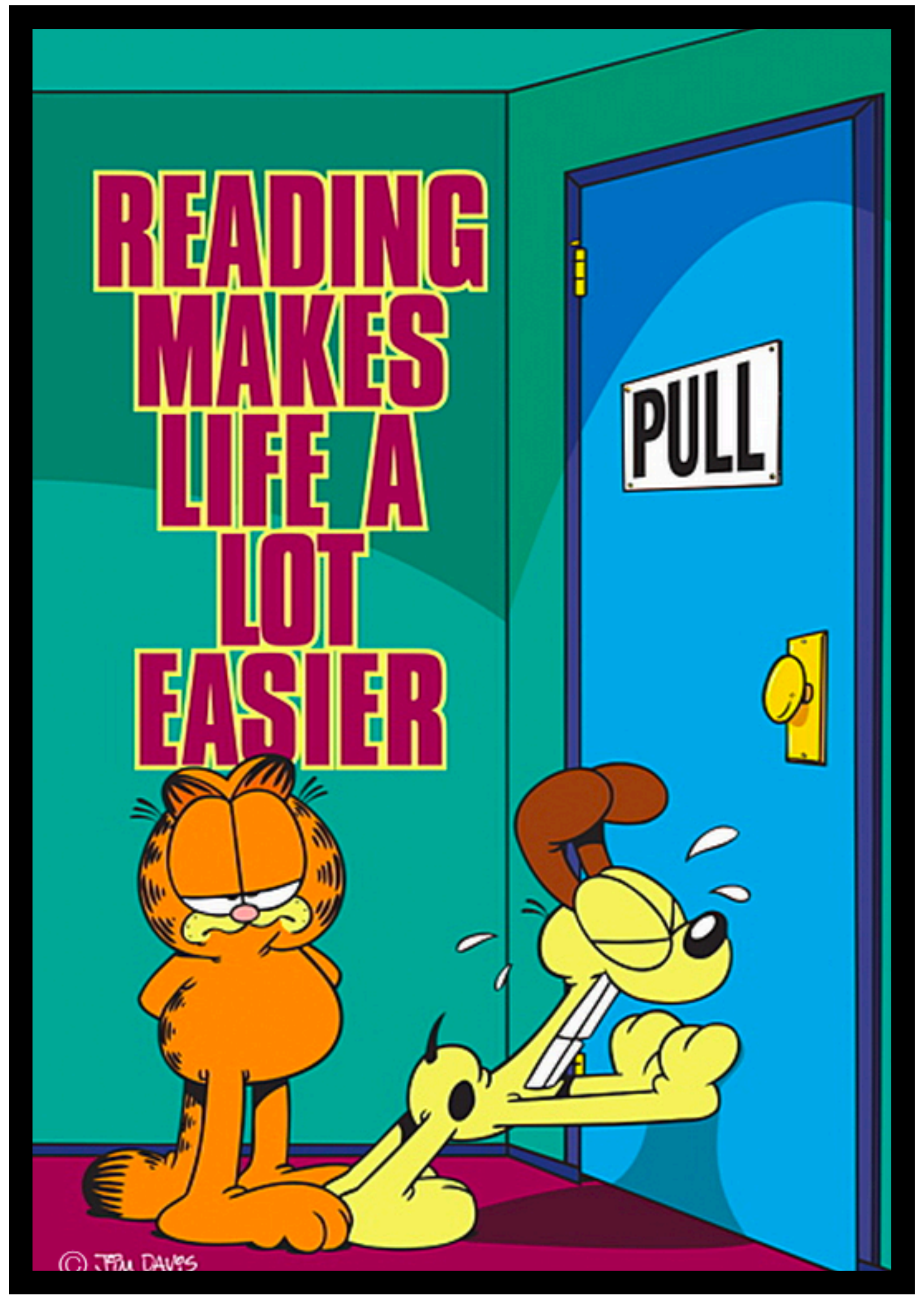


(Image Source: Garfield)

Importance of Data & Controllability in Neural Text Simplification

Wei Xu
School of Interactive Computing
Georgia Institute of Technology
@cocoweixu @cocoxu
https://cocoxu.github.io/

Today's Talk — Automatic Text Simplification

- **Controllable Text Generation**

  Also useful for natural language understanding, etc.

  - Neural semi-Markov CRF for Monolingual Word Alignment (Lan*, Jiang* & Xu, ACL 2021)

  How to incorporate linguistic rules with neural networks?

  - Controllable Text Simplification with Explicit Paraphrasing (Maddela, Alva-Manchego & Xu, NAACL 2021)

- **High-quality Training Data**

  Performance gains from better data are huge!

  - Neural CRF Model for Sentence Alignment in Text Simplification (Jiang, Maddela, Lan, Zhong & Xu, ACL 2020)

Text Simplification

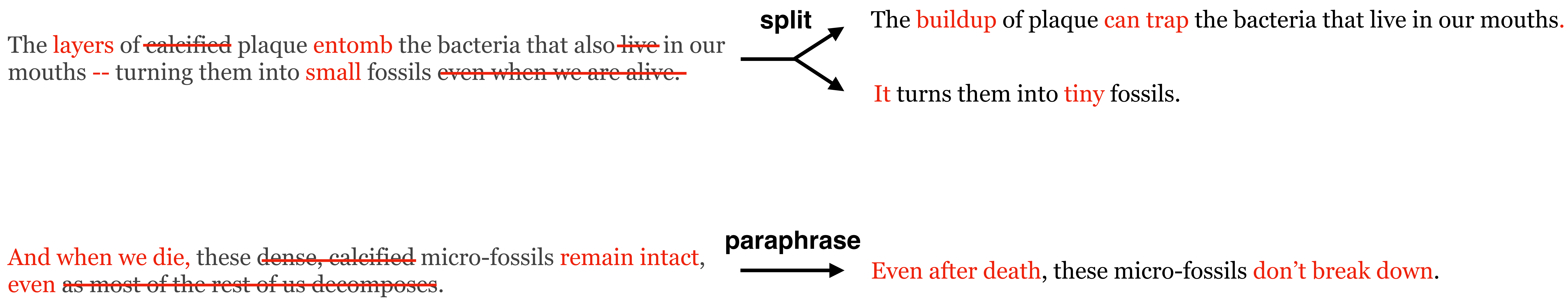
Rewrite complex text into simpler language while retain its original meaning.

The layers of calcified plaque entomb the bacteria that also live in our mouths -- turning them into small fossils even when we are alive.

And when we die, these dense, calcified micro-fossils remain intact, even as most of the rest of us decomposes.

Text Simplification

Rewrite complex text into simpler language while retain its original meaning.



The layers of ~~calcified~~ plaque entomb the bacteria that also ~~live~~ in our mouths -- turning them into small fossils ~~even when we are alive.~~

split

The buildup of plaque can trap the bacteria that live in our mouths.

It turns them into tiny fossils.

And when we die, these ~~dense, calcified~~ micro-fossils remain intact, even ~~as most of the rest of us decomposes~~.

paraphrase

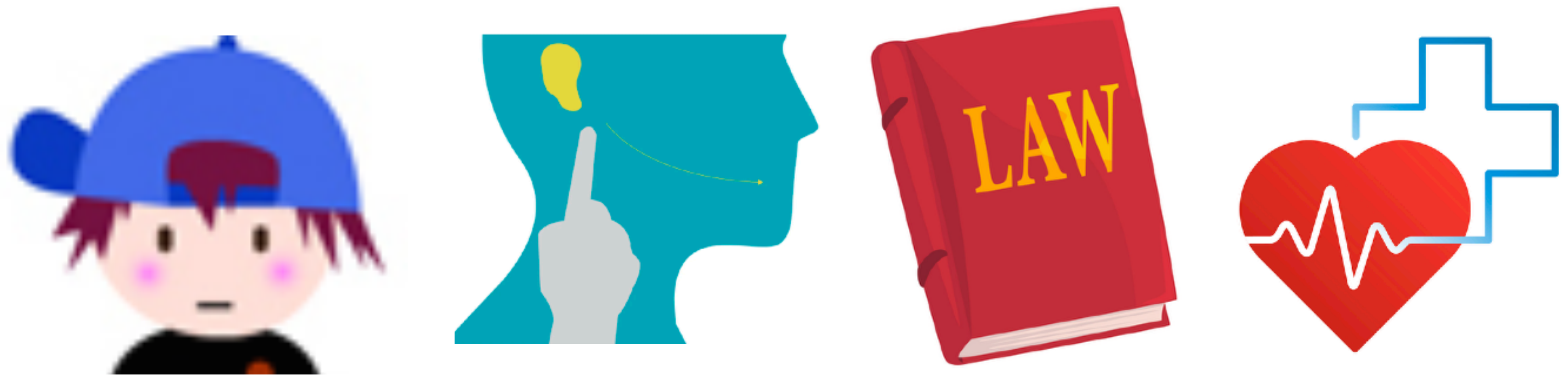
Even after death, these micro-fossils don’t break down.

Why Text Simplification?

It can help a lot of people!

- Children (Leonardo et al., 2018) ← research on education using Newsela data
- Second language learners (Housel et al., 2020)
- Deaf and hard-of-hearing students (Alonzo et al., 2020) ← using our EMNLP 2018 work on lexical simplification
- People with dyslexia (Rello at al., 2013)
- People with autism spectrum disorder (González-Navarro et al., 2014)

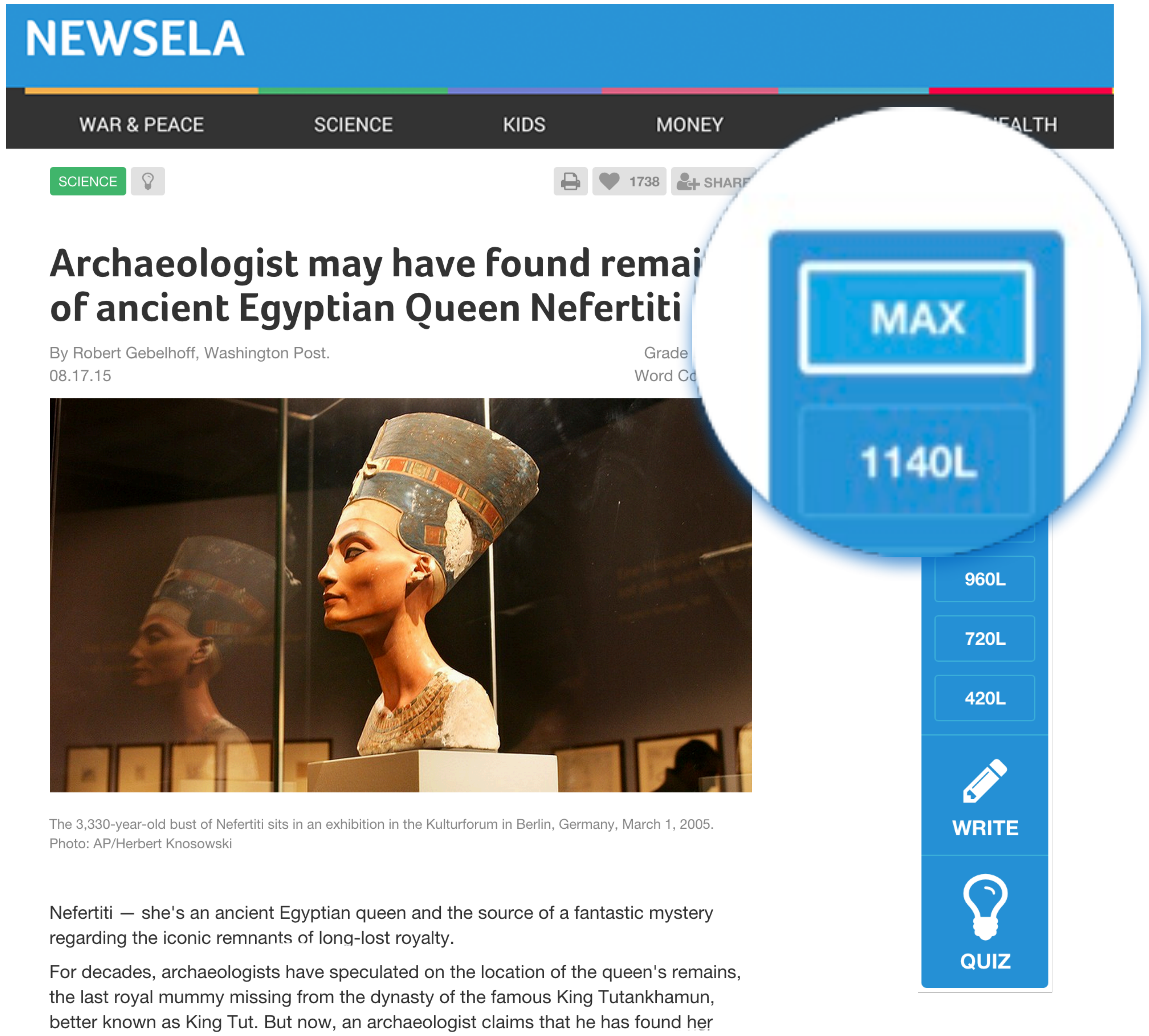
- and many others … e.g., to read medical & legal documents, etc.



Mounica Maddela, Wei Xu. "A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification" (EMNLP 2018)

Human Text Simplification

Professional editors rewrite news articles into 4 different readability levels for grade 3-12 students.



Archaeologist may have found remai
of ancient Egyptian Queen Nefertiti

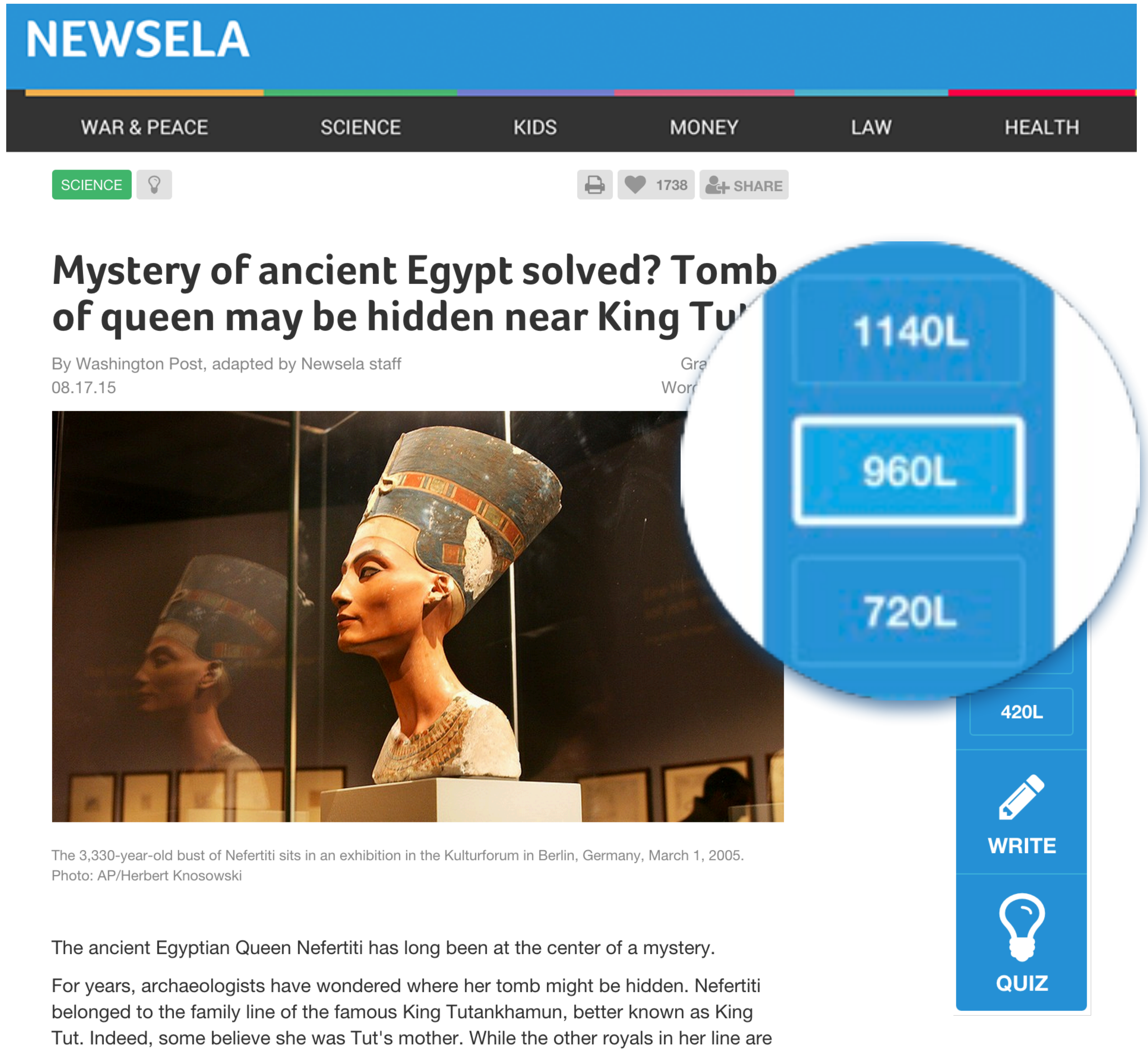
By Robert Gebelhoff, Washington Post.
08.17.15

Grade
Word Co

The 3,330-year-old bust of Nefertiti sits in an exhibition in the Kulturforum in Berlin, Germany, March 1, 2005.
Photo: AP/Herbert Knosowski

Nefertiti — she's an ancient Egyptian queen and the source of a fantastic mystery regarding the iconic remnants of long-lost royalty.

For decades, archaeologists have speculated on the location of the queen's remains, the last royal mummy missing from the dynasty of the famous King Tutankhamun, better known as King Tut. But now, an archaeologist claims that he has found her



Mystery of ancient Egypt solved? Tomb
of queen may be hidden near King Tu

By Washington Post, adapted by Newsela staff
08.17.15

Gra
Wor

The 3,330-year-old bust of Nefertiti sits in an exhibition in the Kulturforum in Berlin, Germany, March 1, 2005.
Photo: AP/Herbert Knosowski

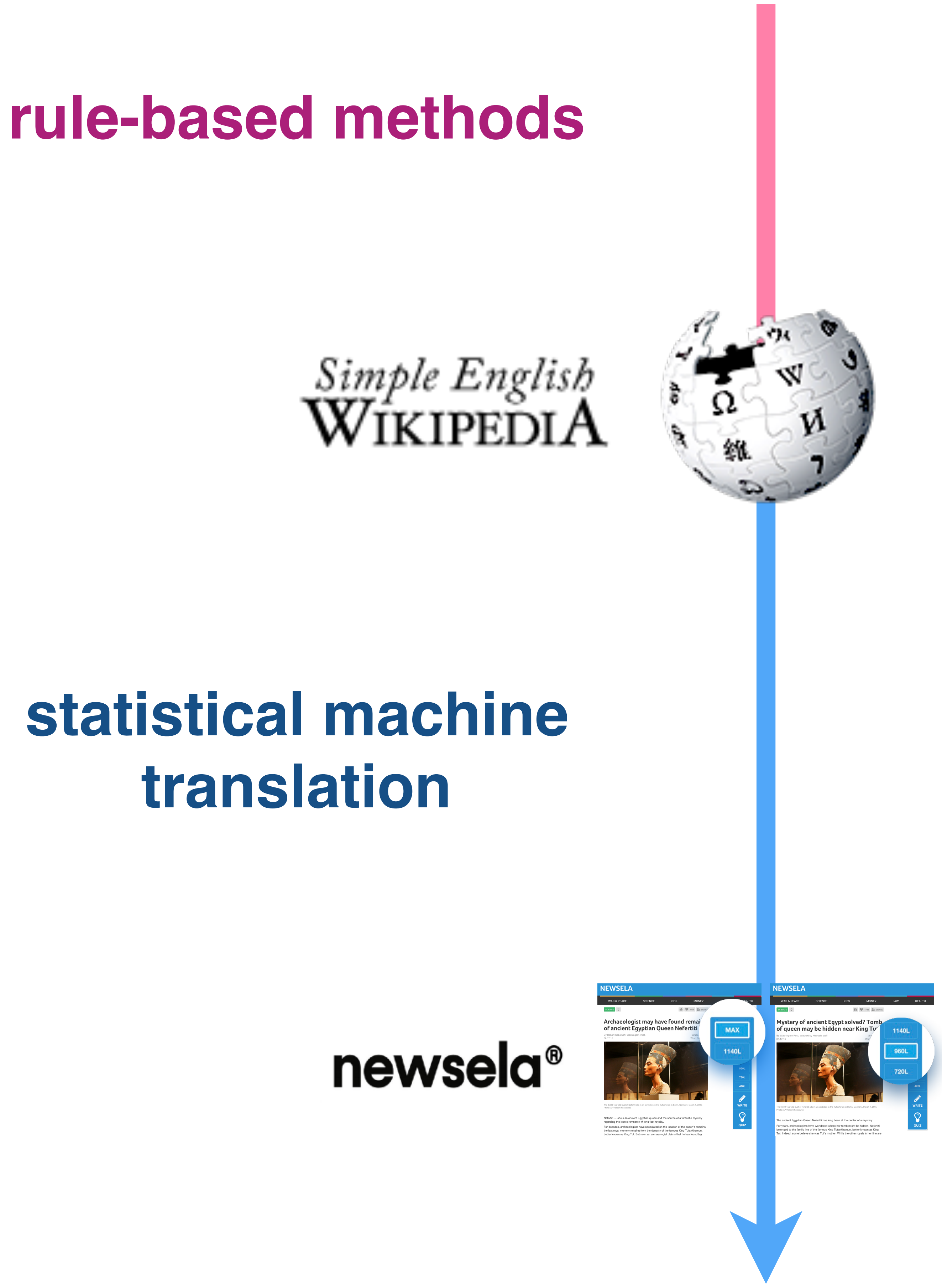
The ancient Egyptian Queen Nefertiti has long been at the center of a mystery.

For years, archaeologists have wondered where her tomb might be hidden. Nefertiti belonged to the family line of the famous King Tutankhamun, better known as King Tut. Indeed, some believe she was Tut's mother. While the other royals in her line are

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” (TACL 2015)
Yang Zhong, Chao Jiang, Wei Xu, Jessy Li. “Discourse Level Factors for Sentence Deletion in Text Simplification” (AAAI 2020)

Automatic Text Simplification

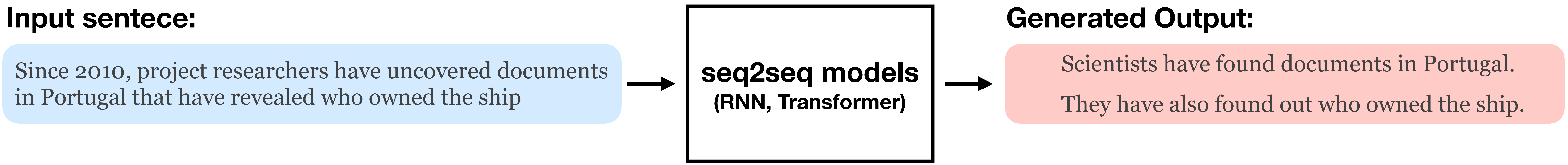
A brief history …



1997	Chandrasekar & Srinivas
1999	Dras (PhD thesis)
2000	Carroll, Minnen, Pearce, Canning, Devlin
2002	Canning (PhD thesis)
2004	Siddharthan (PhD thesis)
2010	**Zhu, Bernhard, Gurevych**
2011	Woodsend & Lapata
2011	Coster & Kauchak
2012	Wubben, van den Bosch, Krahmer
2014	Narayan & Gardent
2014	Siddharthan (Survey)
2014	Angrosh, Nomoto, Siddharthan
2014	Narayan (PhD thesis)
2015	**Xu, Callison-Burch, Napoles** “Problems in Current Text Simplification Research: New Data Can Help” (TACL 2015)
2016	**Xu, Napoles, Pavlick, Chen, Callison-Burch** “Optimizing Statistical Machine Translation for Simplification” (TACL 2016)

Automatic Text Simplification

Now, primarily addressed by sequence-to-sequence neural network models.



Input sentece:

Since 2010, project researchers have uncovered documents in Portugal that have revealed who owned the ship

seq2seq models
(RNN, Transformer)

Generated Output:

Scientists have found documents in Portugal.

They have also found out who owned the ship.

- **Some early works:**

  - LSTM model (Nisioi et al. 2017)
  - Transformer model (Zhao et al. 2018)

Automatic Text Simplification

However, SOTA neural generation models perform mostly deletion.

Input sentece:

According to Ledford, Northrop executives said they would build substantial parts of the bomber in Palmdale, creating about 1,500 jobs.

Generated output:

Programmer-interpreter (Dong et al., 2019)	ledford is a big group of bomber in palmdale.
Rerank (Kriz et al., 2019)	ledford is northrop.
Reinforcement Learning (Zhang & Lapata, 2017)	, said they would build palmdale parts of the substantial in creating.

Automatic Text Simplification

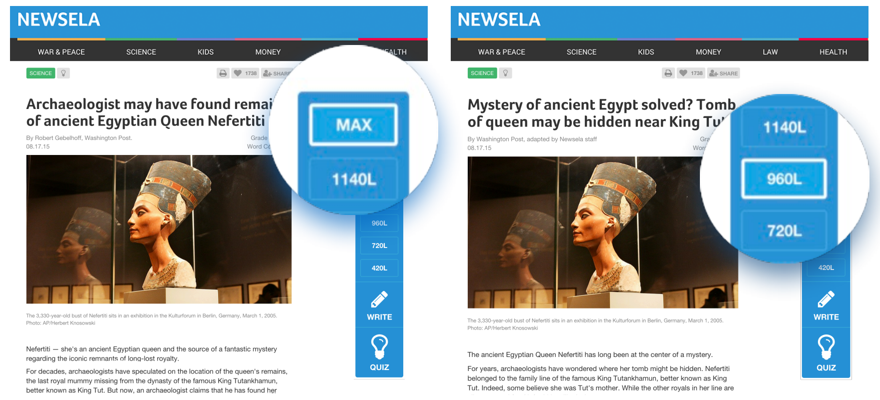
However, SOTA neural generation models perform mostly deletion.

Avg. length of input sentences is 20.7 tokens.

	Output-Length	New-Words	Identical-to-Input	Sentence-Split
Programmer-interpreter (Dong et al., 2019)	10.9	8.4%	4.6%	0%
Rerank (Kriz et al., 2019)	10.8	11.2%	1.2%	0%
Reinforcement Learning (Zhang & Lapata, 2017)	13.8	8.1%	16.8%	0%
Professional Editors	17.9	29.0%	0.0%	30.0%

Text Simplification Data

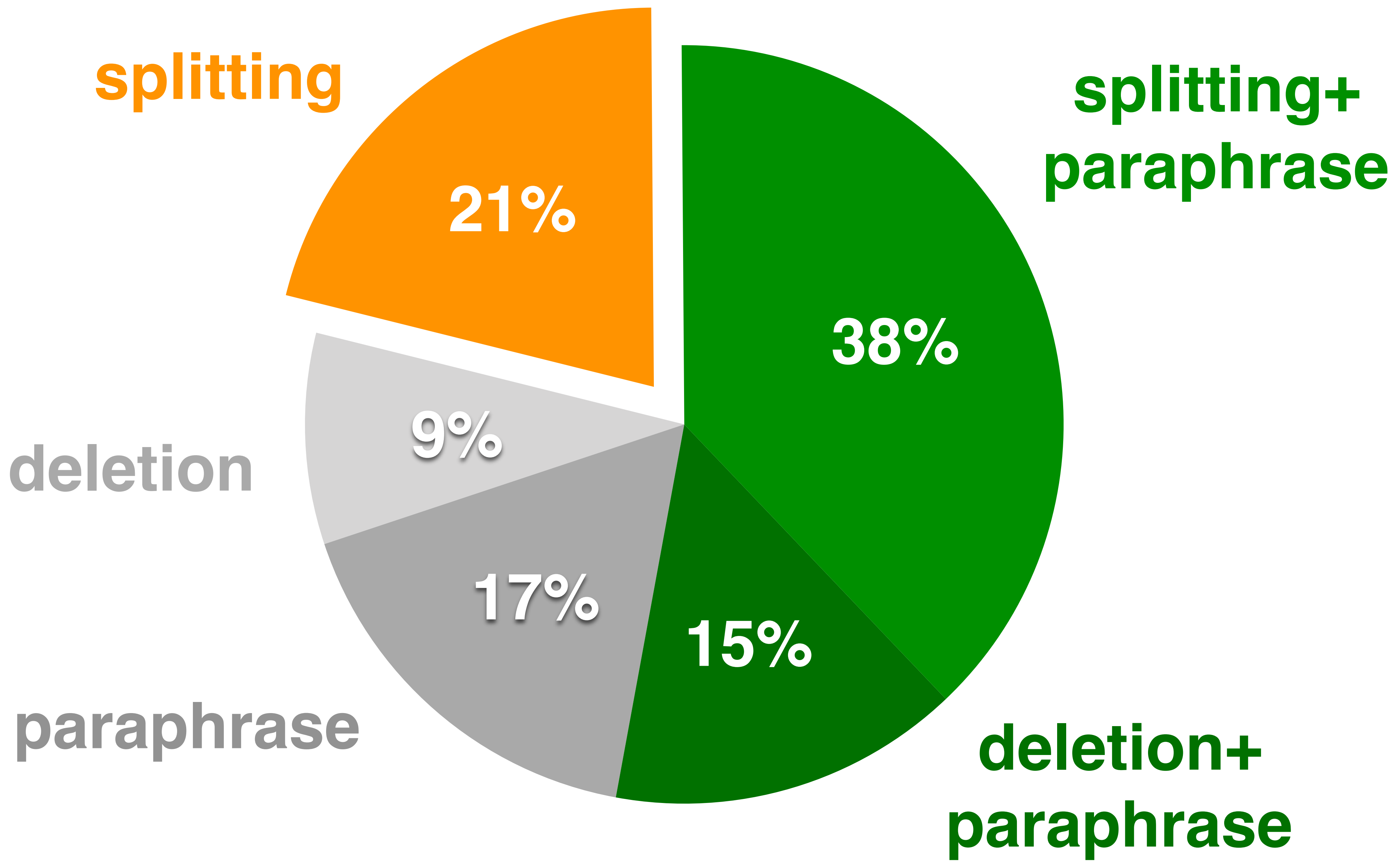
Professional editors use a sophisticated combination of rephrasing, splitting, and deletion.



NEWSELA
Archaeologist may have found remai of ancient Egyptian Queen Nefertiti
MAX
1140L
NEWSELA
Mystery of ancient Egypt solved? Tomb of queen may be hidden near King Tut
1140L
960L
720L

1882 news articles x 4 readability levels

Sentence Alignment →

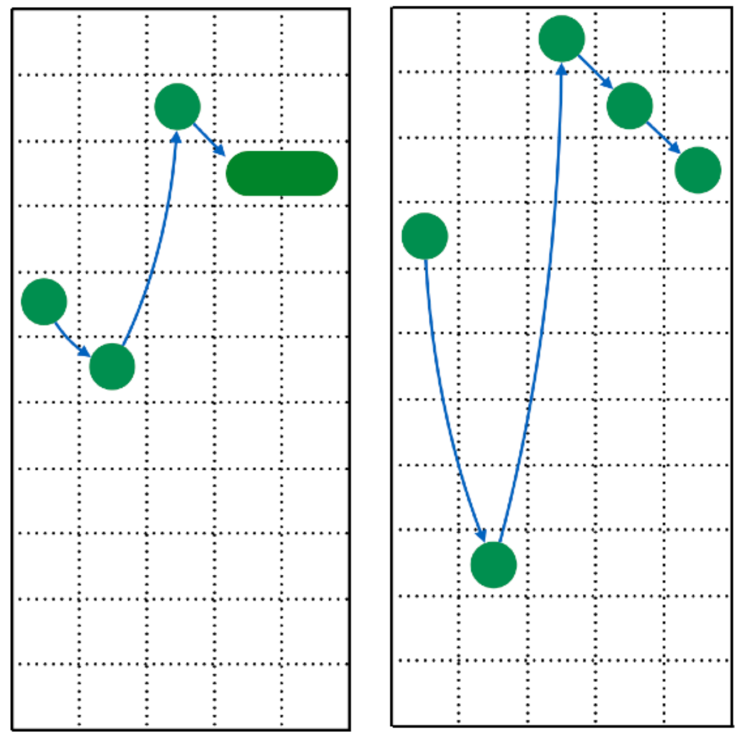


Newsela-Auto Corpus (Jiang et al. 2020)
666k sentence pairs

+ Wiki-Auto Corpus 488k sentence pairs

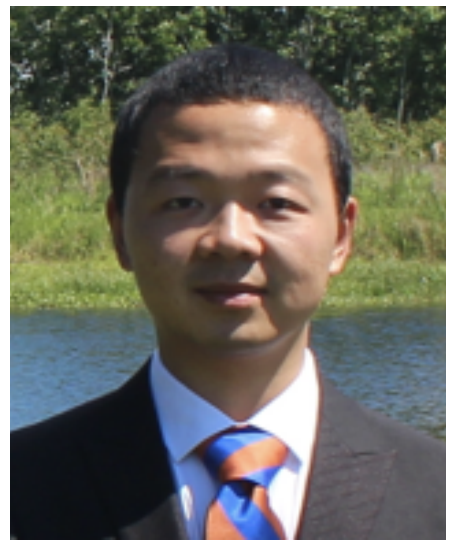
Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, Wei Xu. “Neural CRF Model for Sentence Alignment in Text Simplification” (ACL 2020)

Part 0 — Monolingual Word Alignment



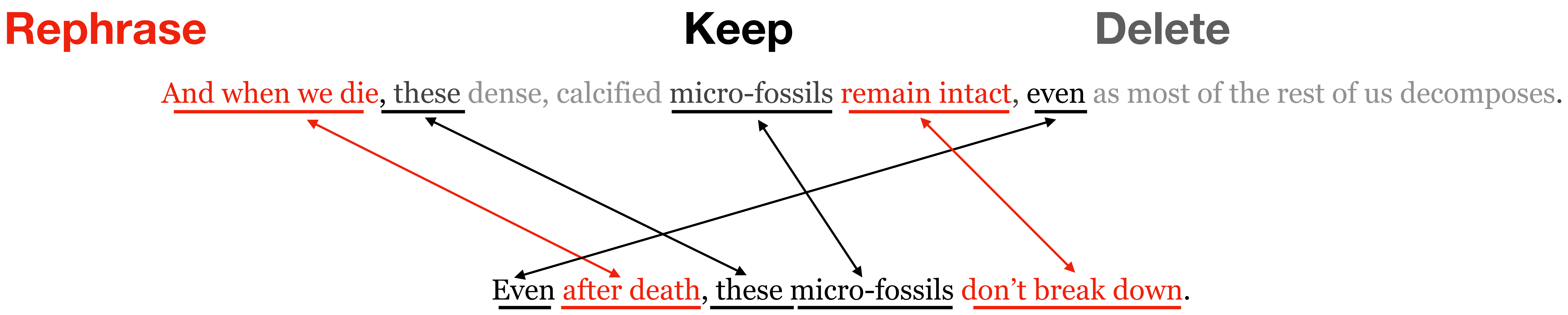
Neural semi-Markov CRF for Monolingual Word Alignment

Wuwei Lan*, Chao Jiang*, Wei Xu (ACL 2021)



Monolingual Word Alignment

Can support not only text-to-text generation tasks, but also natural language understanding tasks.



Rephrase

Keep

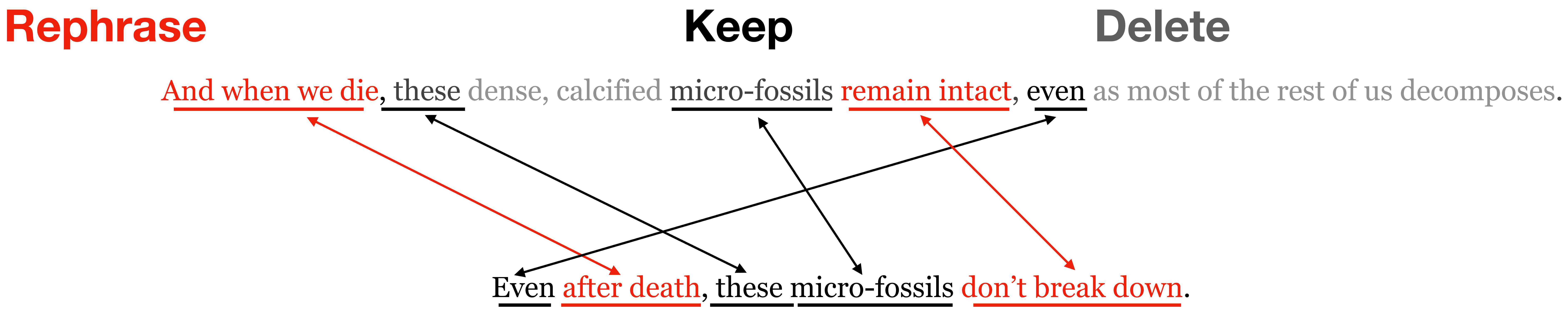
Delete

And when we die, these dense, calcified micro-fossils remain intact, even as most of the rest of us decomposes.

Even after death, these micro-fossils don’t break down.

Monolingual ~~Word~~ Span Alignment

Can support not only text-to-text generation tasks, but also natural language understanding tasks.



Rephrase

Keep

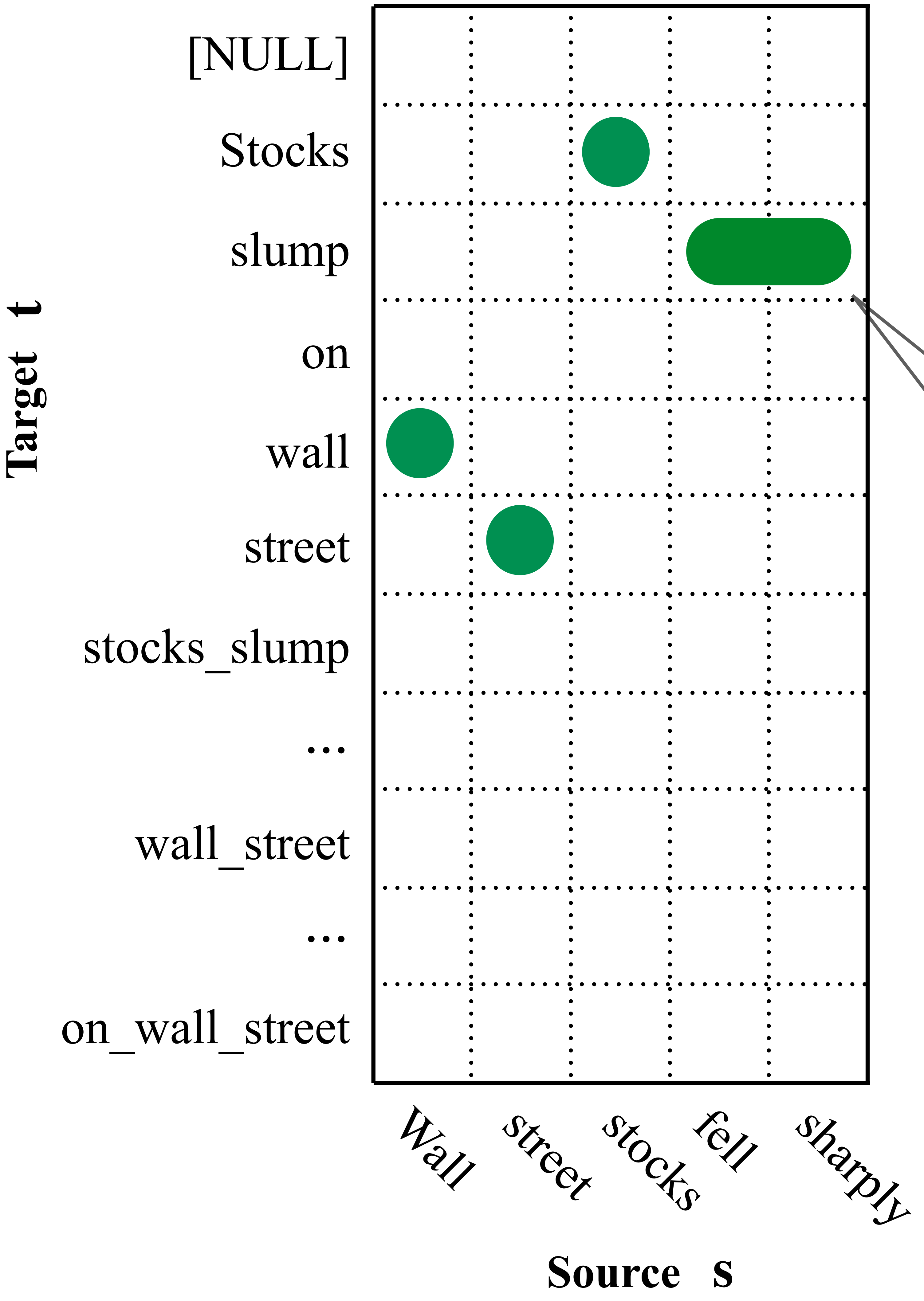
Delete

And when we die, these dense, calcified micro-fossils remain intact, even as most of the rest of us decomposes.

Even after death, these micro-fossils don’t break down.

Semi-CRF Word Alignment Model

Span Interaction Matrix



Span representation based on SpanBERT (Joshi et al. 2020)

$$h_i^s = (e_{start(i)}; e_{end(i)}; attn_i)$$

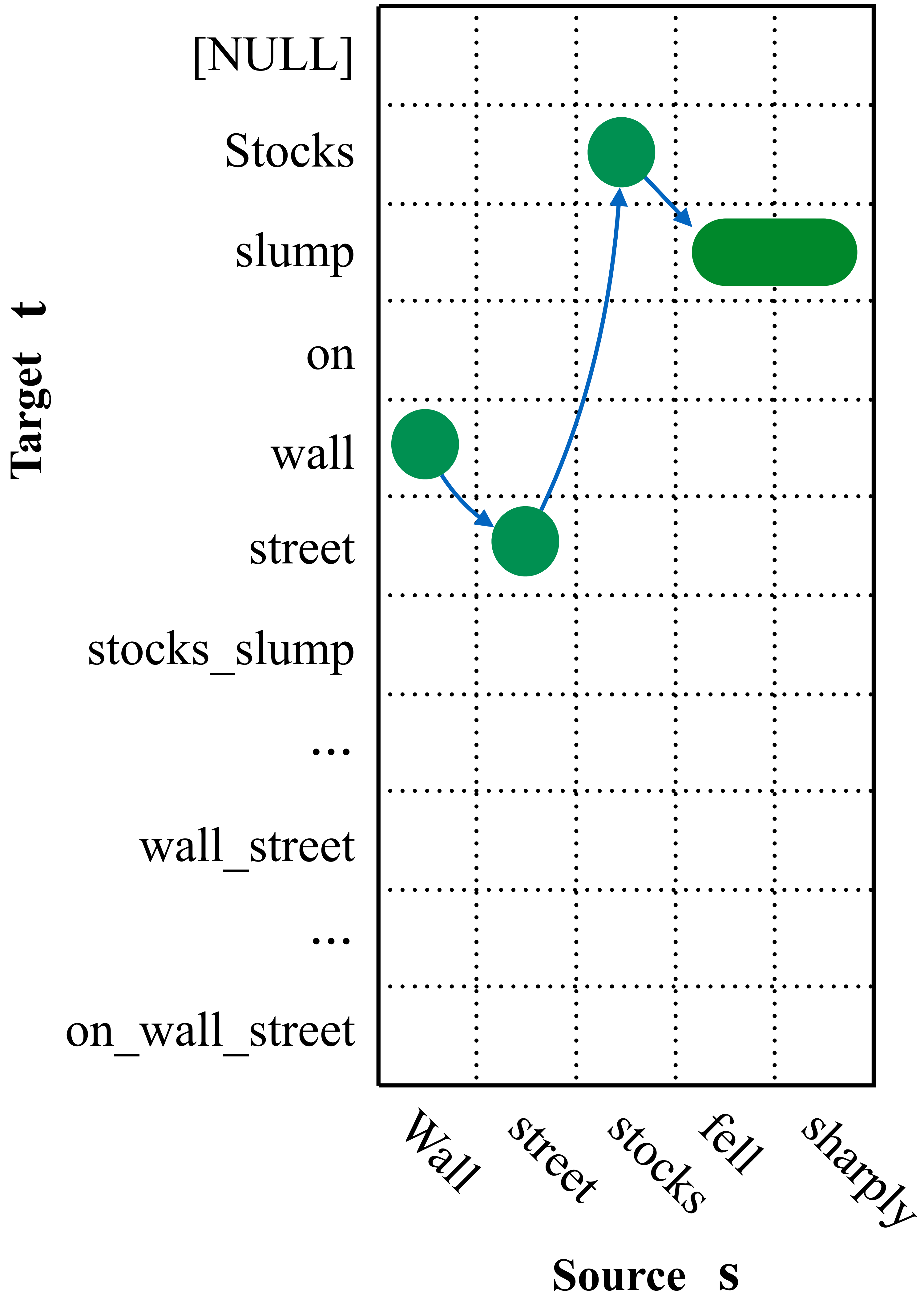
$$score(s_i, t_j) = \text{FFNN}(h_i^s; h_j^t; |h_i^s - h_j^t|; h_i^s \circ h_j^t)$$

2-layer FFNN to capture semantic similarity between (s_i, t_j)

Wuwei Lan*, Chao Jiang*, Wei Xu. Neural Semi-Markov CRF for Monolingual Word Alignment (ACL 2021)

Semi-CRF Word Alignment Model

Alignment Label Transition



semi-Markov Conditional Random Fields for span alignment

$$\Psi(\mathbf{a}, \mathbf{s}, \mathbf{t}) = \sum_i score(s_i, t_{a_i}) + T(a_{i-1}, a_i) + cost(\mathbf{a}, \mathbf{a}^*)$$

Negative Log-likelihood Loss **Hamming Loss**

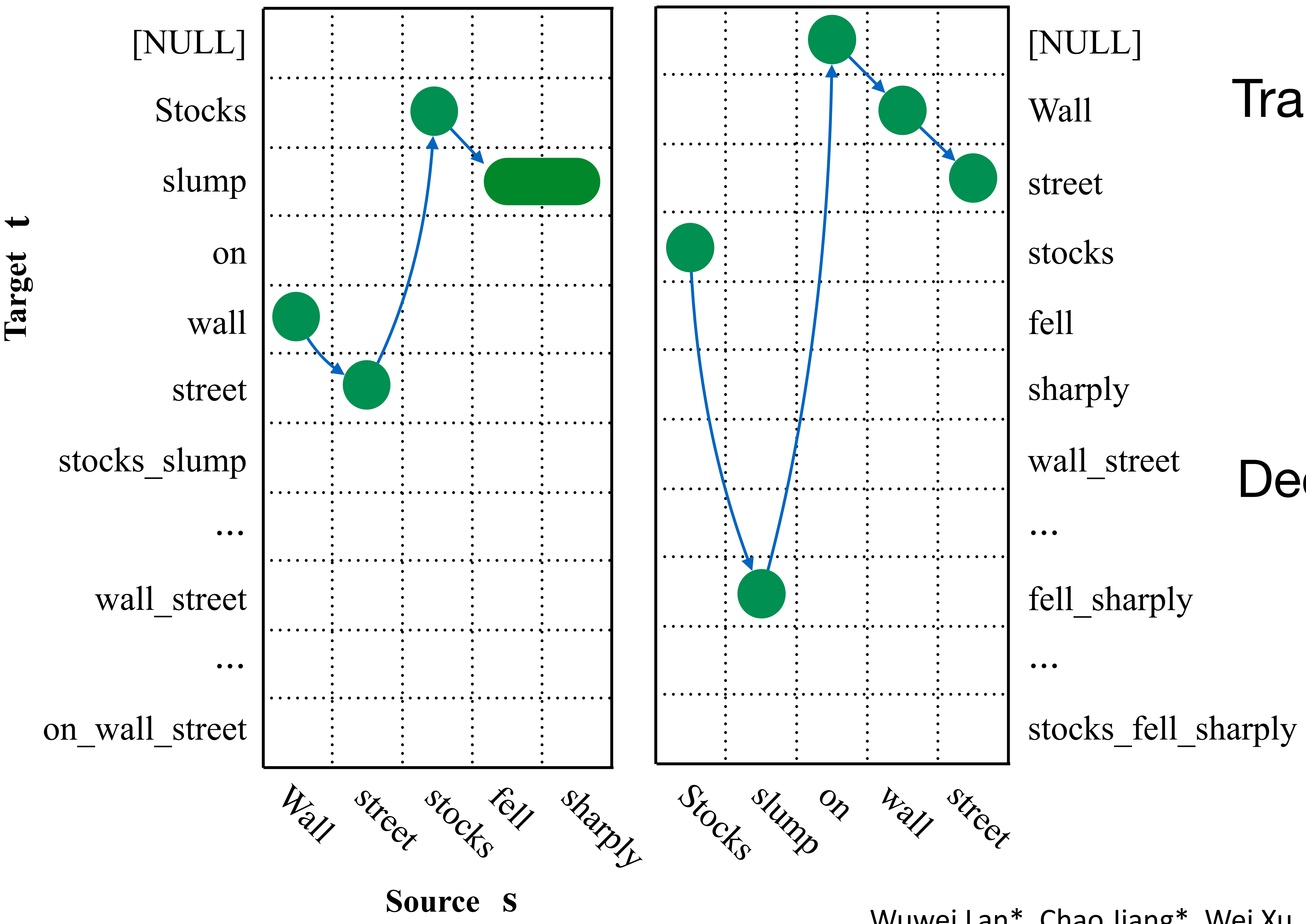
$$P(\mathbf{a} \,|\, \mathbf{s}, \mathbf{t}) = \frac{exp\left(\Psi(\mathbf{a}, \mathbf{s}, \mathbf{t})\right)}{\sum_{\mathbf{a} \in \mathbf{A}} exp\left(\Psi(\mathbf{a}, \mathbf{s}, \mathbf{t})\right)}$$

all possible alignments over variable length spans

Wuwei Lan*, Chao Jiang*, Wei Xu. Neural Semi-Markov CRF for Monolingual Word Alignment (ACL 2021)

Semi-CRF Word Alignment Model

Bi-directional Training / Decoding



Training objective:

$$\sum_{\mathbf{s},\mathbf{t},\mathbf{a}} - logP(\mathbf{a}_{s2t} \,|\, \mathbf{s}, \mathbf{t}) - logP(\mathbf{a}_{t2s} \,|\, \mathbf{t}, \mathbf{s})$$

Source-to-target **Target-to-source**

Decoding:

Viterbi-like Algorithm + Intersect + Expand

Wuwei Lan*, Chao Jiang*, Wei Xu. Neural Semi-Markov CRF for Monolingual Word Alignment (ACL 2021)

Experiments on MultiMWA Benchmark

We annotate a Multi-Genre Monolingual Word Alignment dataset that covers four different text genres.

	In-domain	Out-of-domain		
	MTReference	Newsela	arXiv	Wikipedia
JacanaToken (Yao et al. 2013a)	76.2	79.8	95.8	95.8
JacanaPhrase (Yao et al. 2013b)	75.8	79.4	93.7	94.9
PipelineAligner (Sultan et al. 2014)	74.8	80.3	96.5	97.1
Our Neural CRF aligner	90.8	86.6	95.7	97.0
Our Neural semi-CRF aligner	92.4	87.2	97.3	97.4
	🚀16.2 F1	🚀6.9 F1	🚀0.8 F1	🚀0.3 F1

Wuwei Lan*, Chao Jiang*, Wei Xu. Neural Semi-Markov CRF for Monolingual Word Alignment (ACL 2021)

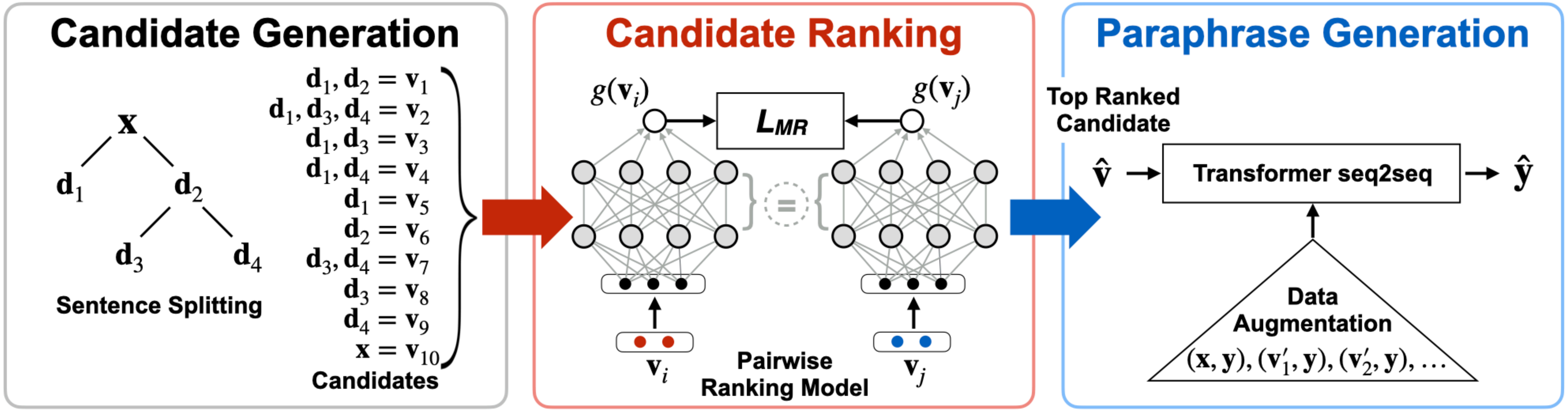
Part 1 — Controllable Generation Model

Controllable Text Simplification with Explicit Paraphrasing

Mounica Maddela, Fernando Alva-Manchego, Wei Xu (NAACL 2021)

Controllable Text Generation

- **Control over 3 edit operations** - deletion, splitting and paraphrasing.
- Incorporate linguistic rules with neural generation models.
- New setup to evaluate generation models's capability over these edit operations.


Candidate Generation
x
$\mathbf{d}_1$
$\mathbf{d}_2$
$\mathbf{d}_3$
$\mathbf{d}_4$
Sentence Splitting
$\mathbf{d}_1, \mathbf{d}_2 = \mathbf{v}_1$
$\mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_4 = \mathbf{v}_2$
$\mathbf{d}_1, \mathbf{d}_3 = \mathbf{v}_3$
$\mathbf{d}_1, \mathbf{d}_4 = \mathbf{v}_4$
$\mathbf{d}_1 = \mathbf{v}_5$
$\mathbf{d}_2 = \mathbf{v}_6$
$\mathbf{d}_3, \mathbf{d}_4 = \mathbf{v}_7$
$\mathbf{d}_3 = \mathbf{v}_8$
$\mathbf{d}_4 = \mathbf{v}_9$
$\mathbf{x} = \mathbf{v}_{10}$
Candidates
Candidate Ranking
$g(\mathbf{v}_i)$
L_{MR}
$g(\mathbf{v}_j)$
=
$\mathbf{v}_i$
$\mathbf{v}_j$
Pairwise Ranking Model
Paraphrase Generation
Top Ranked Candidate
$\hat{\mathbf{v}}$
Transformer seq2seq
$\hat{\mathbf{y}}$
Data Augmentation
$(\mathbf{x}, \mathbf{y}), (\mathbf{v}'_1, \mathbf{y}), (\mathbf{v}'_2, \mathbf{y}), \ldots$

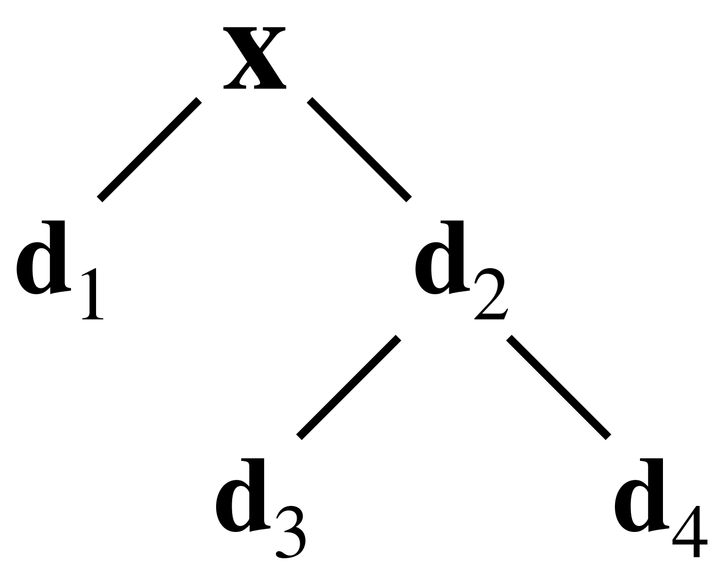
Mounica Maddela, Fernando Alva-Manchego, Wei Xu. "Controllable Text Simplification with Explicit Paraphrasing" (NAACL 2021)

Step 1 —

We use a rule-based method (Niklaus et al., 2019) + a seq2seq model for splitting and deletion.

- 35 hand-crafted grammar rules for English based on Stanford's parser (Socher et al., 2013).
- successfully split 92% of sentences with >= 20 words and make only 6.8% errors.

Candidate Generation



Sentence Splitting

$$\begin{aligned}
\mathbf{d}_1, \mathbf{d}_2 &= \mathbf{v}_1 \\
\mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_4 &= \mathbf{v}_2 \\
\mathbf{d}_1, \mathbf{d}_3 &= \mathbf{v}_3 \\
\mathbf{d}_1, \mathbf{d}_4 &= \mathbf{v}_4 \\
\mathbf{d}_1 &= \mathbf{v}_5 \\
\mathbf{d}_2 &= \mathbf{v}_6 \\
\mathbf{d}_3, \mathbf{d}_4 &= \mathbf{v}_7 \\
\mathbf{d}_3 &= \mathbf{v}_8 \\
\mathbf{d}_4 &= \mathbf{v}_9 \\
\mathbf{x} &= \mathbf{v}_{10}
\end{aligned}$$

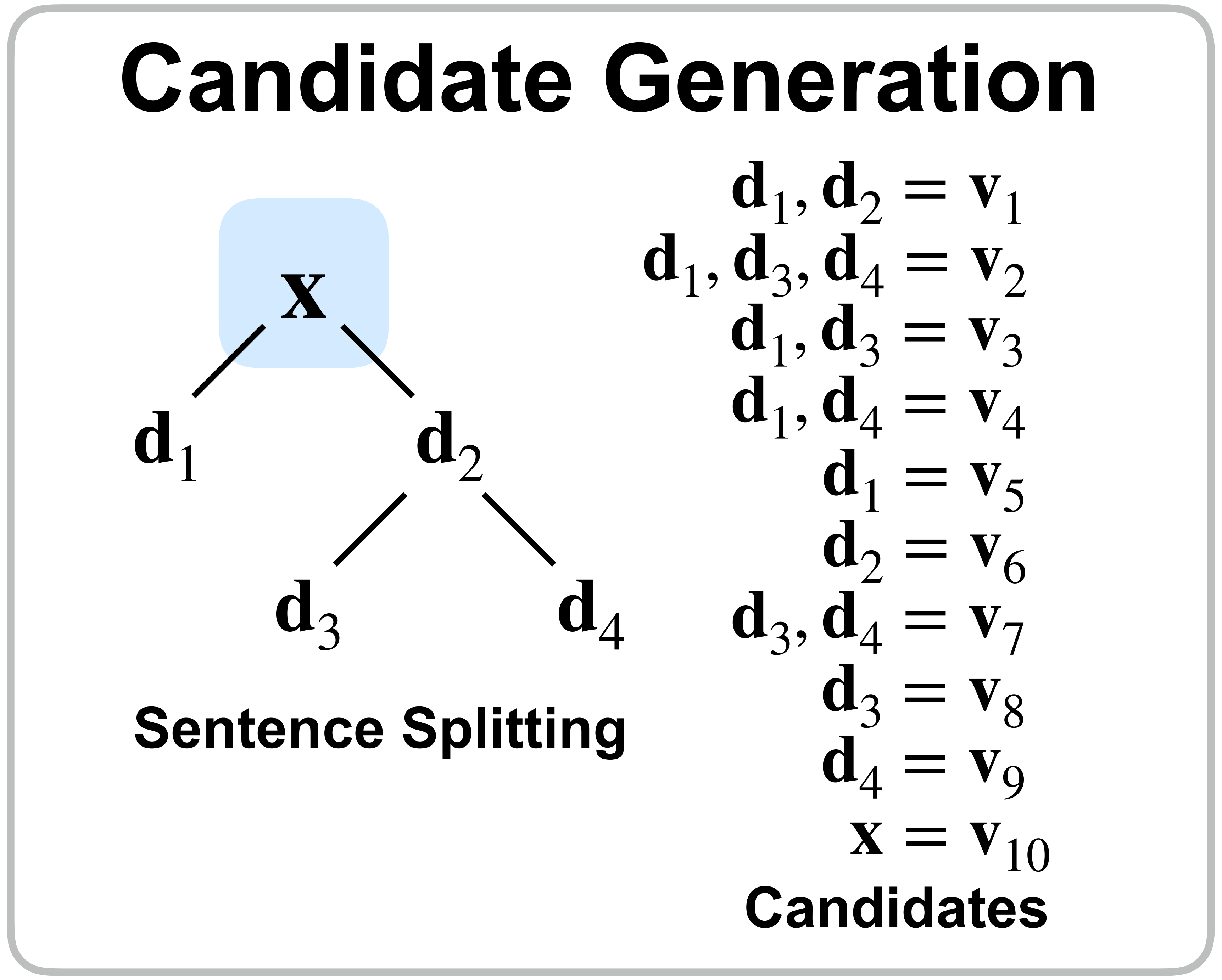
Candidates

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. "Controllable Text Simplification with Explicit Paraphrasing" (NAACL 2021)

Daniel Kim*, Mounica Maddela*, Reno Kriz, Wei Xu, Chris Callison-Burch. "BiSECT: Learning to Split and Rephrase Sentences with Bitexts" (EMNLP 2021)

Step 1 —

We use a rule-based method (Niklaus et al., 2019) + a seq2seq model for splitting and deletion.



Candidate Generation

Sentence Splitting

$$\mathbf{d}_1, \mathbf{d}_2 = \mathbf{v}_1$$
$$\mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_4 = \mathbf{v}_2$$
$$\mathbf{d}_1, \mathbf{d}_3 = \mathbf{v}_3$$
$$\mathbf{d}_1, \mathbf{d}_4 = \mathbf{v}_4$$
$$\mathbf{d}_1 = \mathbf{v}_5$$
$$\mathbf{d}_2 = \mathbf{v}_6$$
$$\mathbf{d}_3, \mathbf{d}_4 = \mathbf{v}_7$$
$$\mathbf{d}_3 = \mathbf{v}_8$$
$$\mathbf{d}_4 = \mathbf{v}_9$$
$$\mathbf{x} = \mathbf{v}_{10}$$

Candidates

Input sentece:

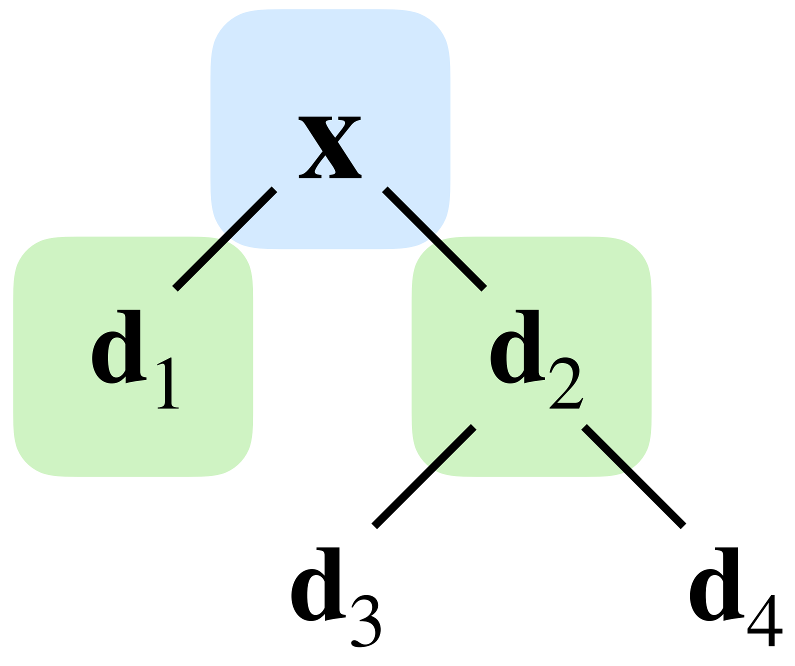
The exhibition, which opened Oct. 8 and runs through Jan. 3, features 27 self-portraits.

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. "Controllable Text Simplification with Explicit Paraphrasing" (NAACL 2021)

Step 1 —

We use a rule-based method (Niklaus et al., 2019) + a seq2seq model for splitting and deletion.

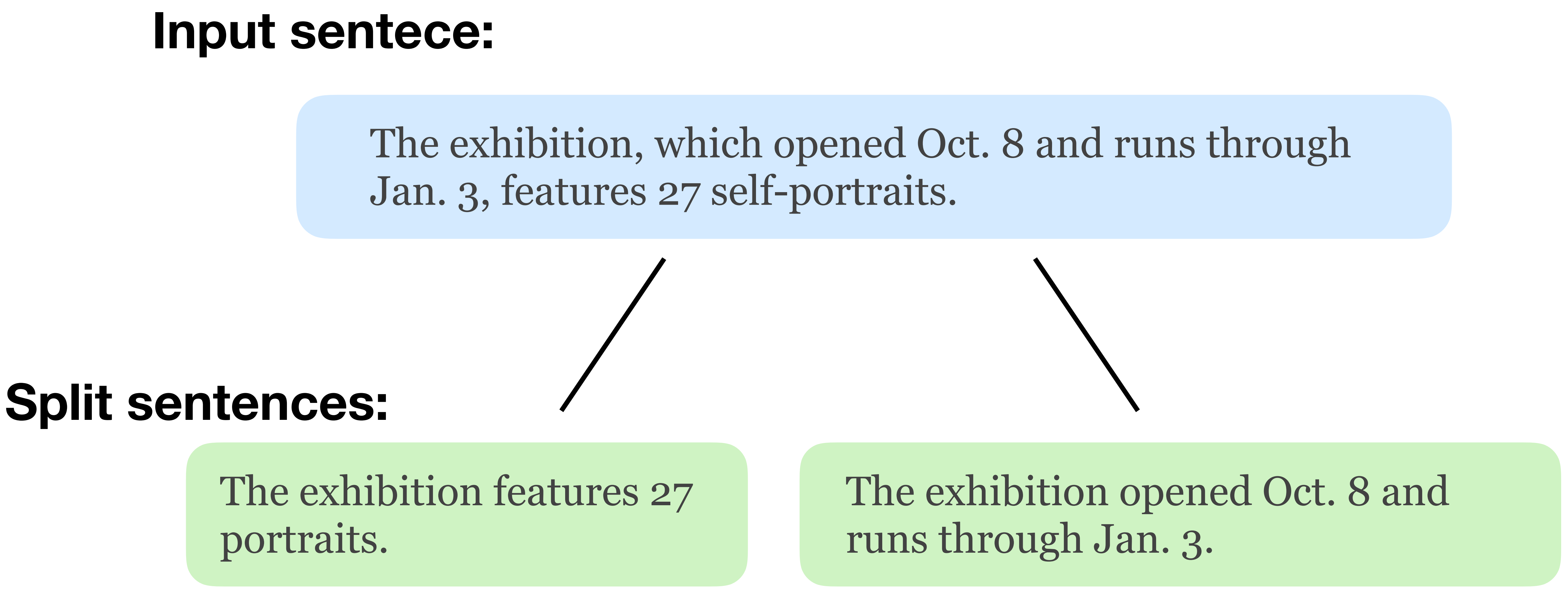
Candidate Generation



Sentence Splitting

$$\begin{aligned}
\mathbf{d}_1, \mathbf{d}_2 &= \mathbf{v}_1 \\
\mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_4 &= \mathbf{v}_2 \\
\mathbf{d}_1, \mathbf{d}_3 &= \mathbf{v}_3 \\
\mathbf{d}_1, \mathbf{d}_4 &= \mathbf{v}_4 \\
\mathbf{d}_1 &= \mathbf{v}_5 \\
\mathbf{d}_2 &= \mathbf{v}_6 \\
\mathbf{d}_3, \mathbf{d}_4 &= \mathbf{v}_7 \\
\mathbf{d}_3 &= \mathbf{v}_8 \\
\mathbf{d}_4 &= \mathbf{v}_9 \\
\mathbf{x} &= \mathbf{v}_{10}
\end{aligned}$$

Candidates



Input sentece:

The exhibition, which opened Oct. 8 and runs through Jan. 3, features 27 self-portraits.

Split sentences:

The exhibition features 27 portraits.

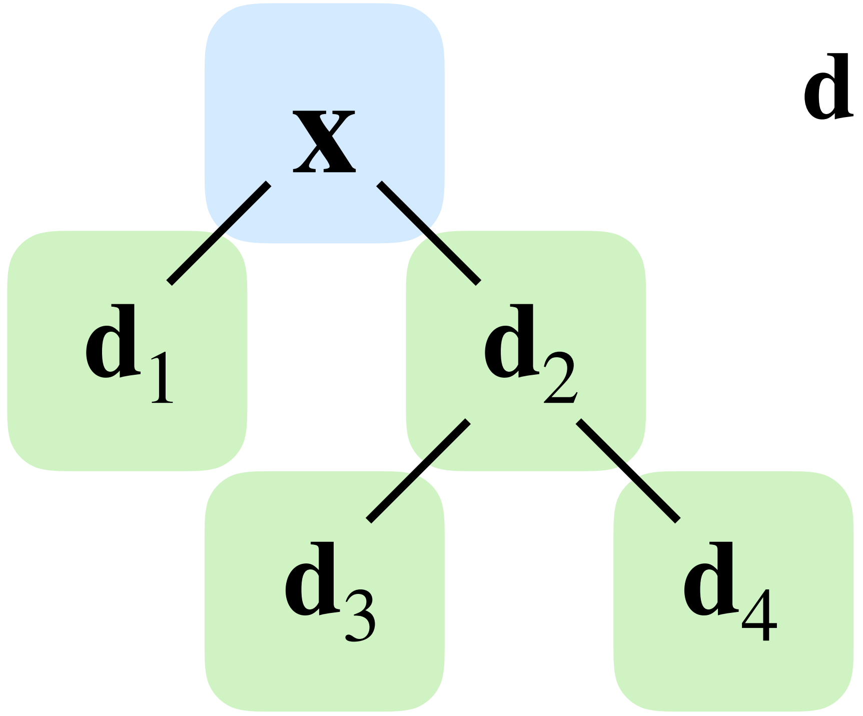
The exhibition opened Oct. 8 and runs through Jan. 3.

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. “Controllable Text Simplification with Explicit Paraphrasing” (NAACL 2021)

Step 1 —

We use a rule-based method (Niklaus et al., 2019) + a seq2seq model for splitting and deletion.

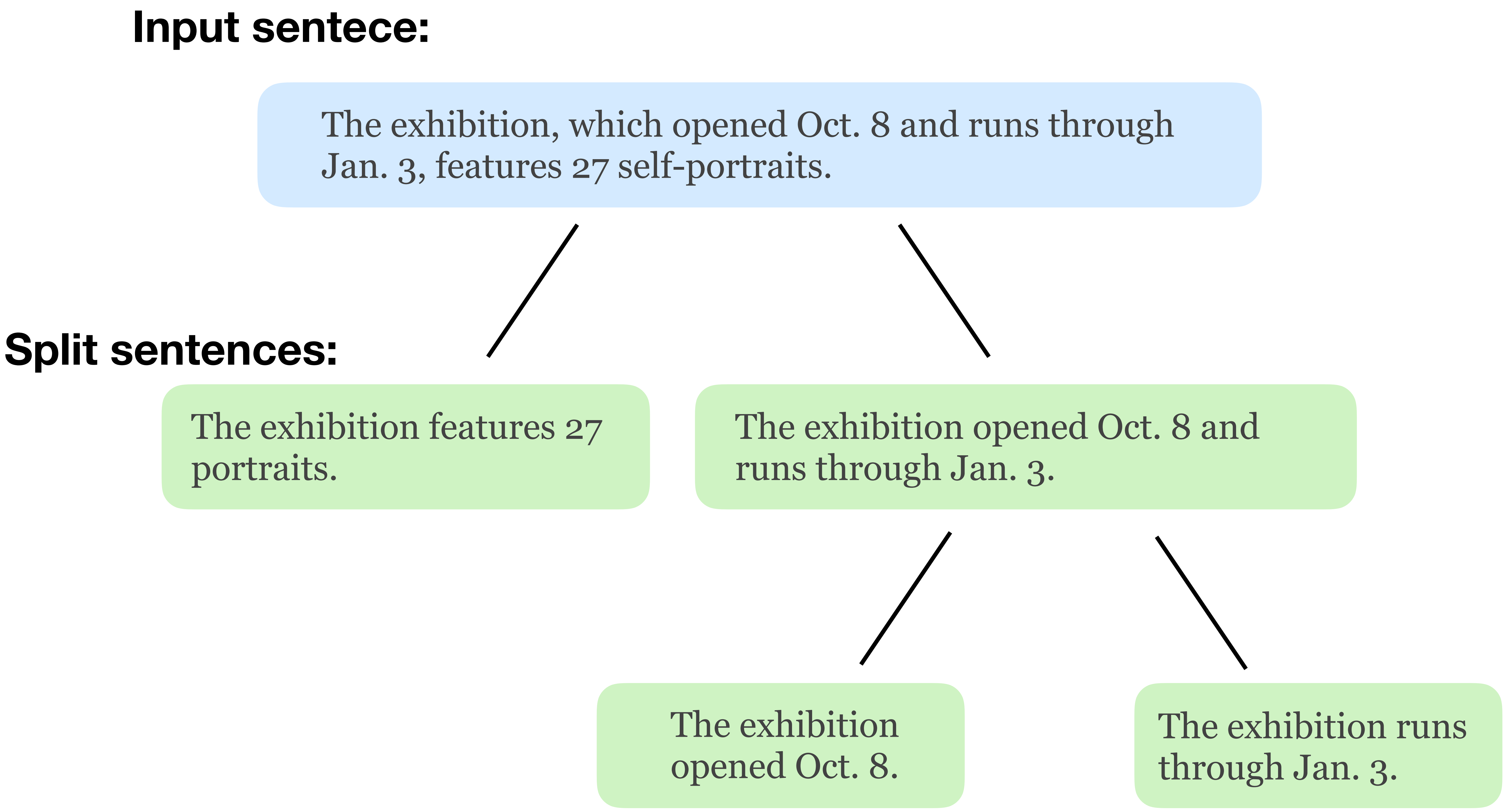
Candidate Generation



Sentence Splitting

$$\begin{aligned}
\mathbf{d}_1, \mathbf{d}_2 &= \mathbf{v}_1\\
\mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_4 &= \mathbf{v}_2\\
\mathbf{d}_1, \mathbf{d}_3 &= \mathbf{v}_3\\
\mathbf{d}_1, \mathbf{d}_4 &= \mathbf{v}_4\\
\mathbf{d}_1 &= \mathbf{v}_5\\
\mathbf{d}_2 &= \mathbf{v}_6\\
\mathbf{d}_3, \mathbf{d}_4 &= \mathbf{v}_7\\
\mathbf{d}_3 &= \mathbf{v}_8\\
\mathbf{d}_4 &= \mathbf{v}_9\\
\mathbf{x} &= \mathbf{v}_{10}
\end{aligned}$$

Candidates



Input sentece:

The exhibition, which opened Oct. 8 and runs through Jan. 3, features 27 self-portraits.

Split sentences:

The exhibition features 27 portraits.

The exhibition opened Oct. 8 and runs through Jan. 3.

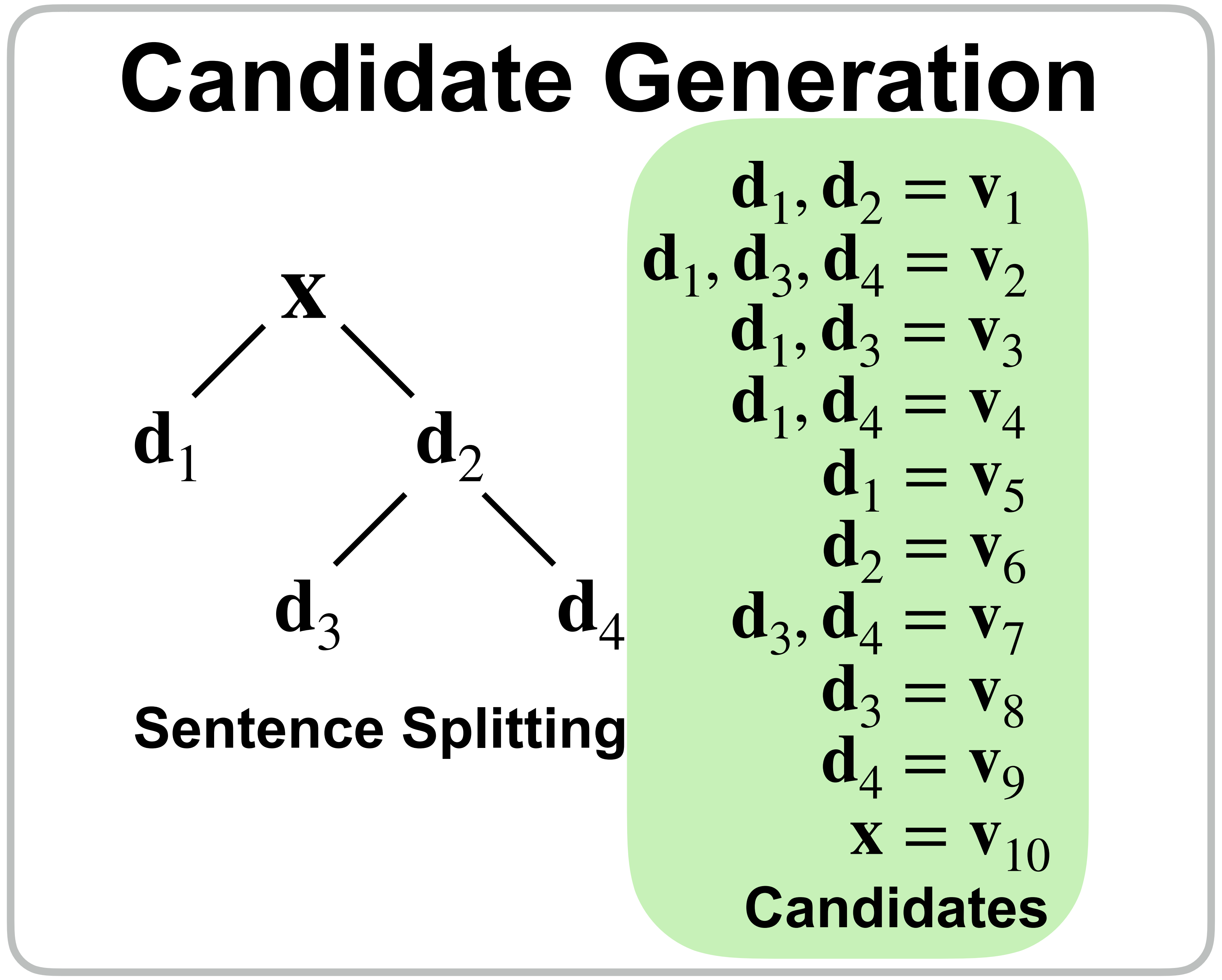
The exhibition opened Oct. 8.

The exhibition runs through Jan. 3.

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. “Controllable Text Simplification with Explicit Paraphrasing” (NAACL 2021)

Step 1 —

We use a rule-based method (Niklaus et al., 2019) + a seq2seq model for splitting and deletion.



Candidate Generation

$\mathbf{x}$, $\mathbf{d}_1$, $\mathbf{d}_2$, $\mathbf{d}_3$, $\mathbf{d}_4$

Sentence Splitting

$$\mathbf{d}_1, \mathbf{d}_2 = \mathbf{v}_1$$
$$\mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_4 = \mathbf{v}_2$$
$$\mathbf{d}_1, \mathbf{d}_3 = \mathbf{v}_3$$
$$\mathbf{d}_1, \mathbf{d}_4 = \mathbf{v}_4$$
$$\mathbf{d}_1 = \mathbf{v}_5$$
$$\mathbf{d}_2 = \mathbf{v}_6$$
$$\mathbf{d}_3, \mathbf{d}_4 = \mathbf{v}_7$$
$$\mathbf{d}_3 = \mathbf{v}_8$$
$$\mathbf{d}_4 = \mathbf{v}_9$$
$$\mathbf{x} = \mathbf{v}_{10}$$

Candidates

Candidates:

The exhibition features 27 portraits. The exhibition opened Oct. 8 and runs through Jan. 3.

The exhibition opened Oct. 8 and runs through Jan. 3.

The exhibition features 27 portraits.

The exhibition opened Oct. 8. The exhibition runs through Jan. 3.

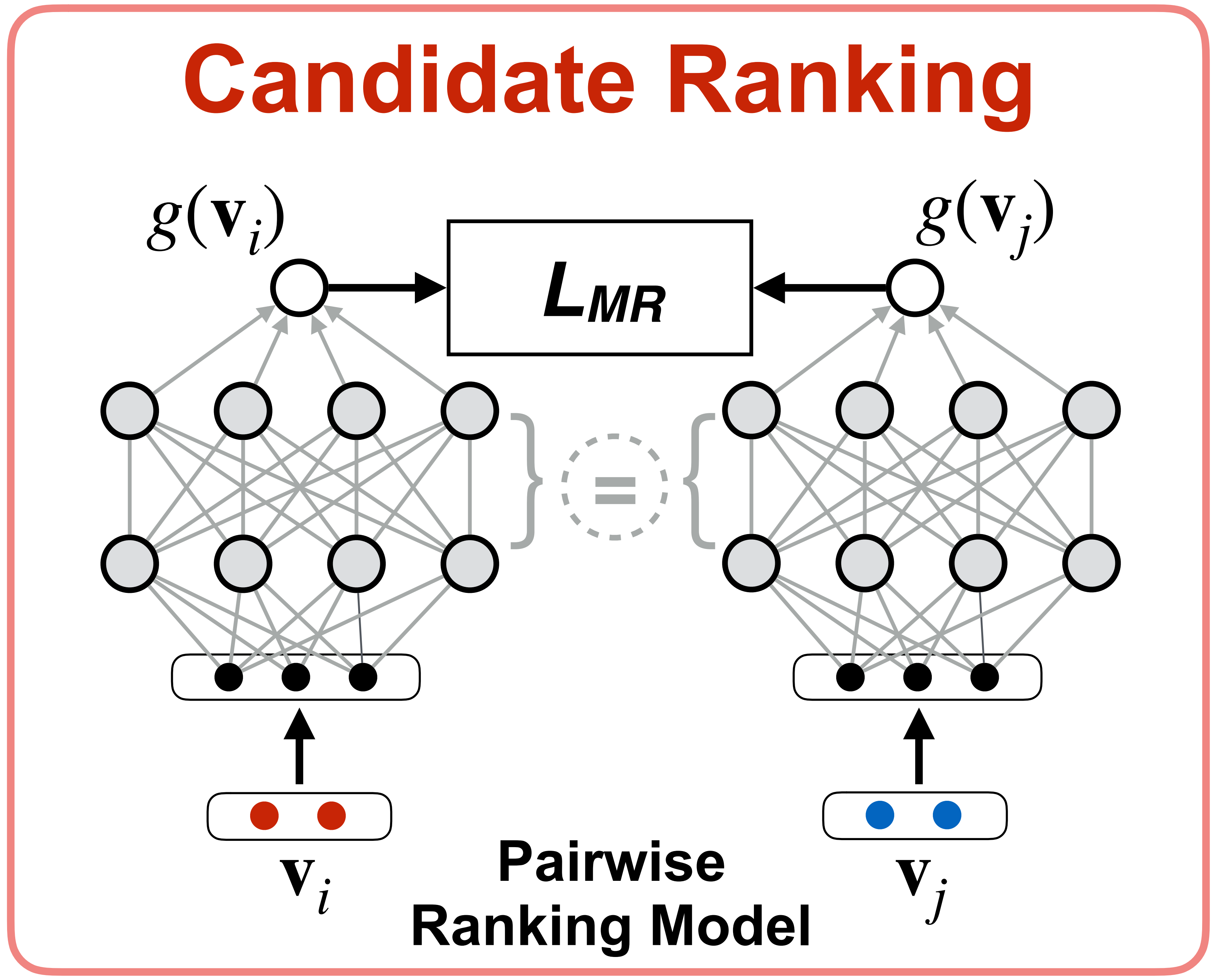
The exhibition features 27 portraits. The exhibition opened Oct. 8.

… (and more)

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. “Controllable Text Simplification with Explicit Paraphrasing” (NAACL 2021)

Step 2 —

Then, we rank all the intermediate outputs (after splitting & deletion).



Candidates:

The exhibition features 27 portraits. The exhibition opened Oct. 8 and runs through Jan. 3.

The exhibition opened Oct. 8 and runs through Jan. 3.

The exhibition features 27 portraits.

The exhibition opened Oct. 8. The exhibition runs through Jan. 3.

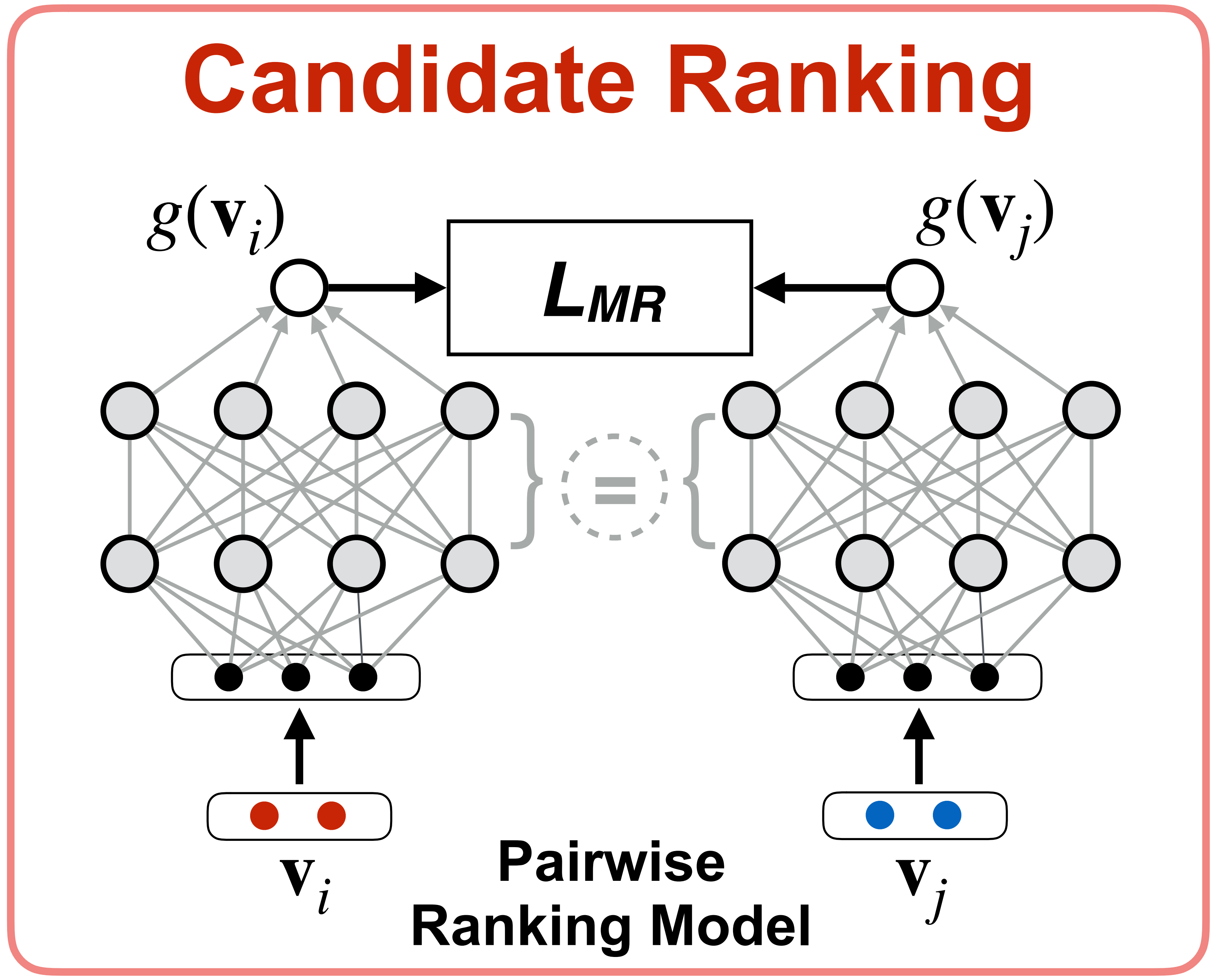
The exhibition features 27 portraits. The exhibition opened Oct. 8.

... (and more)

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. "Controllable Text Simplification with Explicit Paraphrasing" (NAACL 2021)

Step 2 —

Then, we rank all the intermediate outputs (after splitting & deletion).



Candidates:

The exhibition opened Oct. 8. The exhibition runs through Jan. 3.

The exhibition opened Oct. 8 and runs through Jan. 3.

The exhibition features 27 portraits. The exhibition opened Oct. 8.

The exhibition features 27 portraits. The exhibition opened Oct. 8 and runs through Jan. 3.

The exhibition features 27 portraits.

... (and more)

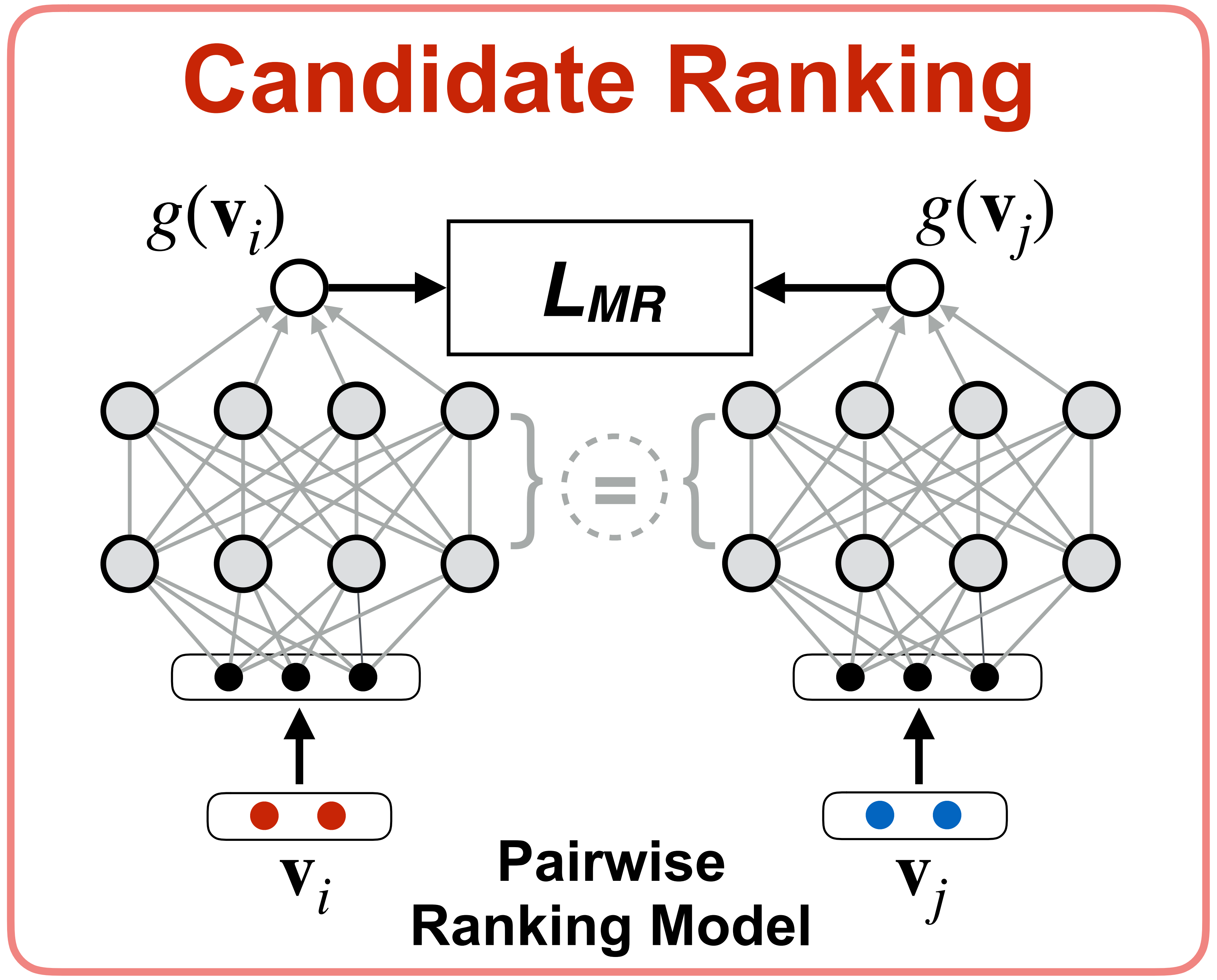
Human reference:

The show started Oct. 8. It ends. Jan 3.

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. “Controllable Text Simplification with Explicit Paraphrasing” (NAACL 2021)

Step 2 —

During training, we access each candidate using BERTScore (Zhang et al. 2019) with length penalty.



Scoring function:

target compression ratio

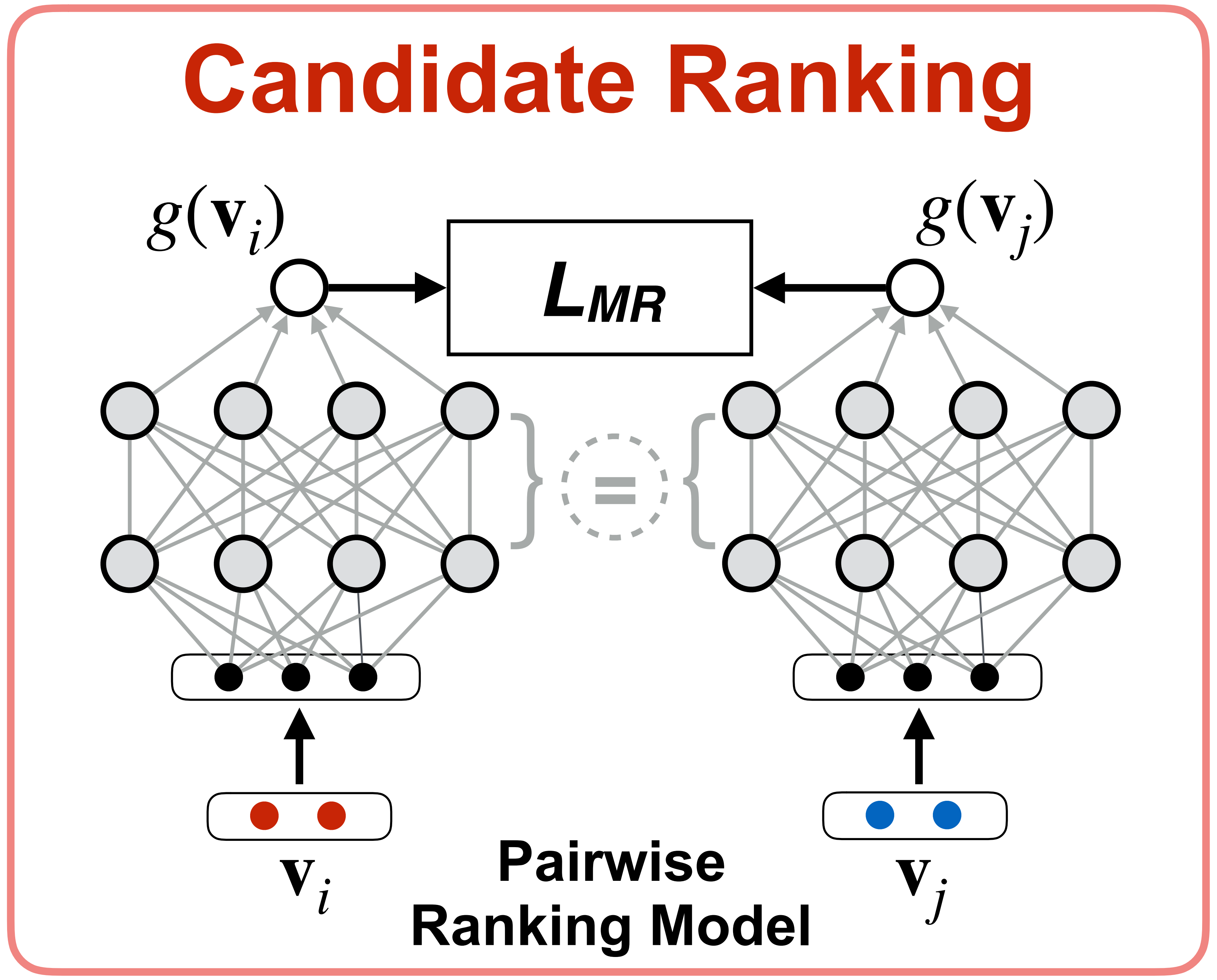
$$g^*(\mathbf{v}_i, \mathbf{y}) = e^{-\lambda||\phi_{\mathbf{v}_i} - \phi_{\mathbf{y}}||} \times BERTScore(\mathbf{v}_i, \mathbf{y})$$

candidate reference

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. “Controllable Text Simplification with Explicit Paraphrasing” (NAACL 2021)

Step 2 —

During training, we access each candidate using BERTScore (Zhang et al. 2019) with length penalty.



Loss function:

$$L_{MR} = \frac{1}{m}\sum_{k=1}^{m}\frac{1}{n_k^2}\sum_{i=1}^{n_k}\sum_{j=1,i\neq j}^{n_k}\max(0, 1 - l_{ij}^k d_{ij}^k)$$

$$d_{ij}^k = g(\mathbf{v}_i^k) - g(\mathbf{v}_j^k)$$

$$l_{ij}^k = sign\left(g^*(\mathbf{v}_i^k, \mathbf{y}^k) - g^*(\mathbf{v}_j^k, \mathbf{y}^k)\right)$$

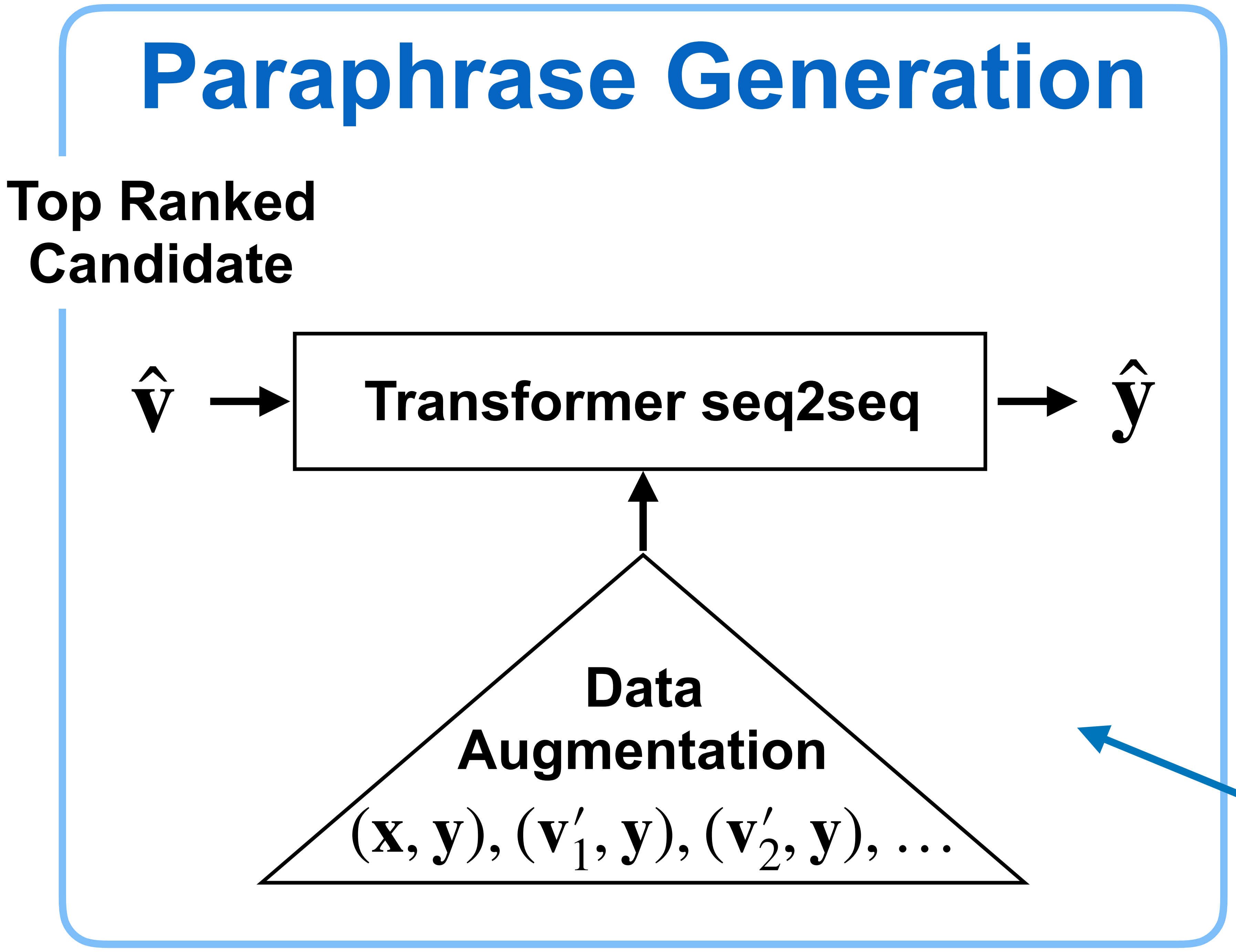
Length-penalized BERTScore

Features: number of words in v_i and x, compression ratio of v_i with respect to x, Jaccard similarity between v_i and x, the rules applied on x to obtain v_i, and the number of rule applications.

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. “Controllable Text Simplification with Explicit Paraphrasing” (NAACL 2021)

Step 3 —

Finally, we have a paraphrase generation model trained with augmented training data.
(some selected candidates, in addition to the original input, are paired with the human reference)

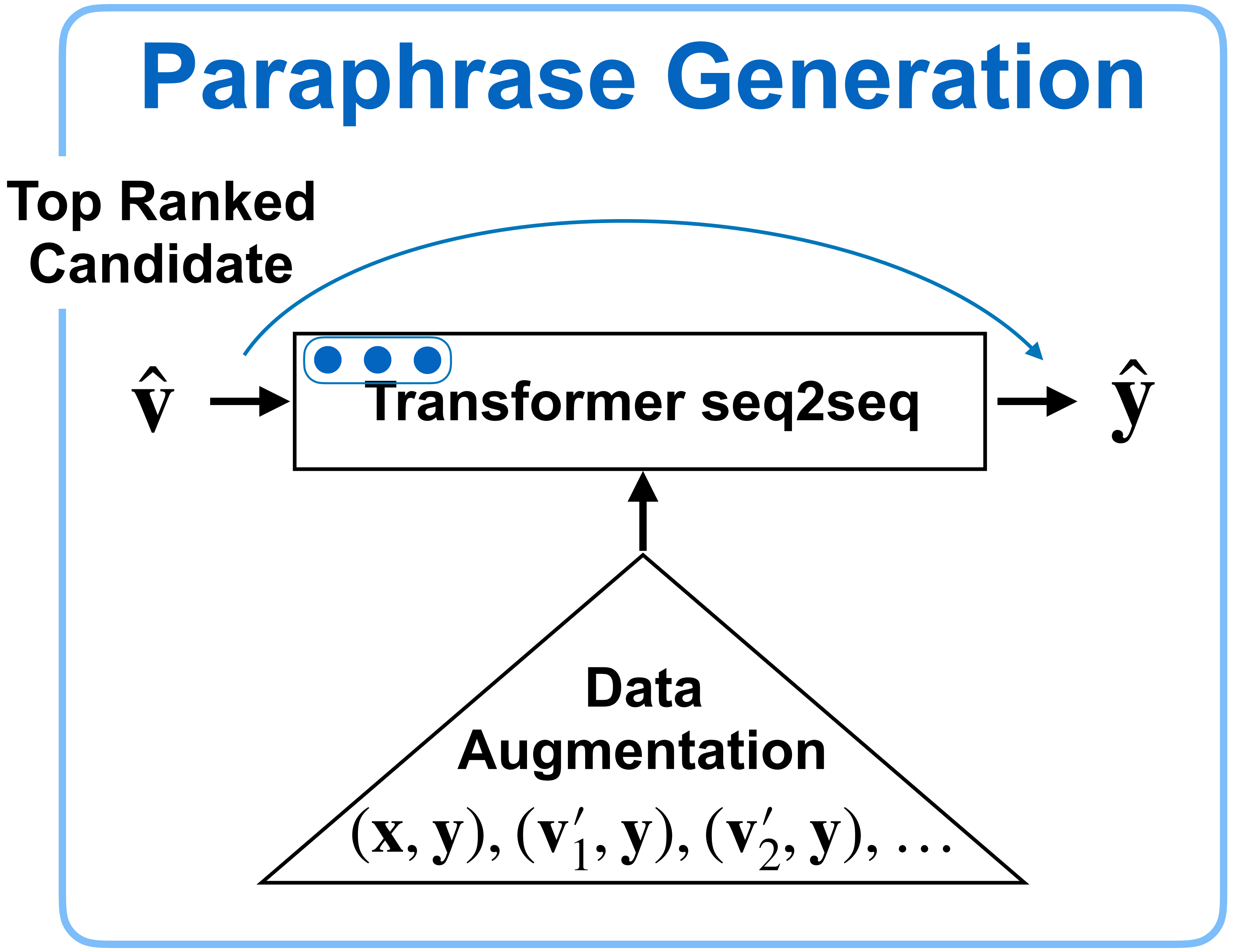


Training a specific generation model that focuses on generating more diverse paraphrases.

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. “Controllable Text Simplification with Explicit Paraphrasing” (NAACL 2021)

Step 3 —

Finally, we have a paraphrase generation model trained with augmented training data.
(some selected candidates, in addition to the original input, are paired with the human reference)



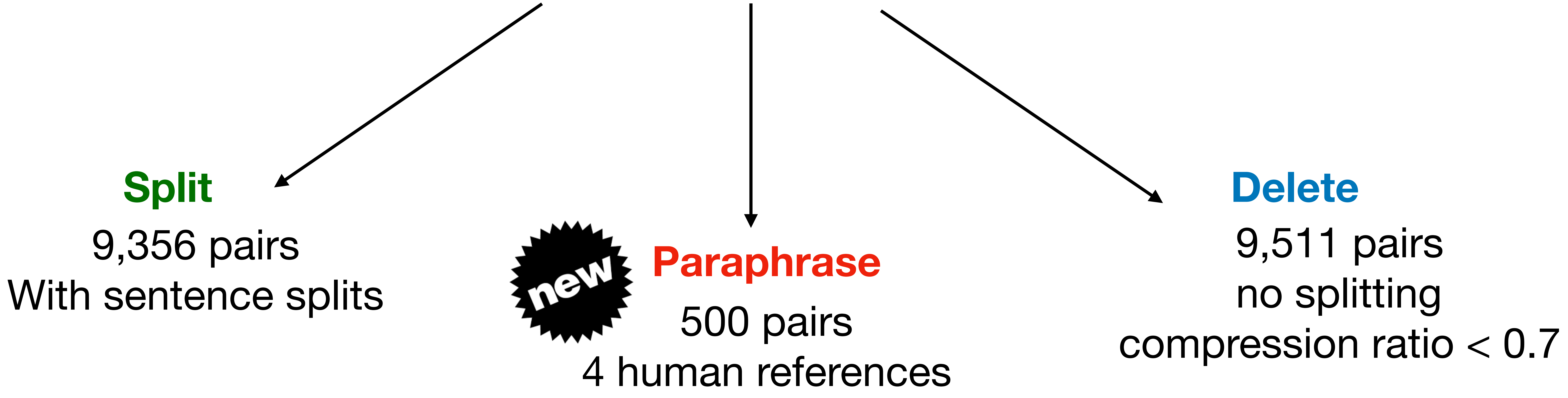
Additional control over the degree of paraphrasing:

- A copy-control token as soft constraint.
- An auxiliary task (whether a word should be copied) using a monolingual word aligner to derive noisy training labels.

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. "Controllable Text Simplification with Explicit Paraphrasing" (NAACL 2021)

Experiments on Text Simplification

- Evaluation setup
  - Standard Evaluation on **Newsela-Auto** and **Wikipedia-Auto** (Jiang et al. 2020).
  - Edit-focused Evaluation on different sections of test set (Our work).



Split

9,356 pairs

With sentence splits

Paraphrase

500 pairs

4 human references

Delete

9,511 pairs

no splitting

compression ratio < 0.7

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. "Controllable Text Simplification with Explicit Paraphrasing" (NAACL 2021)

Controllable Text Generation

We can control the degree of sentence splitting, deletion, and paraphrasing.

Input: Experts say China's air pollution exacts a tremendous toll on human health.

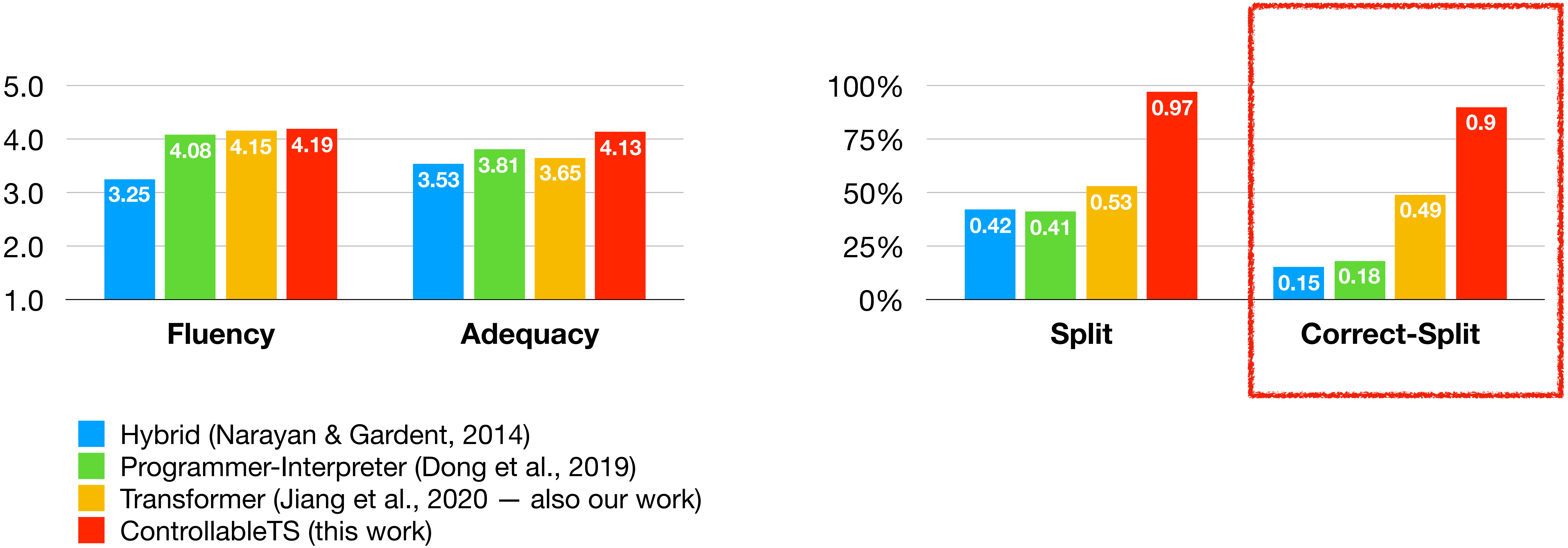
Reference: China's air pollution is very unhealthy.

Our Model (*cp = 0.6*)	experts say china's air pollution **is a big problem for** human health.
Our Model (*cp = 0.7*)	experts say china's air pollution **can cause a lot of damage on** human health.
Our Model (*cp = 0.8*)	experts say china's air pollution **is** a **huge** toll on human health.
Hybrid-NG	experts say **government's** air pollution exacts a tremendous toll on human health.
LSTM	experts say china's air pollution exacts a tremendous toll on human health.
Transformer	experts say china's air pollution exacts a tremendous **effect** on human health.
EditNTS	experts say china's air pollution **can cause** human health.

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. "Controllable Text Simplification with Explicit Paraphrasing" (NAACL 2021)

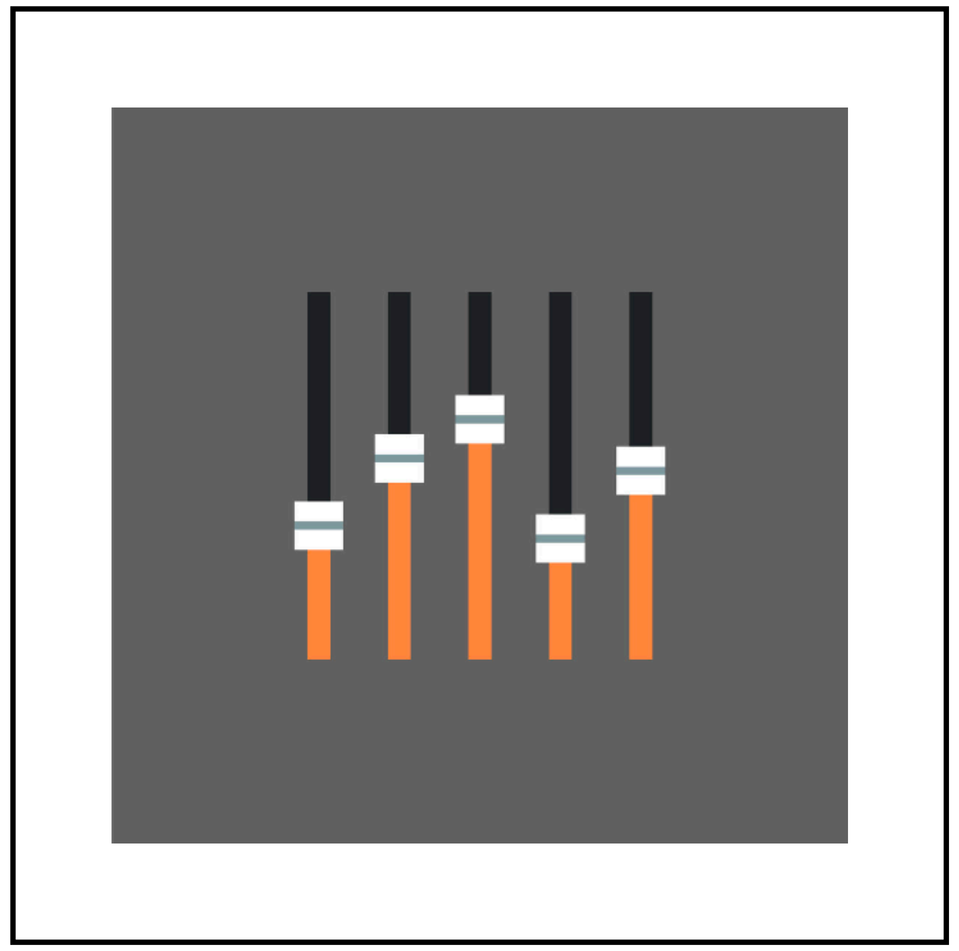
More Syntactic Transformations

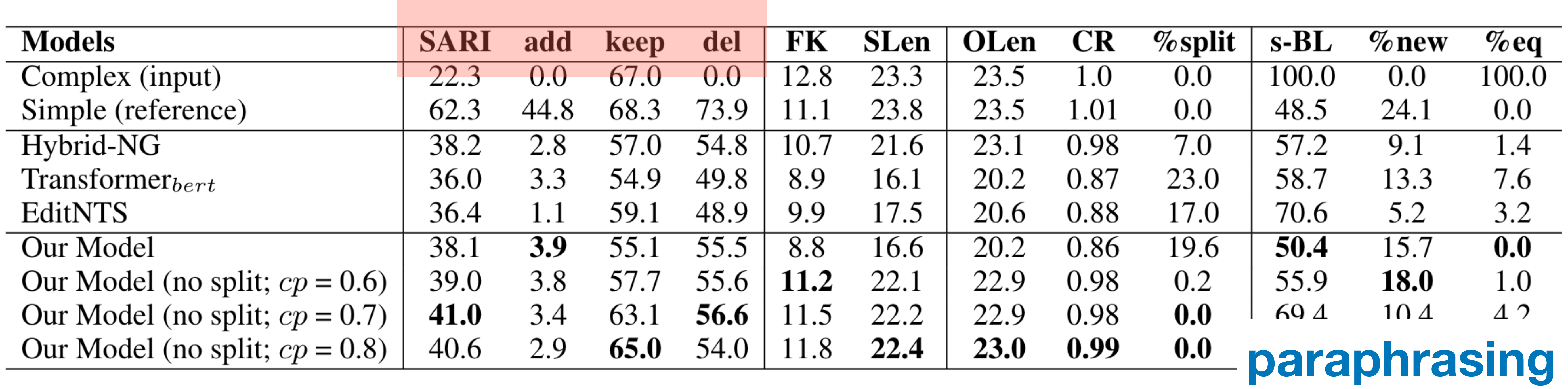
Human evaluation (1-5 Likert scale) on sentences where simplification involves splitting.



Mounica Maddela, Fernando Alva-Manchego, Wei Xu. "Controllable Text Simplification with Explicit Paraphrasing" (NAACL 2021)

Controllable Generation & Evaluation

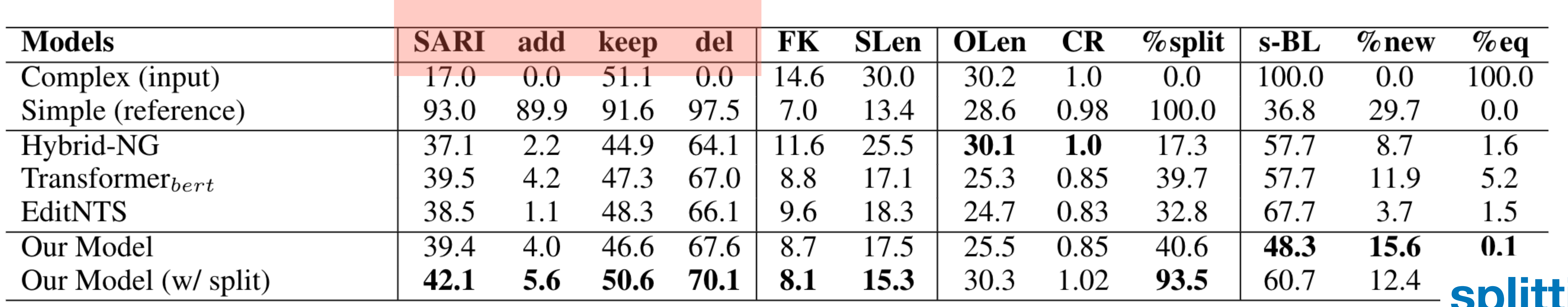




Models	SARI	add	keep	del	FK	SLen	OLen	CR	%split	s-BL	%new	%eq
Complex (input)	22.3	0.0	67.0	0.0	12.8	23.3	23.5	1.0	0.0	100.0	0.0	100.0
Simple (reference)	62.3	44.8	68.3	73.9	11.1	23.8	23.5	1.01	0.0	48.5	24.1	0.0
Hybrid-NG	38.2	2.8	57.0	54.8	10.7	21.6	23.1	0.98	7.0	57.2	9.1	1.4
Transformer$_{bert}$	36.0	3.3	54.9	49.8	8.9	16.1	20.2	0.87	23.0	58.7	13.3	7.6
EditNTS	36.4	1.1	59.1	48.9	9.9	17.5	20.6	0.88	17.0	70.6	5.2	3.2
Our Model	38.1	**3.9**	55.1	55.5	8.8	16.6	20.2	0.86	19.6	**50.4**	15.7	**0.0**
Our Model (no split; *cp* = 0.6)	39.0	3.8	57.7	55.6	**11.2**	22.1	22.9	0.98	0.2	55.9	**18.0**	1.0
Our Model (no split; *cp* = 0.7)	**41.0**	3.4	63.1	**56.6**	11.5	22.2	22.9	0.98	**0.0**	69.4	10.4	4.2
Our Model (no split; *cp* = 0.8)	40.6	2.9	**65.0**	54.0	11.8	**22.4**	**23.0**	**0.99**	**0.0**			

paraphrasing

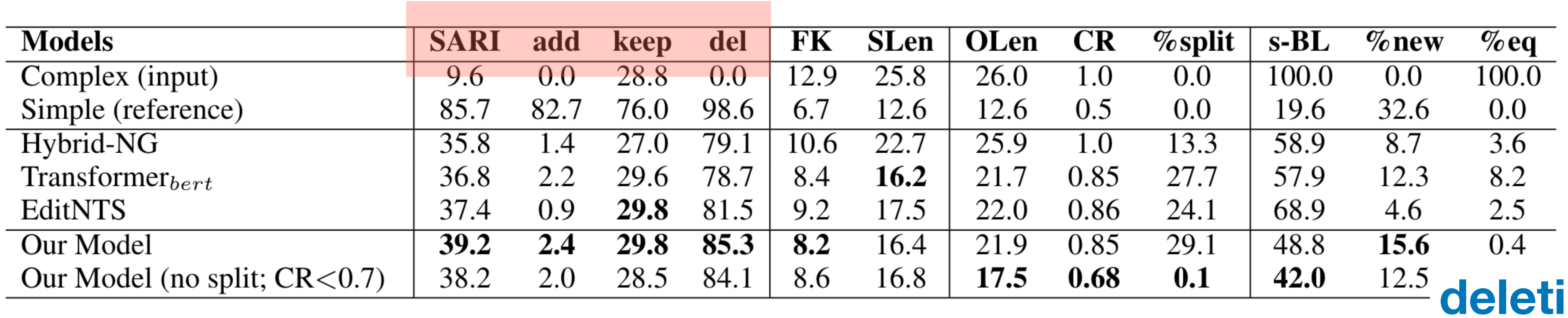
Table 2: Automatic evaluation results on NEWSELA-TURK that focuses on paraphrasing (500 complex sentences with 4 human written paraphrases). We control the extent of paraphrasing of our models by specifying the percentage of words to be copied (*cp*) from the input as a soft constraint.



Models	SARI	add	keep	del	FK	SLen	OLen	CR	%split	s-BL	%new	%eq
Complex (input)	17.0	0.0	51.1	0.0	14.6	30.0	30.2	1.0	0.0	100.0	0.0	100.0
Simple (reference)	93.0	89.9	91.6	97.5	7.0	13.4	28.6	0.98	100.0	36.8	29.7	0.0
Hybrid-NG	37.1	2.2	44.9	64.1	11.6	25.5	**30.1**	**1.0**	17.3	57.7	8.7	1.6
Transformer$_{bert}$	39.5	4.2	47.3	67.0	8.8	17.1	25.3	0.85	39.7	57.7	11.9	5.2
EditNTS	38.5	1.1	48.3	66.1	9.6	18.3	24.7	0.83	32.8	67.7	3.7	1.5
Our Model	39.4	4.0	46.6	67.6	8.7	17.5	25.5	0.85	40.6	**48.3**	**15.6**	**0.1**
Our Model (w/ split)	**42.1**	**5.6**	**50.6**	**70.1**	**8.1**	**15.3**	30.3	1.02	**93.5**	60.7	12.4	

splitting

Table 3: Automatic evaluation results on a subset of NEWSELA-AUTO test set that focuses on splitting (9,356 complex-simple sentence pairs with splitting). Our model chooses only candidate simplifications that have undergone splitting during the ranking step of the pipeline.



Models	SARI	add	keep	del	FK	SLen	OLen	CR	%split	s-BL	%new	%eq
Complex (input)	9.6	0.0	28.8	0.0	12.9	25.8	26.0	1.0	0.0	100.0	0.0	100.0
Simple (reference)	85.7	82.7	76.0	98.6	6.7	12.6	12.6	0.5	0.0	19.6	32.6	0.0
Hybrid-NG	35.8	1.4	27.0	79.1	10.6	22.7	25.9	1.0	13.3	58.9	8.7	3.6
Transformer$_{bert}$	36.8	2.2	29.6	78.7	8.4	**16.2**	21.7	0.85	27.7	57.9	12.3	8.2
EditNTS	37.4	0.9	**29.8**	81.5	9.2	17.5	22.0	0.86	24.1	68.9	4.6	2.5
Our Model	**39.2**	**2.4**	**29.8**	**85.3**	**8.2**	16.4	21.9	0.85	29.1	48.8	**15.6**	0.4
Our Model (no split; CR<0.7)	38.2	2.0	28.5	84.1	8.6	16.8	**17.5**	**0.68**	**0.1**	**42.0**	12.5	

deletion

Table 4: Automatic evaluation results on a subset of NEWSELA-AUTO test set that focuses on deletion (9,511 complex-simple sentence pairs with compression ratio < 0.7 and no sentence splits). Our model selects only candidates with similar compression ratio and no splits during ranking.

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. “Controllable Text Simplification with Explicit Paraphrasing” (NAACL 2021)

Part 1.5 — Automatic Evaluation Metric

🥻

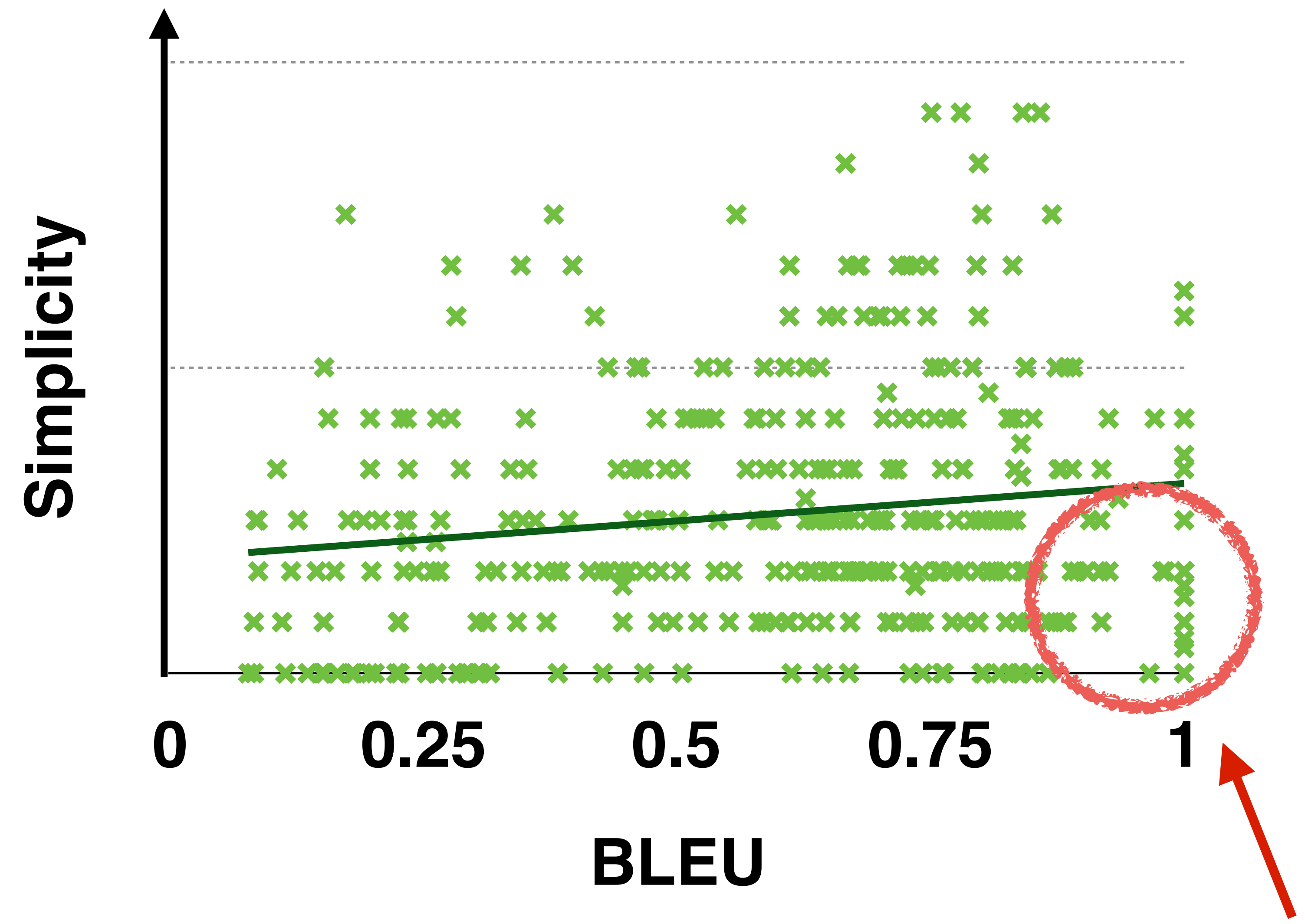
Optimizing Statistical Machine Translation for Simplification
Xu et al. (TACL 2016)

BLEU is not for Simplification

If a text generation model simply output the input unchanged, it gets perfect grammar, perfect meaning preservation, and very high BLEU score.

Human Evaluation (1-5 Likert scale)

- ● **Grammaticality / Fluency**
- ● **Meaning preservation / Adequacy**
- ● **Simplicity**

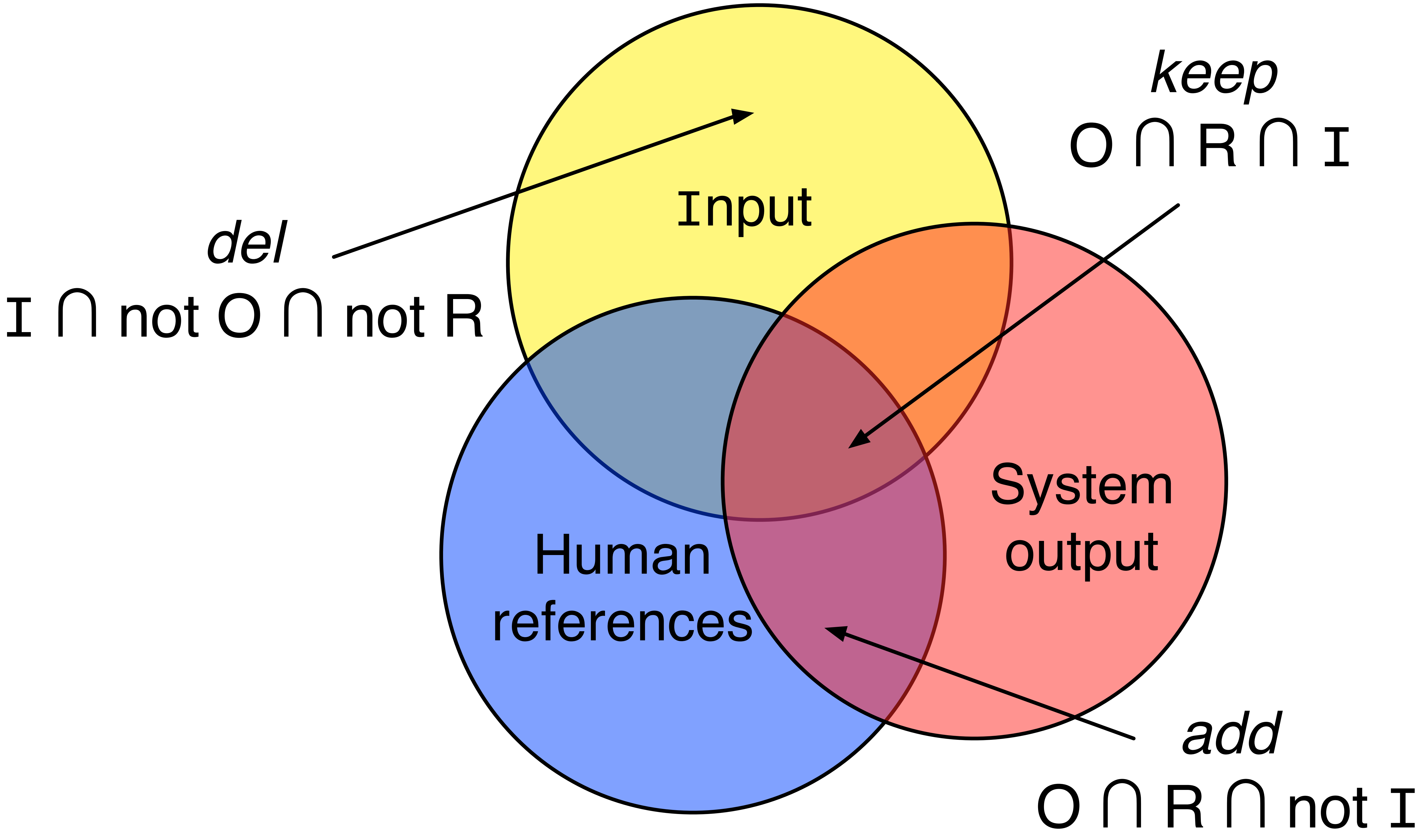


low Simplicity high BLEU

Wei Xu, Courtney Napoles, Ellie Pavlick, Chris Callison-Burch. “Optimizing Statistical Machine Translation for Simplification” (TACL 2016)

SARI Metric

It compares system output against references and against the input sentence.



$$p_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \overline{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(O \cap \overline{I})}$$

$$r_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \overline{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(R \cap \overline{I})}$$

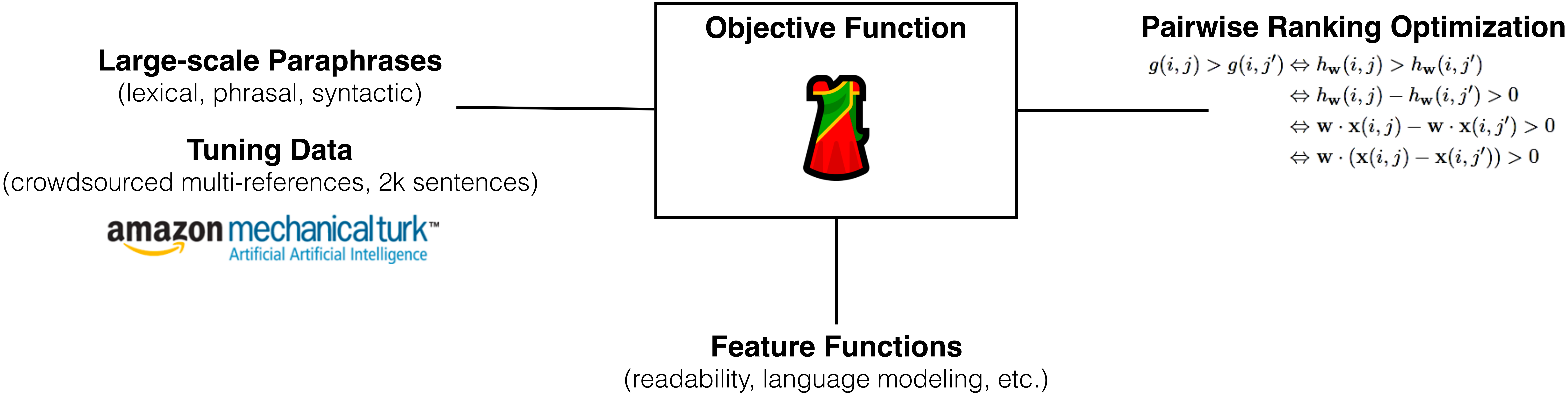
$$\text{SARI} = d_1 F_{add} + d_2 F_{keep} + d_3 P_{del}$$

$$d_1 = d_2 = d_3 = 1/3$$

Wei Xu, Courtney Napoles, Ellie Pavlick, Chris Callison-Burch. “Optimizing Statistical Machine Translation for Simplification” (TACL 2016)

SARI Metric + Turk Corpus

SARI can also be used as (part of) the training objective/reward function.



Objective Function

Large-scale Paraphrases
(lexical, phrasal, syntactic)

Tuning Data
(crowdsourced multi-references, 2k sentences)

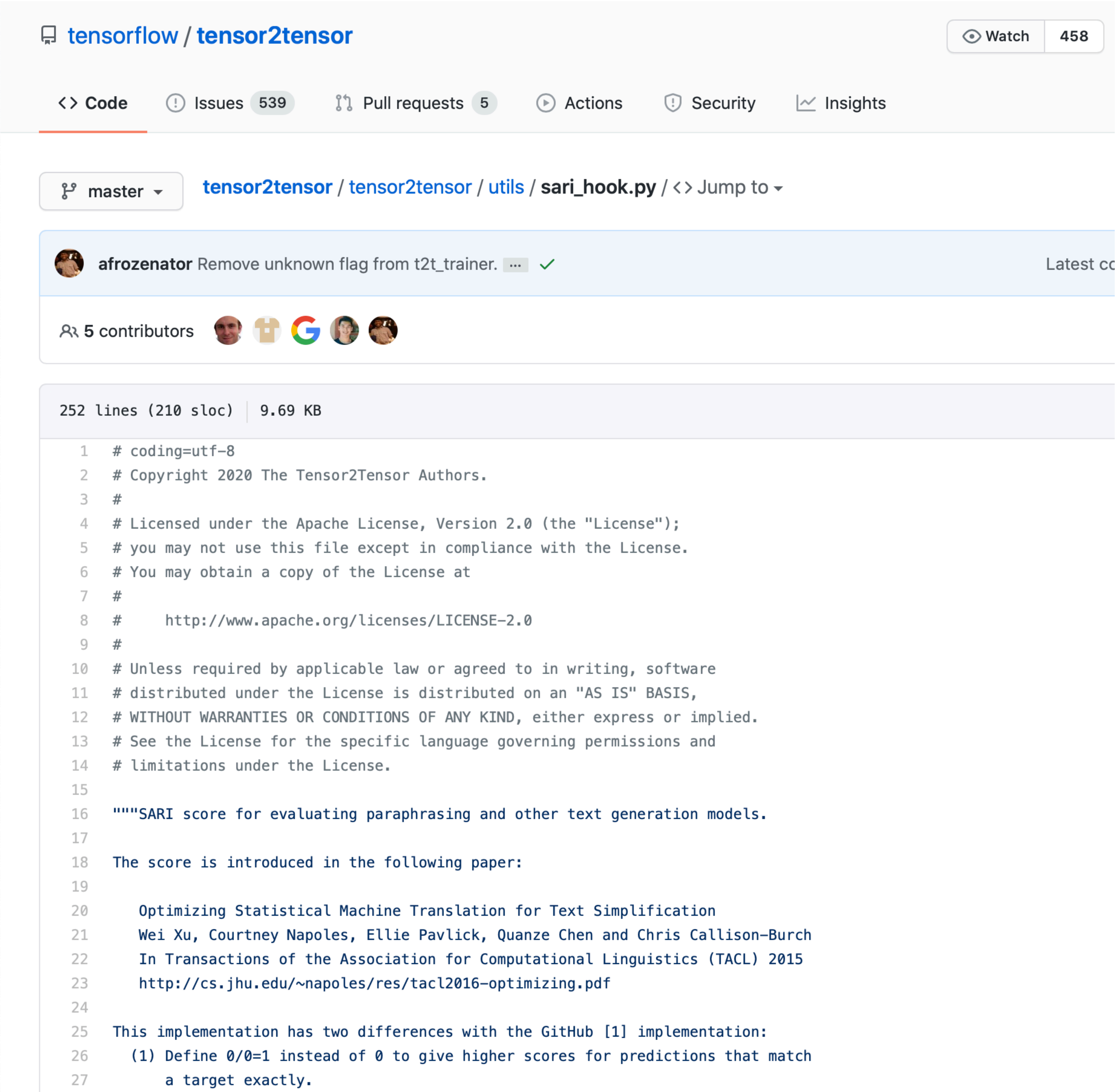
amazon mechanical turk™
Artificial Artificial Intelligence

Pairwise Ranking Optimization

$$\begin{aligned} g(i,j) > g(i,j') &\Leftrightarrow h_{\mathbf{w}}(i,j) > h_{\mathbf{w}}(i,j') \\ &\Leftrightarrow h_{\mathbf{w}}(i,j) - h_{\mathbf{w}}(i,j') > 0 \\ &\Leftrightarrow \mathbf{w} \cdot \mathbf{x}(i,j) - \mathbf{w} \cdot \mathbf{x}(i,j') > 0 \\ &\Leftrightarrow \mathbf{w} \cdot (\mathbf{x}(i,j) - \mathbf{x}(i,j')) > 0 \end{aligned}$$

Feature Functions
(readability, language modeling, etc.)

Wei Xu, Courtney Napoles, Ellie Pavlick, Chris Callison-Burch. “Optimizing Statistical Machine Translation for Simplification” (TACL 2016)


tensorflow / tensor2tensor
Watch 458
Code
Issues 539
Pull requests 5
Actions
Security
Insights
master
tensor2tensor / tensor2tensor / utils / sari_hook.py / Jump to
afrozenator Remove unknown flag from t2t_trainer. …
Latest co
5 contributors

```
252 lines (210 sloc) | 9.69 KB
```

```
# coding=utf-8
# Copyright 2020 The Tensor2Tensor Authors.
#
# Licensed under the Apache License, Version 2.0 (the "License");
# you may not use this file except in compliance with the License.
# You may obtain a copy of the License at
#
#     http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.

"""SARI score for evaluating paraphrasing and other text generation models.

The score is introduced in the following paper:

   Optimizing Statistical Machine Translation for Text Simplification
   Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen and Chris Callison-Burch
   In Transactions of the Association for Computational Linguistics (TACL) 2015
   http://cs.jhu.edu/~napoles/res/tacl2016-optimizing.pdf

This implementation has two differences with the GitHub [1] implementation:
  (1) Define 0/0=1 instead of 0 to give higher scores for predictions that match
      a target exactly.
```

SARI is added to TensorFlow by Google AI group in Feb 2019.

Google

TensorFlow

Now, also in

HUGGING FACE

SARI Metric

It compares **s**ystem output **a**gainst **r**eferences and against the **i**nput sentence.

Beyond text simplification …

“Leveraging Pre-trained Checkpoints for Sequence Generation Tasks”

[Sascha Rothe, Shashi Narayan, Aliaksei Severyn - TACL 2020]

← using SARI for sentence splitting and fusion

“Decontextualization: Making Sentences Stand-Alone”

[Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, Michael Collins - TACL 2021]

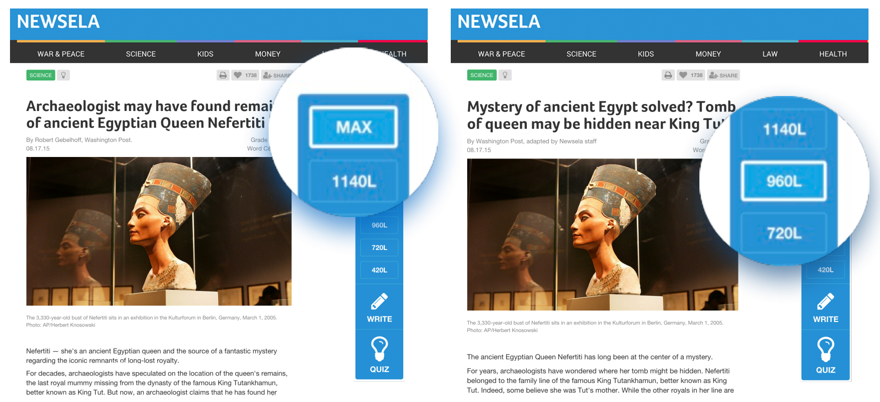
← using SARI for sentence decontextualization: taking a sentence together with its context and rewriting it to be interpretable out of context, while preserving its meaning

“Evidence-based Factual Error Correction”

[James Thorne, Andreas Vlachos - ACL 2021]

← using SARI for revising claims based on facts correlates well with human judgements!

Part 2 — High-quality Training Data



Archaeologist may have found remai
of ancient Egyptian Queen Nefertiti

By Robert Gebelhoff, Washington Post.
08.17.15

The 3,330-year-old bust of Nefertiti sits in an exhibition in the Kulturforum in Berlin, Germany, March 1, 2005.
Photo: AP/Herbert Knosowski

Nefertiti — she's an ancient Egyptian queen and the source of a fantastic mystery regarding the iconic remnants of long-lost royalty.

For decades, archaeologists have speculated on the location of the queen's remains, the last royal mummy missing from the dynasty of the famous King Tutankhamun, better known as King Tut. But now, an archaeologist claims that he has found her

Mystery of ancient Egypt solved? Tomb
of queen may be hidden near King Tu

By Washington Post, adapted by Newsela staff
08.17.15

The 3,330-year-old bust of Nefertiti sits in an exhibition in the Kulturforum in Berlin, Germany, March 1, 2005.
Photo: AP/Herbert Knosowski

The ancient Egyptian Queen Nefertiti has long been at the center of a mystery.

For years, archaeologists have wondered where her tomb might be hidden. Nefertiti belonged to the family line of the famous King Tutankhamun, better known as King Tut. Indeed, some believe she was Tut's mother. While the other royals in her line are

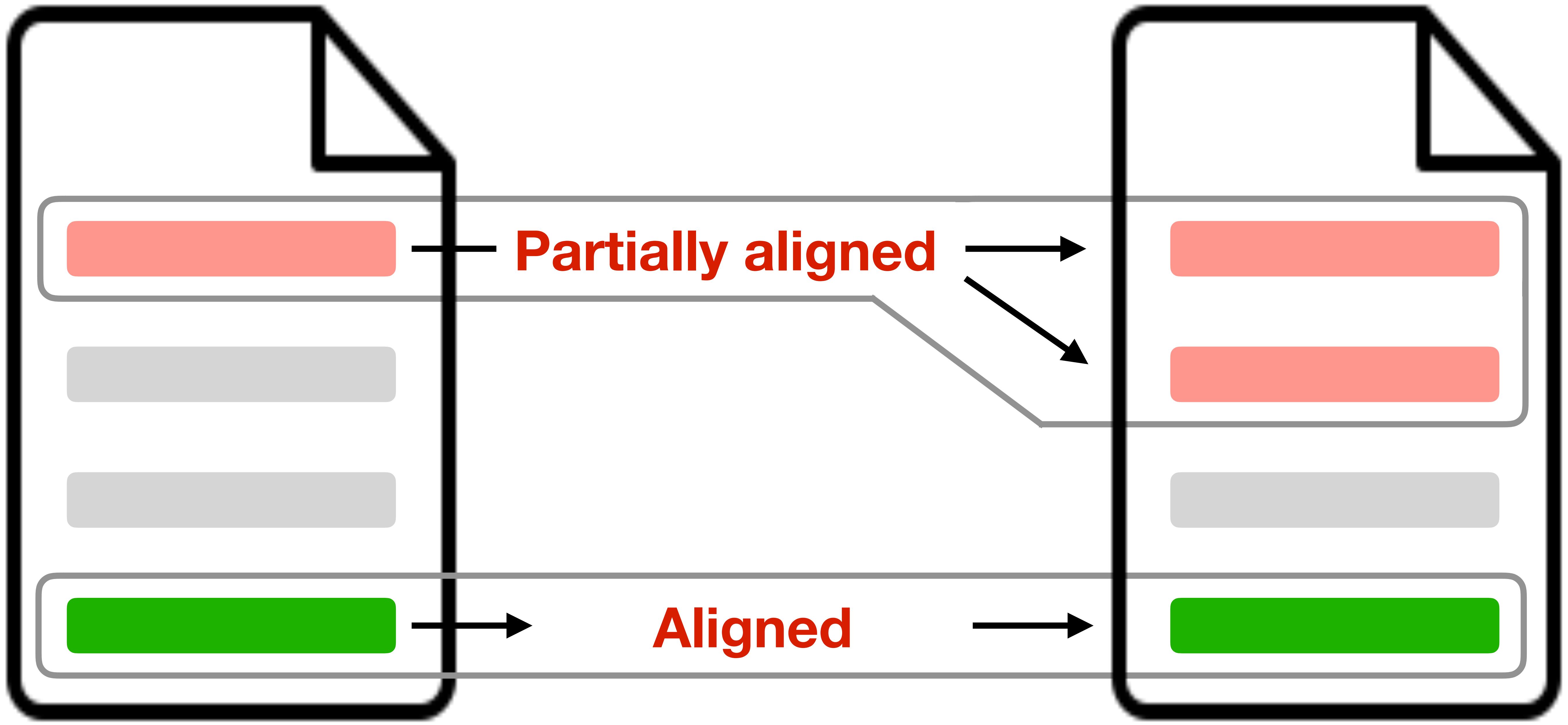
Simple English
WIKIPEDIA

Neural CRF Model for Sentence Alignment in Text Simplification

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, Wei Xu (ACL 2020)

Automatic Text Simplification

- Primarily addressed by sequence-to-sequence models.
- **Training corpus** are complex-simple sentence pairs extracted by **aligning parallel articles**.



Partially aligned

Aligned

Complex article

Simple article

WIKIPEDIA
The Free Encyclopedia

newsela
(Original article)

Simple English
WIKIPEDIA

newsela
(Simplified article)

Our Solution for Sentence Alignment

- Two high-quality manually annotated sentence alignment datasets (20k / 10k sentence pairs).
- Structure prediction + BERT$_{finetune}$ → A neural CRF alignment model.

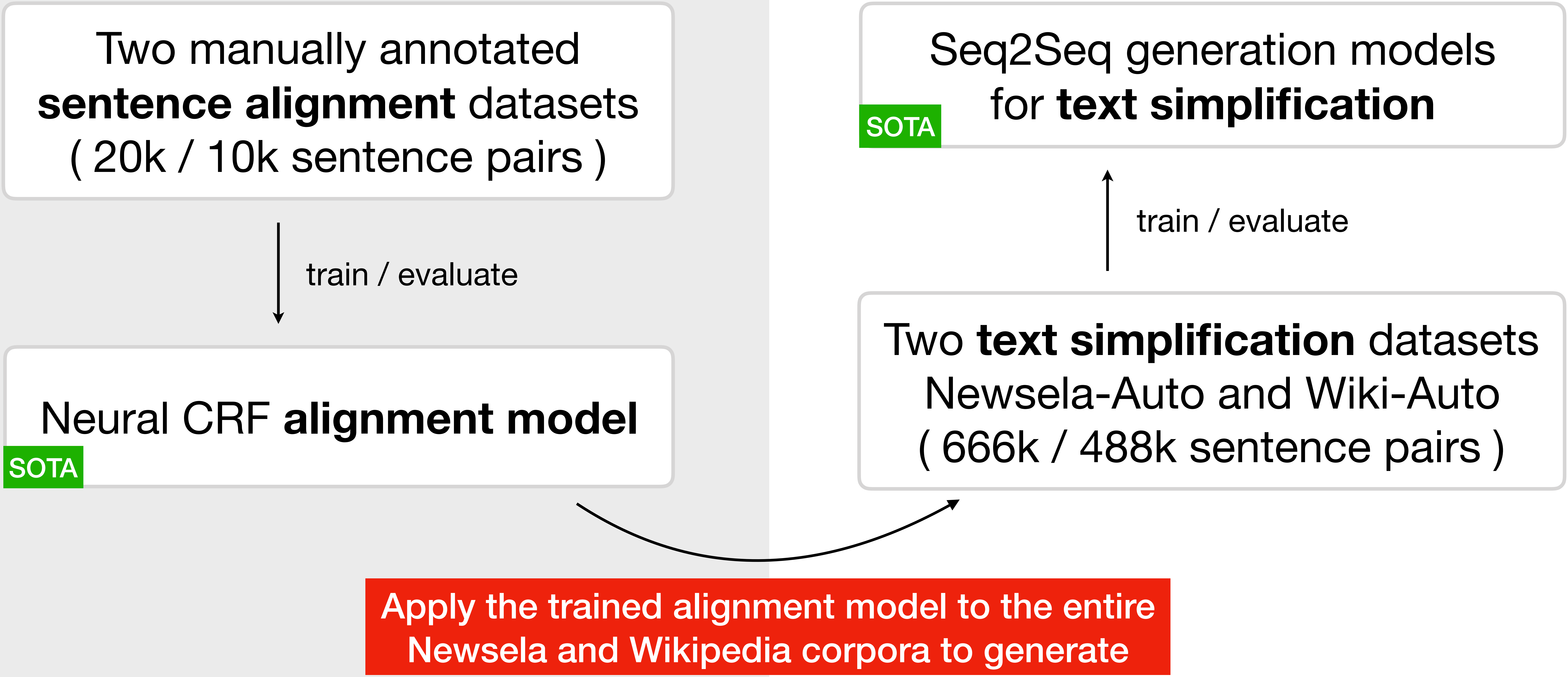
		aligned + partial vs. others*		
		Precision	Recall	F1
Greedy	JaccardAlign (Xu et al., 2015)	98.66	67.58	80.22
Dynamic Programming	MASSAlign (Paetzold et al., 2017)	95.49	82.27	88.39
Greedy	CATS (Štajner et al., 2018)	88.56	91.31	89.92
Threshold	BERT$_{finetune}$	94.99	89.62	92.22
Threshold	BERT$_{finetune}$ + paragraph alignment	98.05	88.63	93.10
CRF	Our CRF aligner	97.86	91.31	95.59

+5.7

* Results are on the manually annotated Newsela dataset.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, Wei Xu. "Neural CRF Model for Sentence Alignment in Text Simplification" (ACL 2020)

Our Work



Two manually annotated **sentence alignment** datasets (20k / 10k sentence pairs)

train / evaluate

Neural CRF **alignment model**

SOTA

Apply the trained alignment model to the entire Newsela and Wikipedia corpora to generate

Two **text simplification** datasets Newsela-Auto and Wiki-Auto (666k / 488k sentence pairs)

train / evaluate

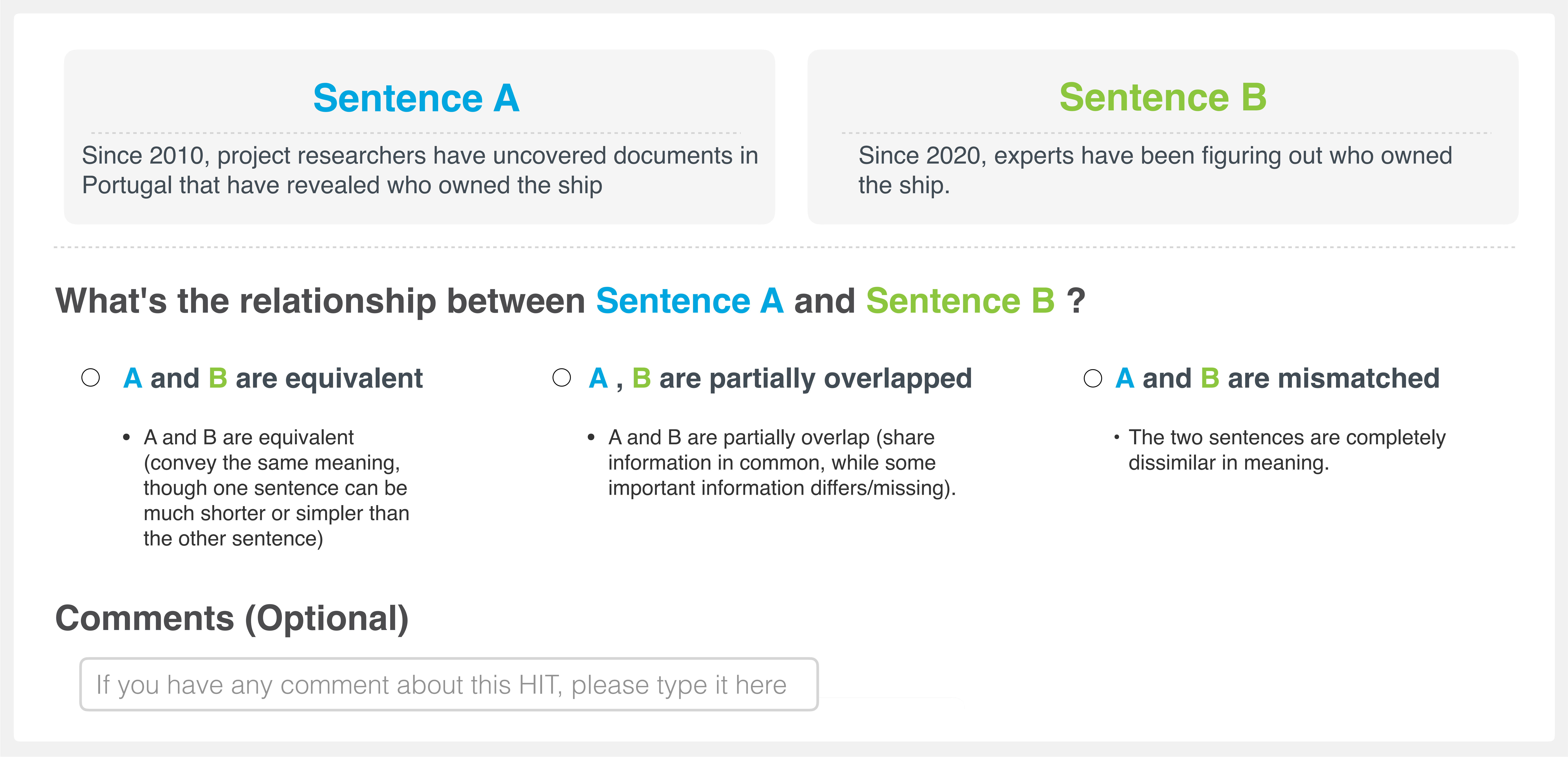
Seq2Seq generation models for **text simplification**

SOTA

Sentence Alignment

Text Simplification

Crowdsourcing Annotation Interface



Sentence A

Since 2010, project researchers have uncovered documents in Portugal that have revealed who owned the ship

Sentence B

Since 2020, experts have been figuring out who owned the ship.

What's the relationship between Sentence A and Sentence B ?

○ A and B are equivalent

- A and B are equivalent (convey the same meaning, though one sentence can be much shorter or simpler than the other sentence)

○ A , B are partially overlapped

- A and B are partially overlap (share information in common, while some important information differs/missing).

○ A and B are mismatched

- The two sentences are completely dissimilar in meaning.

Comments (Optional)

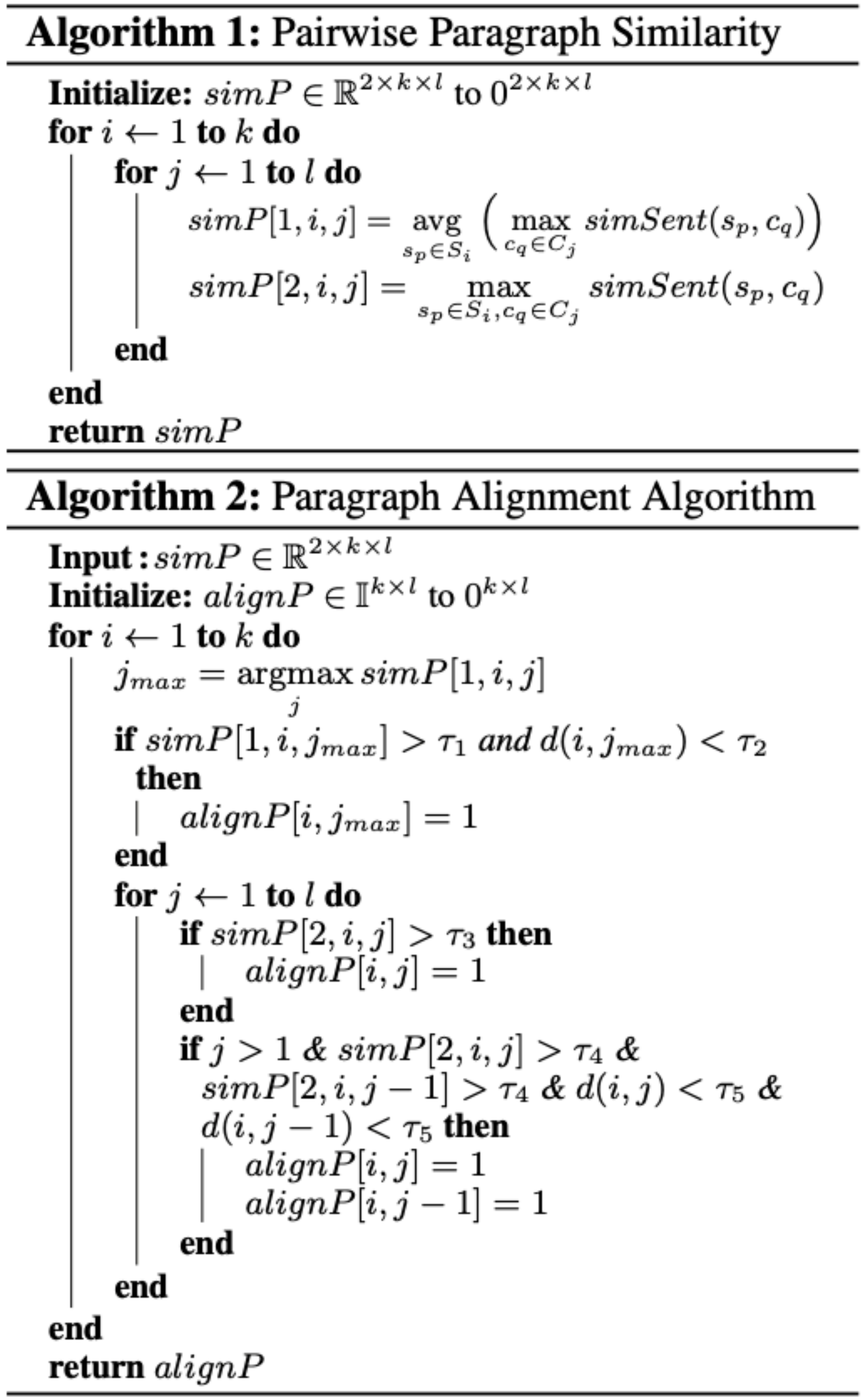
If you have any comment about this HIT, please type it here

Neural CRF Alignment Model

Step 1: Paragraph alignment algorithm

- Based on sentence similarity and vicinity information.
- Significantly improve alignment accuracy (+3 points in precision)

Step 2: Sentence alignment model



Algorithm 1: Pairwise Paragraph Similarity

```
Initialize: $simP \in \mathbb{R}^{2\times k\times l}$ to $0^{2\times k\times l}$
for $i \leftarrow 1$ to $k$ do
    for $j \leftarrow 1$ to $l$ do
        $simP[1,i,j] = \underset{s_p \in S_i}{\text{avg}} \left( \max_{c_q \in C_j} simSent(s_p, c_q) \right)$
        $simP[2,i,j] = \max_{s_p \in S_i, c_q \in C_j} simSent(s_p, c_q)$
    end
end
return $simP$
```

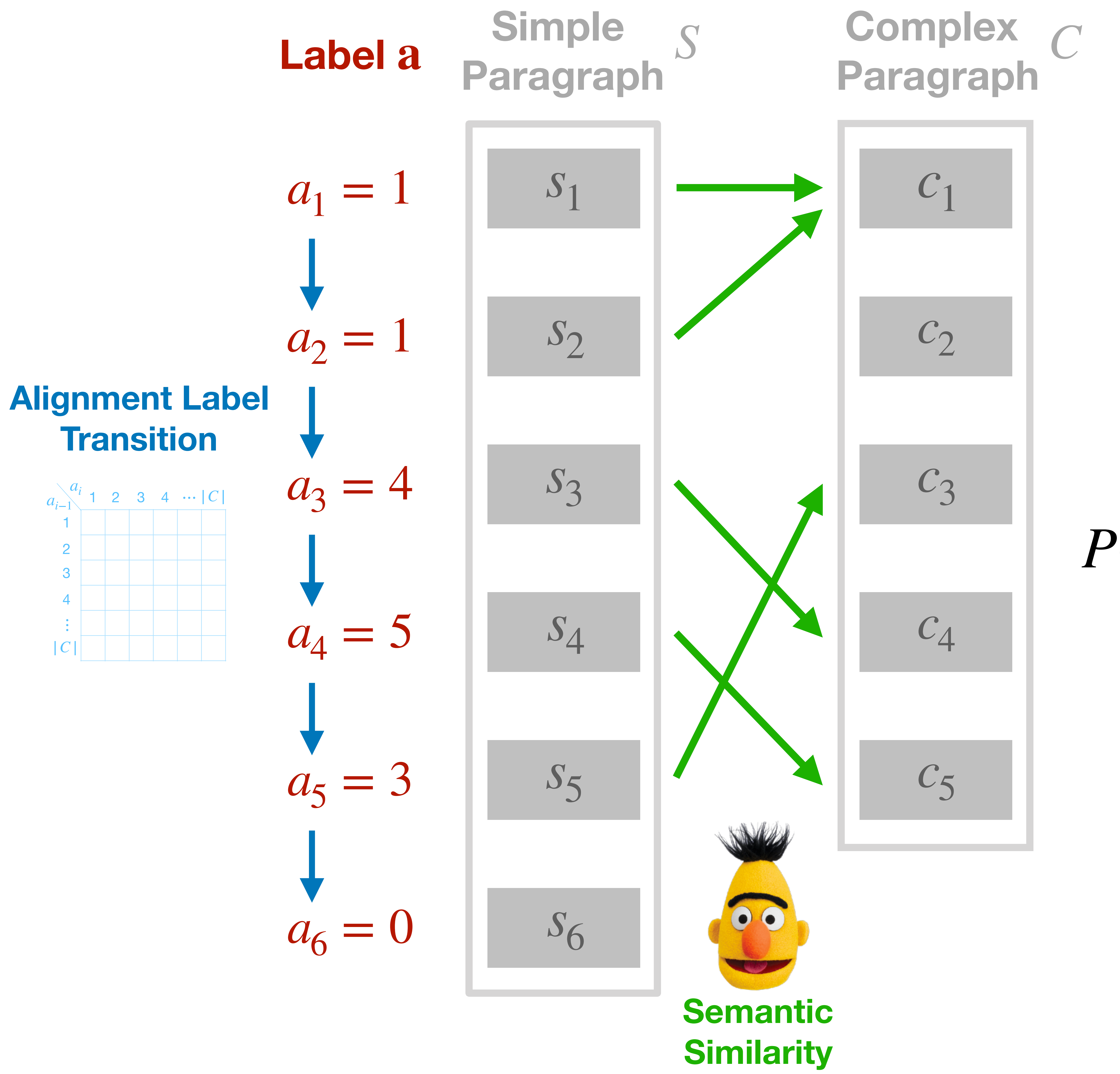
Algorithm 2: Paragraph Alignment Algorithm

```
Input: $simP \in \mathbb{R}^{2\times k\times l}$
Initialize: $alignP \in \mathbb{I}^{k\times l}$ to $0^{k\times l}$
for $i \leftarrow 1$ to $k$ do
    $j_{max} = \underset{j}{\text{argmax}}\, simP[1,i,j]$
    if $simP[1,i,j_{max}] > \tau_1$ and $d(i,j_{max}) < \tau_2$
     then
        $alignP[i,j_{max}] = 1$
    end
    for $j \leftarrow 1$ to $l$ do
        if $simP[2,i,j] > \tau_3$ then
            $alignP[i,j] = 1$
        end
        if $j > 1$ & $simP[2,i,j] > \tau_4$ &
          $simP[2,i,j-1] > \tau_4$ & $d(i,j) < \tau_5$ &
          $d(i,j-1) < \tau_5$ then
            $alignP[i,j] = 1$
            $alignP[i,j-1] = 1$
        end
    end
end
return $alignP$
```

Screenshots of paragraph alignment algorithm

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, Wei Xu. “Neural CRF Model for Sentence Alignment in Text Simplification” (ACL 2020)

Neural CRF Alignment Model



Label a

Simple Paragraph S

Complex Paragraph C

$a_1 = 1$

$a_2 = 1$

$a_3 = 4$

$a_4 = 5$

$a_5 = 3$

$a_6 = 0$

s_1 s_2 s_3 s_4 s_5 s_6

c_1 c_2 c_3 c_4 c_5

Alignment Label Transition

Semantic Similarity

Semantic Similarity

Alignment Label Transition

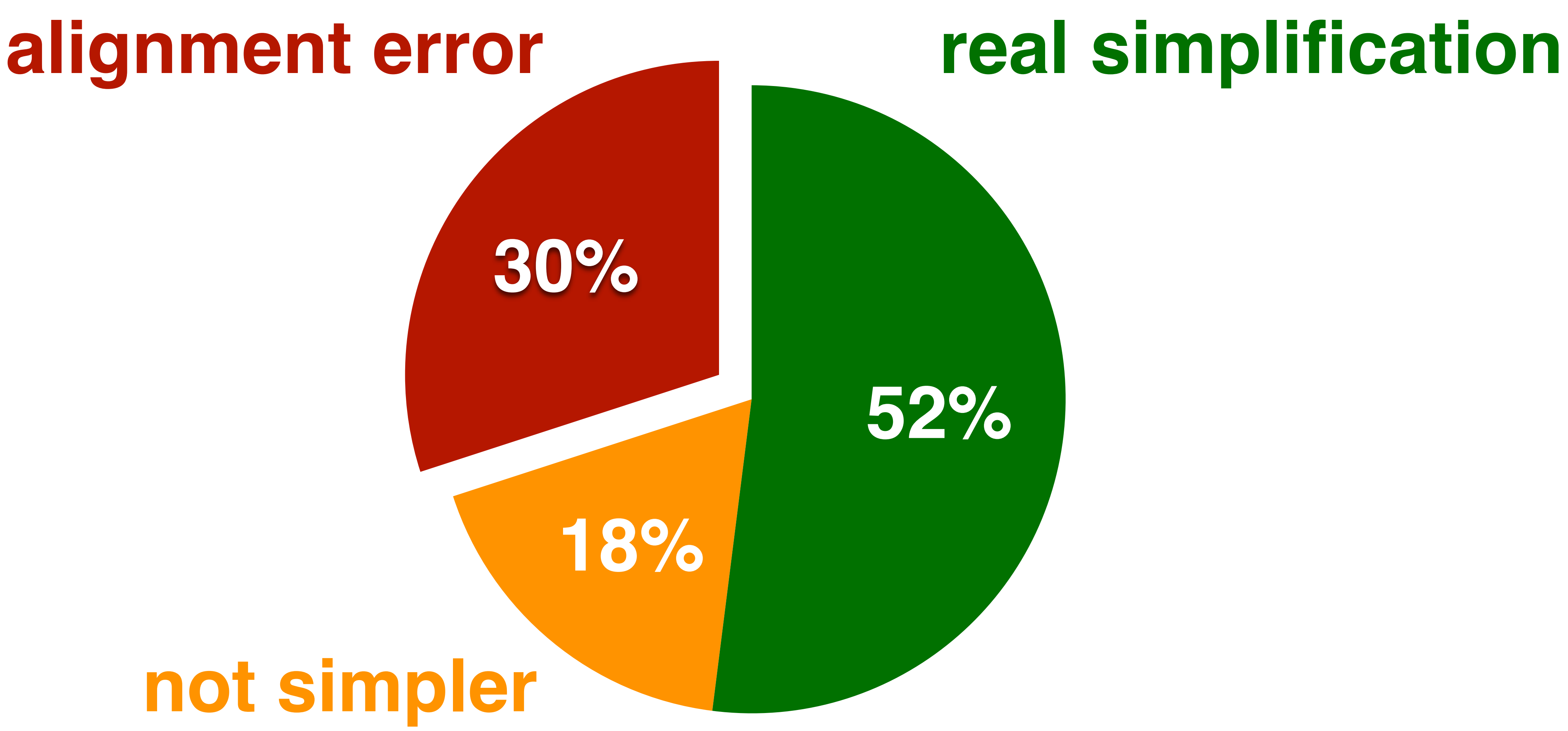
$$\Psi(\mathbf{a}, S, C) = \sum_{i=0}^{|S|} sim(s_i, c_{a_i}) + T(a_i, a_{i-1})$$

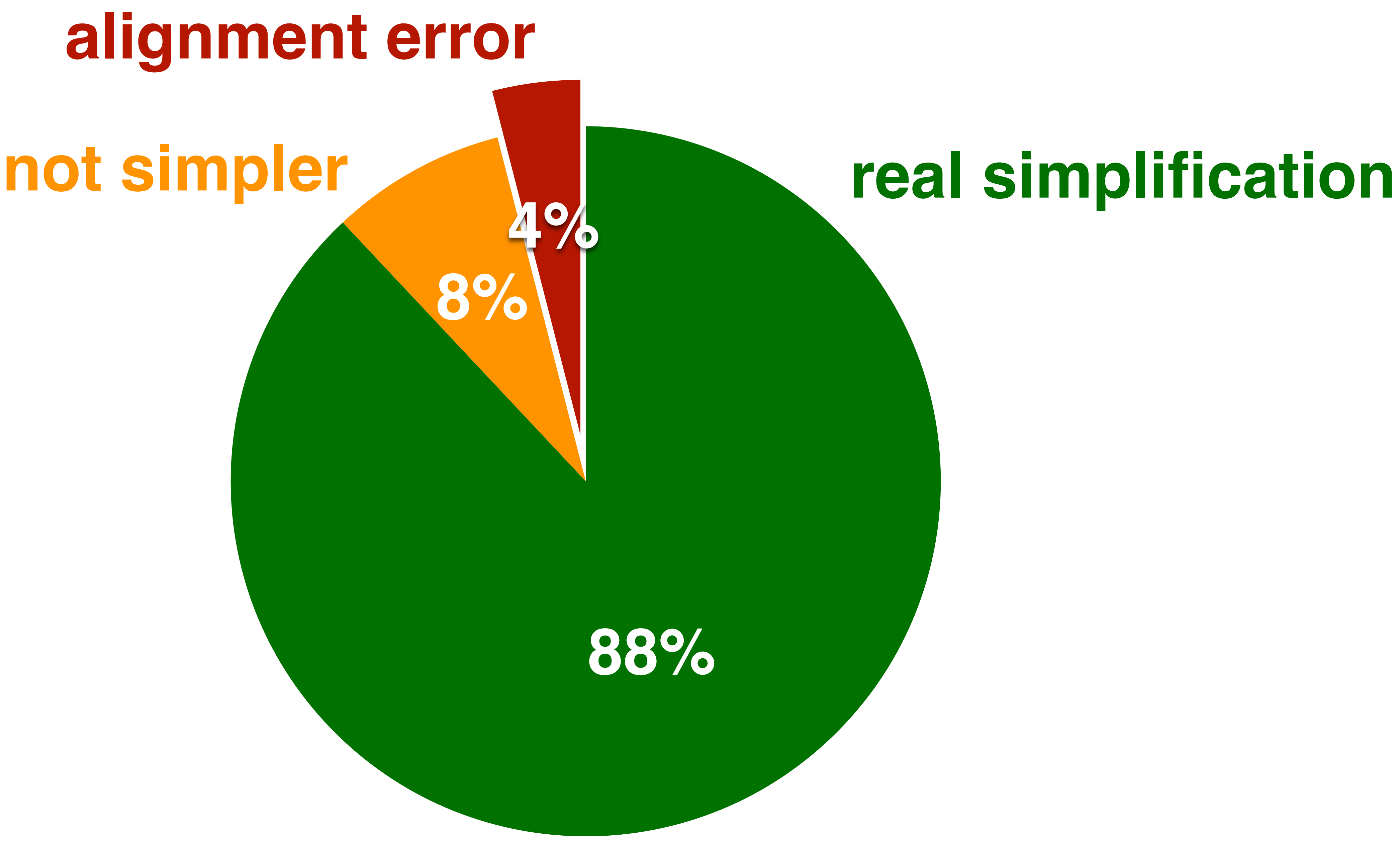
$$P(\mathbf{a} \,|\, S, C) = \frac{\exp(\Psi(\mathbf{a}, S, C))}{\sum_{\mathbf{a} \in \mathbf{A}} \exp(\Psi(\mathbf{a}, S, C))}$$

Linear-chain CRF

all possible alignments (dynamic programming)

New Corpora Contain Way Fewer Errors*





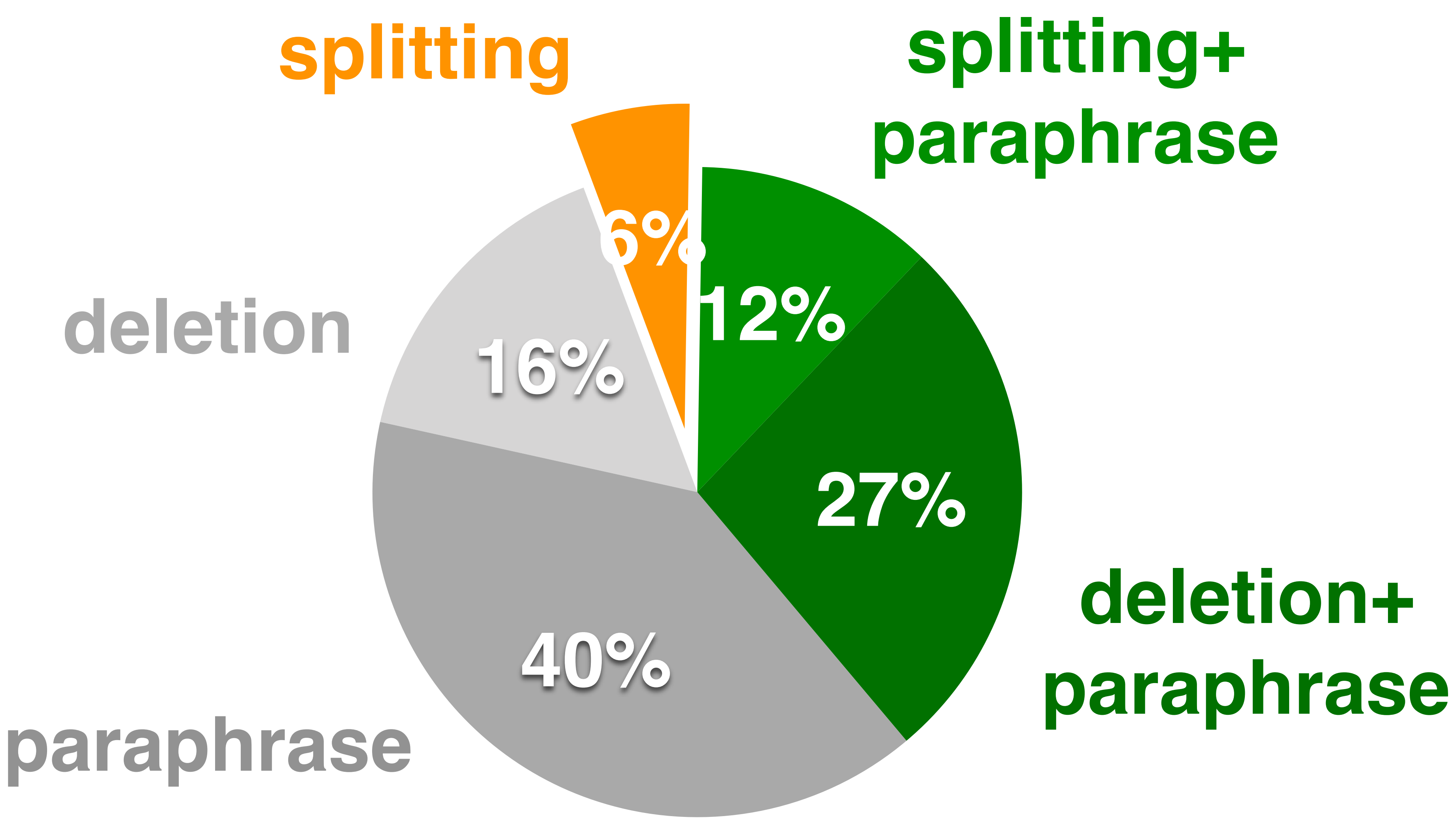
Wiki-Large
(Zhang and Lapata, 2017)

Wiki-Auto (our work)
1.6 times larger — 488k sentence pairs

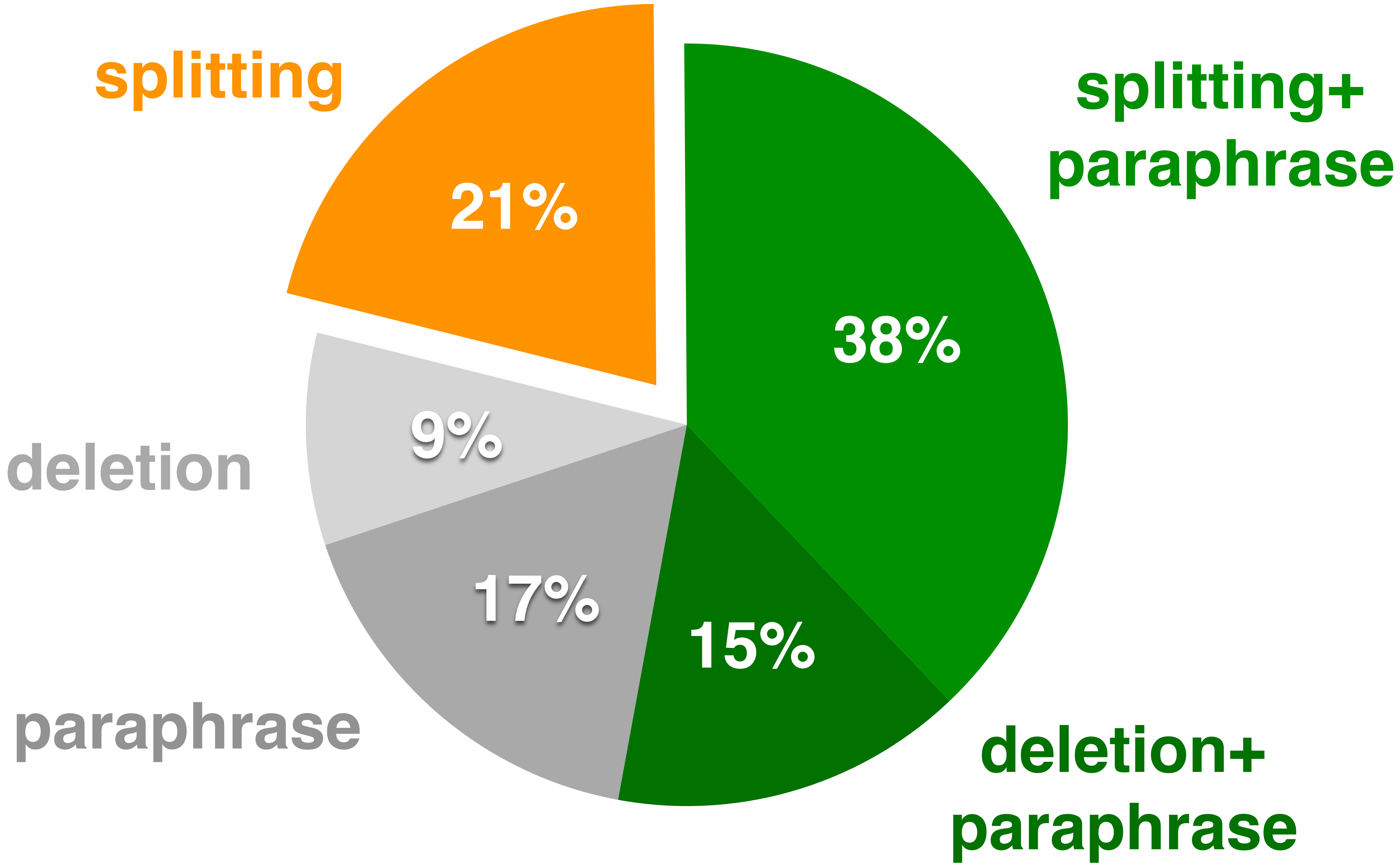
Wiki-Auto has 75% less defective pairs (alignment error + not simpler).

* Based on manual inspection on 100 random sampled sentences from each dataset.

New Corpora Contain More High-quality Simplification*



Newsela
(Xu et al., 2015)



Newsela-Auto (this work)
4.7 times larger — 666k sentence pairs

Newsela-Auto has much more splitting and complex re-writes.

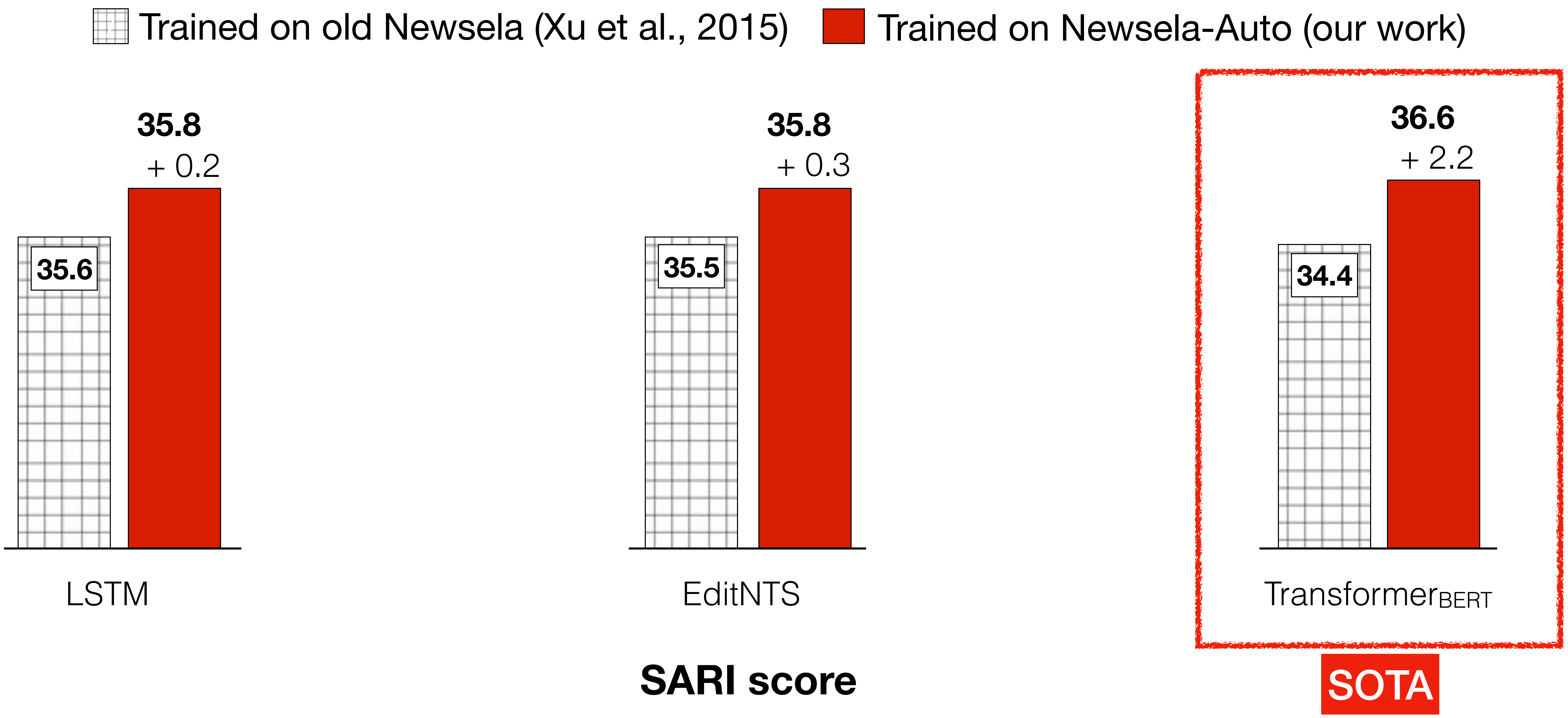
* Based on manual inspection on 100 random sampled sentences from each dataset.

Experiments on Text Simplification

- Transformer$_{BERT}$ (Rothe, Narayan, Severyn, 2020)
- Baseline models
  - LSTM
  - EditNTS (Dong et al., 2019)
  - Rerank (Kriz et al., 2019)
- Datasets
  - This work: **Newsela-Auto** and **Wiki-Auto**
  - Previously existing datasets: Newsela (Xu et al., 2015) and Wiki-Large (Zhang & Lapata, 2017)

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, Wei Xu. "Neural CRF Model for Sentence Alignment in Text Simplification" (ACL 2020)

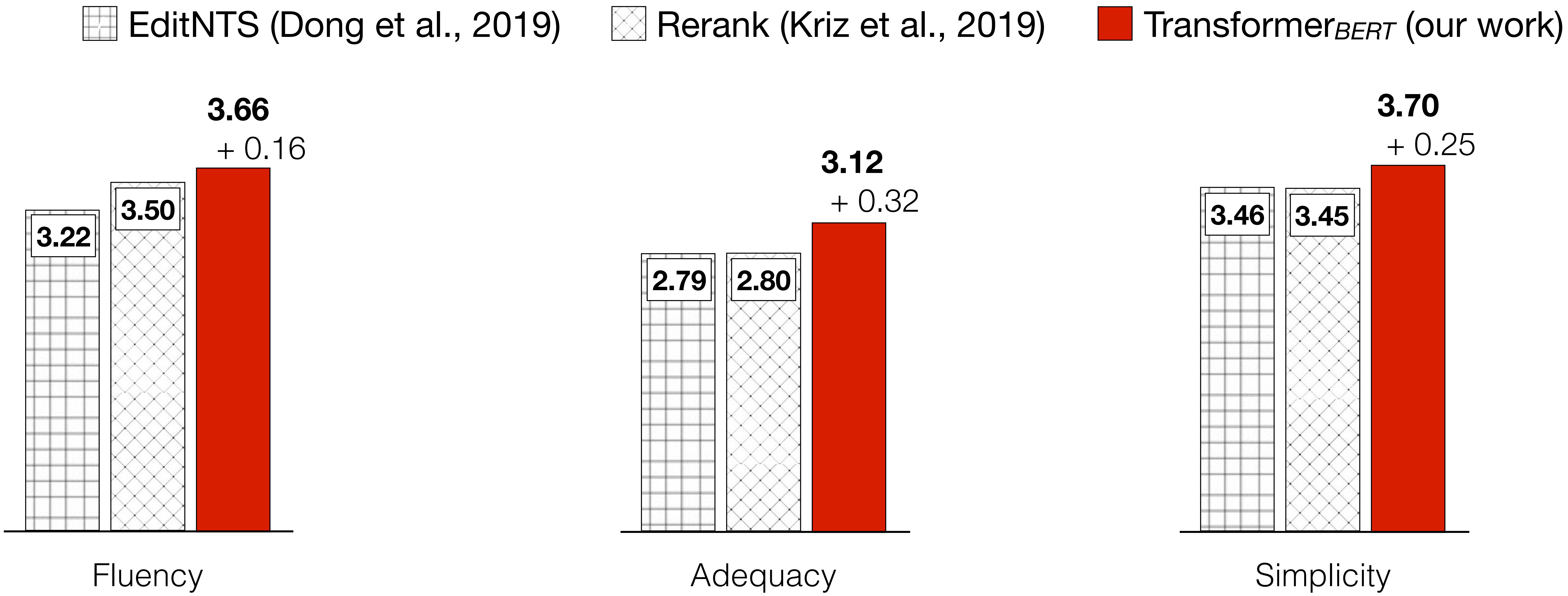
Automatic Evaluation on Text Simplification*



SARI score

* Evaluate on the Newsela-Auto (this work) test set.

Human Evaluation on Text Simplification*

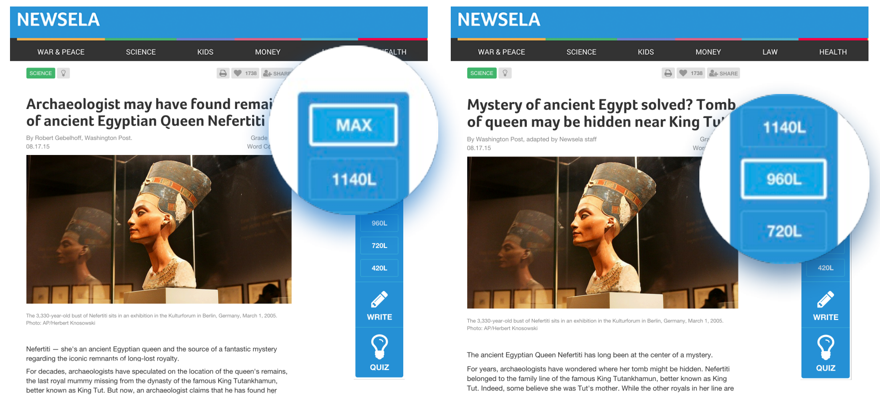


1-5 Likert Scale

* Evaluate on the Old Newsela (Xu et al., 2015) test set.

Open Source

Code and data are available at - https://github.com/chaojiang06/wiki-auto



Archaeologist may have found remai of ancient Egyptian Queen Nefertiti

By Robert Gebelhoff, Washington Post.
08.17.15

The 3,330-year-old bust of Nefertiti sits in an exhibition in the Kulturforum in Berlin, Germany, March 1, 2005. Photo: AP/Herbert Knosowski

Nefertiti — she's an ancient Egyptian queen and the source of a fantastic mystery regarding the iconic remnants of long-lost royalty.

For decades, archaeologists have speculated on the location of the queen's remains, the last royal mummy missing from the dynasty of the famous King Tutankhamun, better known as King Tut. But now, an archaeologist claims that he has found her

Mystery of ancient Egypt solved? Tomb of queen may be hidden near King Tu

By Washington Post, adapted by Newsela staff
08.17.15

The 3,330-year-old bust of Nefertiti sits in an exhibition in the Kulturforum in Berlin, Germany, March 1, 2005. Photo: AP/Herbert Knosowski

The ancient Egyptian Queen Nefertiti has long been at the center of a mystery.

For years, archaeologists have wondered where her tomb might be hidden. Nefertiti belonged to the family line of the famous King Tutankhamun, better known as King Tut. Indeed, some believe she was Tut's mother. While the other royals in her line are

Simple English
WIKIPEDIA

Neural CRF Model for Sentence Alignment in Text Simplification

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, Wei Xu (ACL 2020)

Take Aways

- **Controllable Generation Model**

  Also useful for semantics and natural language understanding.

  - Neural semi-Markov CRF for Monolingual Word Alignment (Lan*, Jiang* & Xu, ACL 2021)

  How to incorporate linguistic rules with neural networks?

  - Controllable Text Simplification with Explicit Paraphrasing (Maddela, Alva-Manchego & Xu, NAACL 2021)

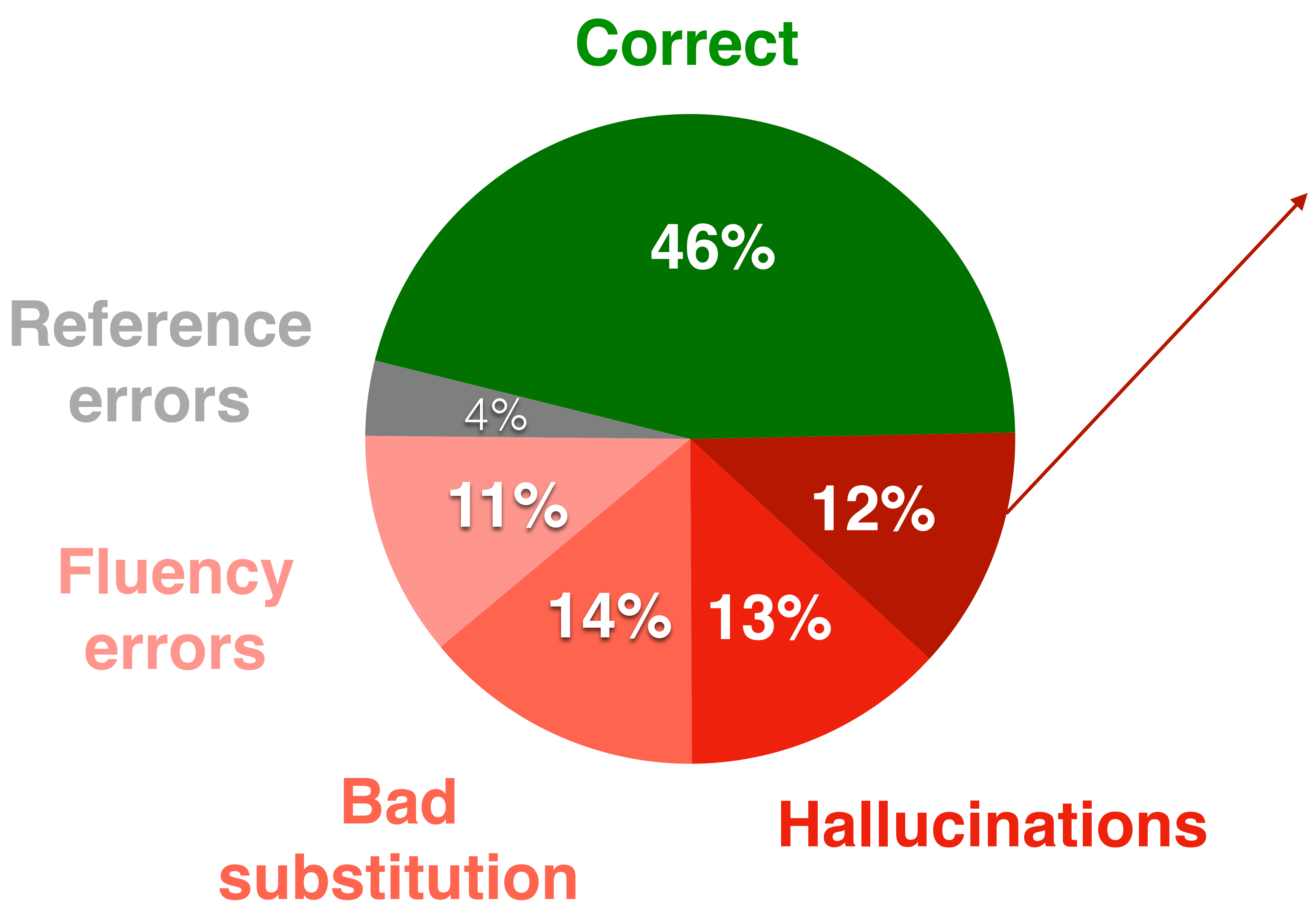
- **High-quality Training Data**

  Performance gains from better data are huge!

  - Neural CRF Model for Sentence Alignment in Text Simplification (Jiang, Maddela, Lan, Zhong & Xu, ACL 2020)
  - Discourse Level Factors for Sentence Deletion in Text Simplification (Zhong, Jiang, Xu & Li, AAAI 2020)
  - A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification (Maddela & Xu, EMNLP 2018)
  - Optimizing Statistical Machine Translation for Text Simplification (Xu et al., TACL 2016)
  - Problems in Current Text Simplification Research: New Data Can Help (Xu et al., TACL 2015)

What lie in the future? Here is an error analysis.

Manually inspected 100 simplifications by our model from the **Newsela-Auto** test set.



Anaphora resolution / Entailment

Complex	Sea slugs dubbed sacoglossans are some of the most remarkable biological burglars on the planet
Simple	The sea slugs known as sacoglossans are some of the most interesting burglars on the planet.
Our Model	Scientists say **these** are some of the most **interesting creatures** on the planet.

Check out the code/data at https://github.com/mounicam/controllable_simplification

Thank you!

https://cocoxu.github.io/

thank u 4 ur time

thanku

I am grateful

thanking you

thx

thanks a lot

appreciate it

gratitude

gramercies

3x

tyvm

thanks

say thanks

thank you very much

thnx

thanks a ton

wawwww thankkkkkkkkkkk you alottttttttttt!

I can no other answer make but thanks, and thanks, and ever thanks.

NSF DARPA IARPA criteo figure eight NVIDIA