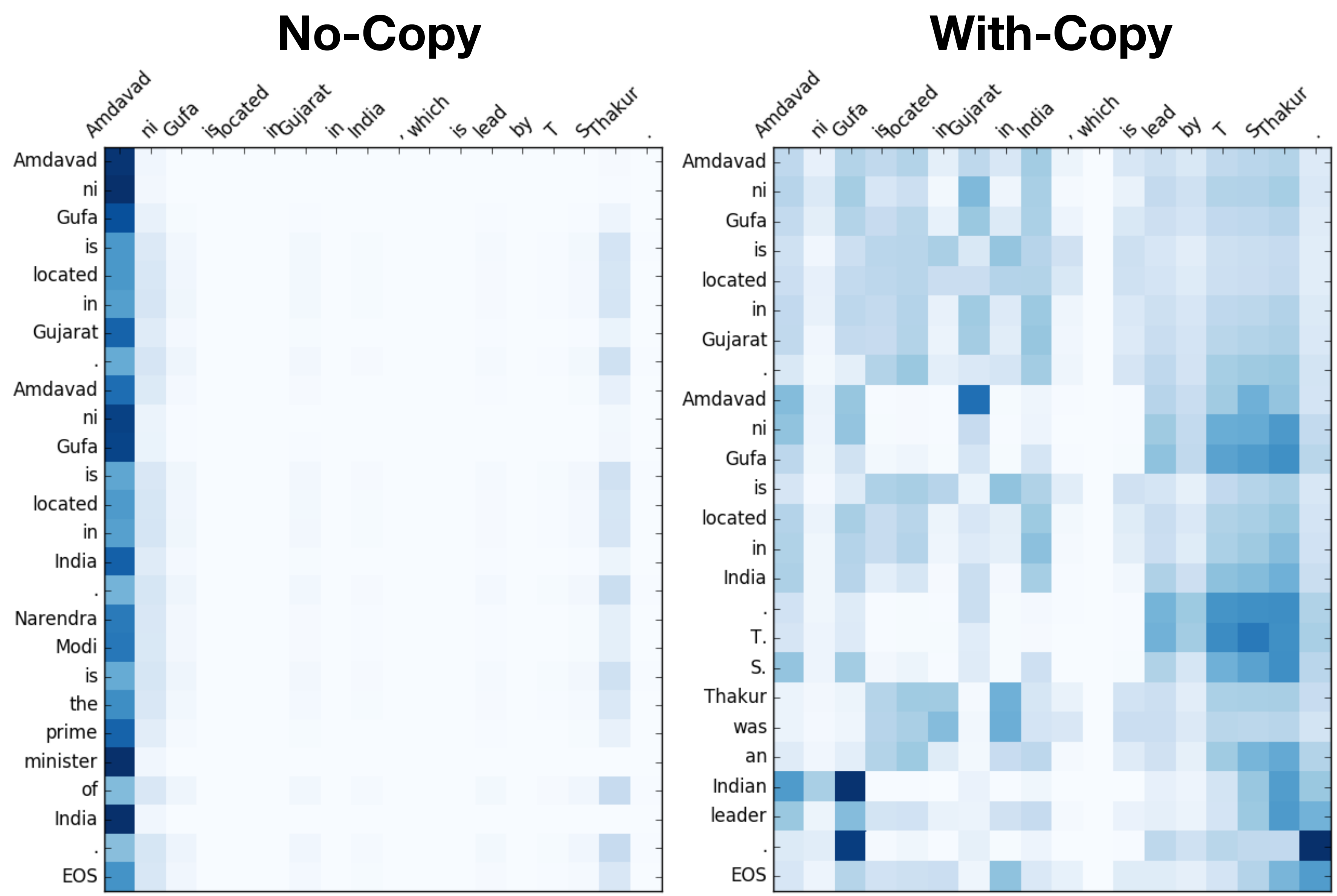
- Baseline seq2seq models completely break (BLEU < 7) on the new split
- Copy mechanism helps to generalize
- Much lower than the original benchmark - memorization was crucial for the high BLEU



Copying and Attention

Copying and Attention

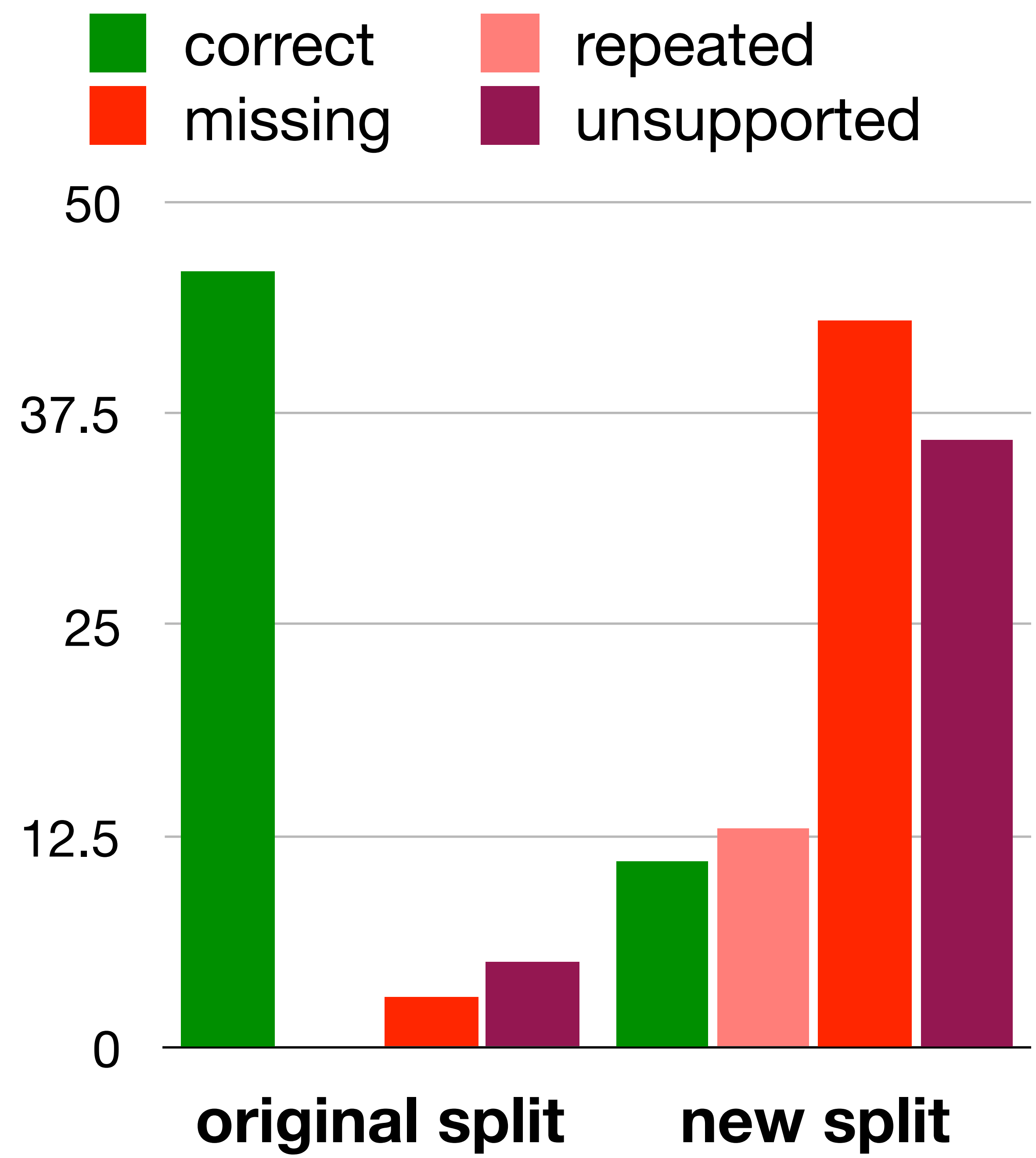
The copy-enhanced models spread the attention across the input tokens while improving results



Error Analysis

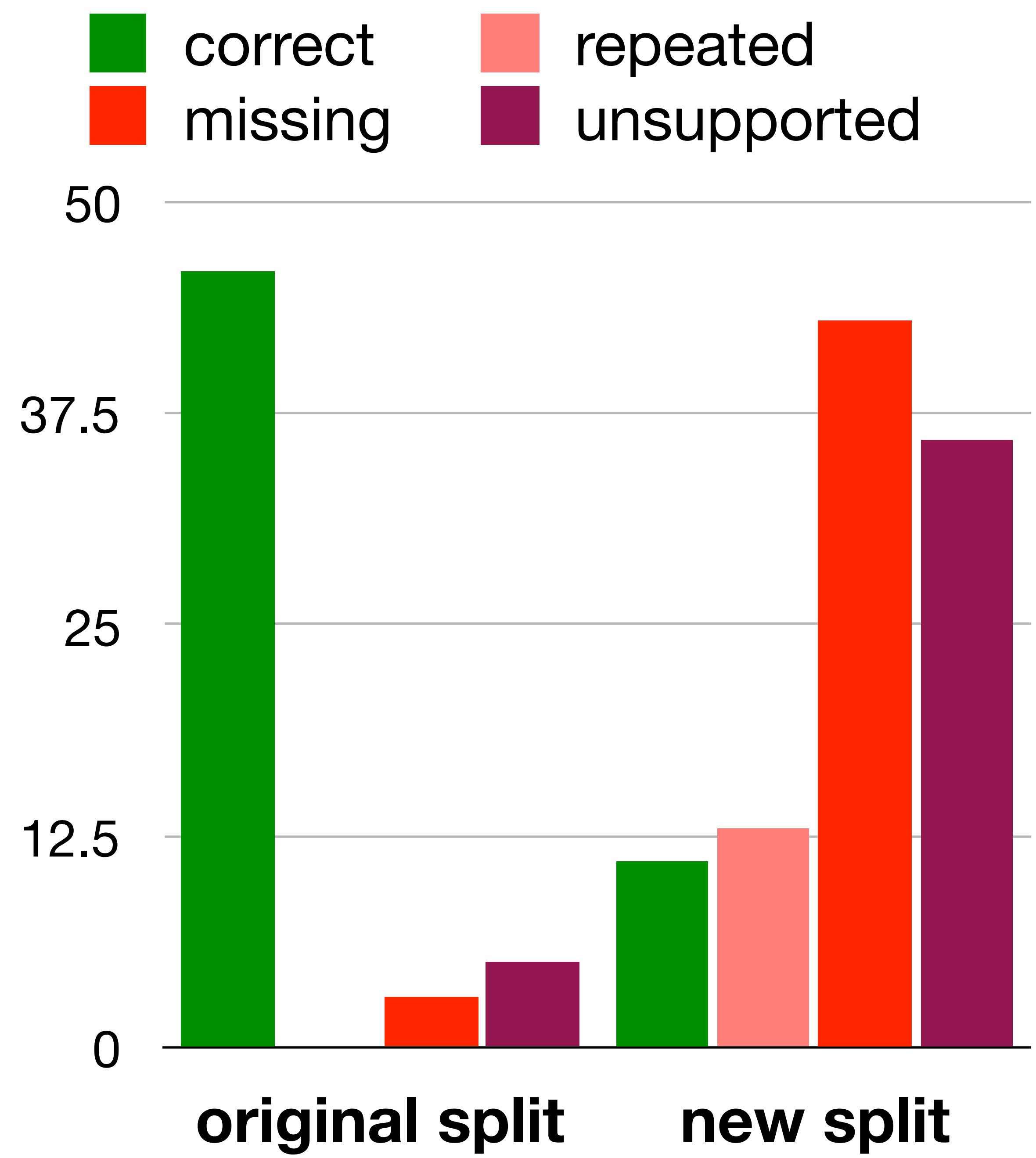
Error Analysis

- On the original split the models did very well (due to memorization) with up to 91% correct simple sentences



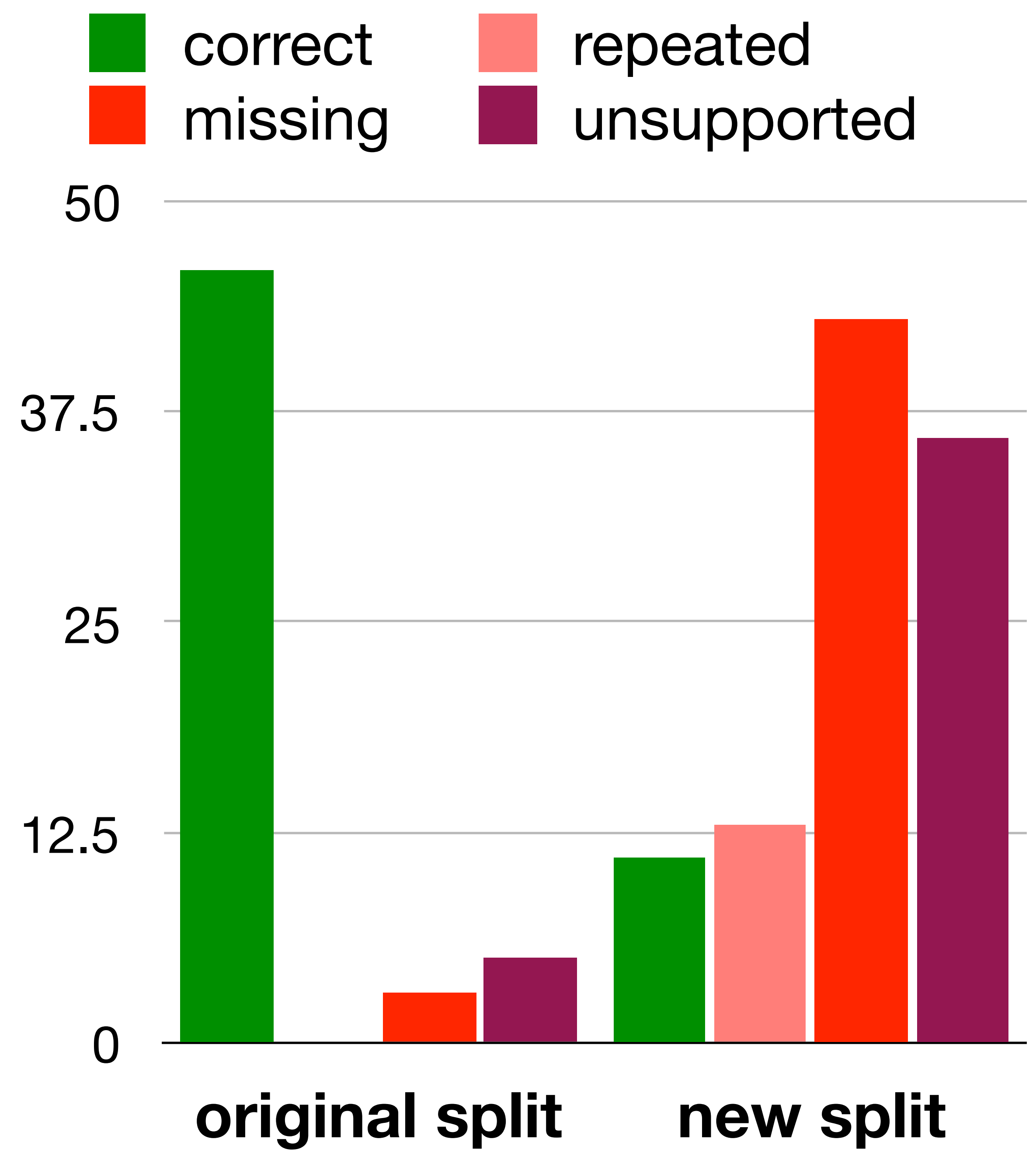
Error Analysis

- On the original split the models did very well (due to memorization) with up to 91% correct simple sentences
- On the new benchmark the best model got only up to 20% correct simple sentences



Error Analysis

- On the original split the models did very well (due to memorization) with up to 91% correct simple sentences
- On the new benchmark the best model got only up to 20% correct simple sentences
- The task is much more challenging then previously demonstrated



Conclusions

Conclusions

- Simple neural models seem to perform well due to **memorization**

Conclusions

- Simple neural models seem to perform well due to **memorization**
- We propose a **more challenging data split** for the task to discourage this

Conclusions

- Simple neural models seem to perform well due to **memorization**
- We propose a **more challenging data split** for the task to discourage this
  - A similar update was proposed by Narayan et al. in parallel to our work (**WebSplit v1.0**)

Conclusions

- Simple neural models seem to perform well due to **memorization**
- We propose a **more challenging data split** for the task to discourage this
  - A similar update was proposed by Narayan et al. in parallel to our work (**WebSplit v1.0**)
- We perform automatic evaluation and error analysis on the new benchmarks, showing that the task is **still far from being solved**

More Broadly

More Broadly

- **Creating datasets is hard!**

More Broadly

- **Creating datasets is hard!**
  - Think how models can “cheat"

More Broadly

- **Creating datasets is hard!**
  - Think how models can “cheat"
  - Create a challenging evaluation environment to capture generalization

More Broadly

- **Creating datasets is hard!**
    - Think how models can “cheat"
    - Create a challenging evaluation environment to capture generalization
    - Look for leakage of train to dev/test

More Broadly

- **Creating datasets is hard!**
  - Think how models can “cheat"
  - Create a challenging evaluation environment to capture generalization
  - Look for leakage of train to dev/test
- **Numbers can be misleading!**

More Broadly

- **Creating datasets is hard!**
  - Think how models can “cheat"
  - Create a challenging evaluation environment to capture generalization
  - Look for leakage of train to dev/test
- **Numbers can be misleading!**
  - Look at the data

More Broadly

- **Creating datasets is hard!**
  - Think how models can “cheat"
  - Create a challenging evaluation environment to capture generalization
  - Look for leakage of train to dev/test
- **Numbers can be misleading!**
  - Look at the data
  - Look at the model

More Broadly

- **Creating datasets is hard!**
  - Think how models can “cheat"
  - Create a challenging evaluation environment to capture generalization
  - Look for leakage of train to dev/test
- **Numbers can be misleading!**
  - Look at the data
  - Look at the model
  - Error analysis

Thank You!

Link to code and data is available in the paper :)