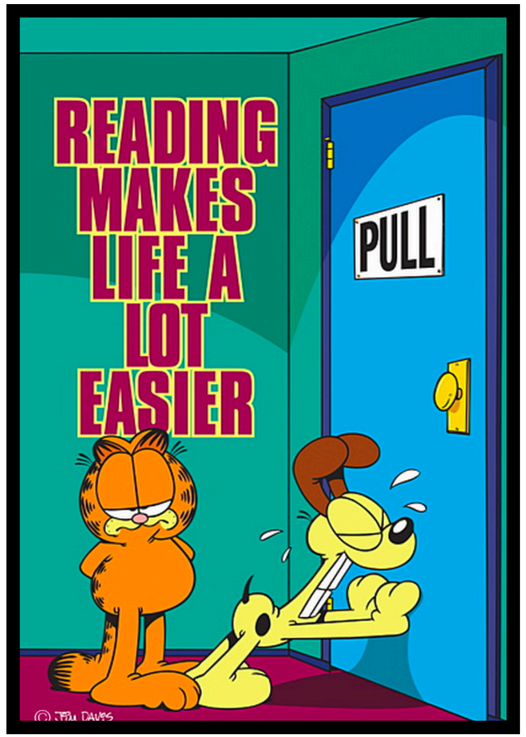


(Image Source: Garfield)

Automatic Text Simplification

Wei Xu
School of Interactive Computing
Georgia Institute of Technology
@cocoweixu
wei.xu@cc.gatech.edu
https://cocoxu.github.io/

Georgia Tech

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.

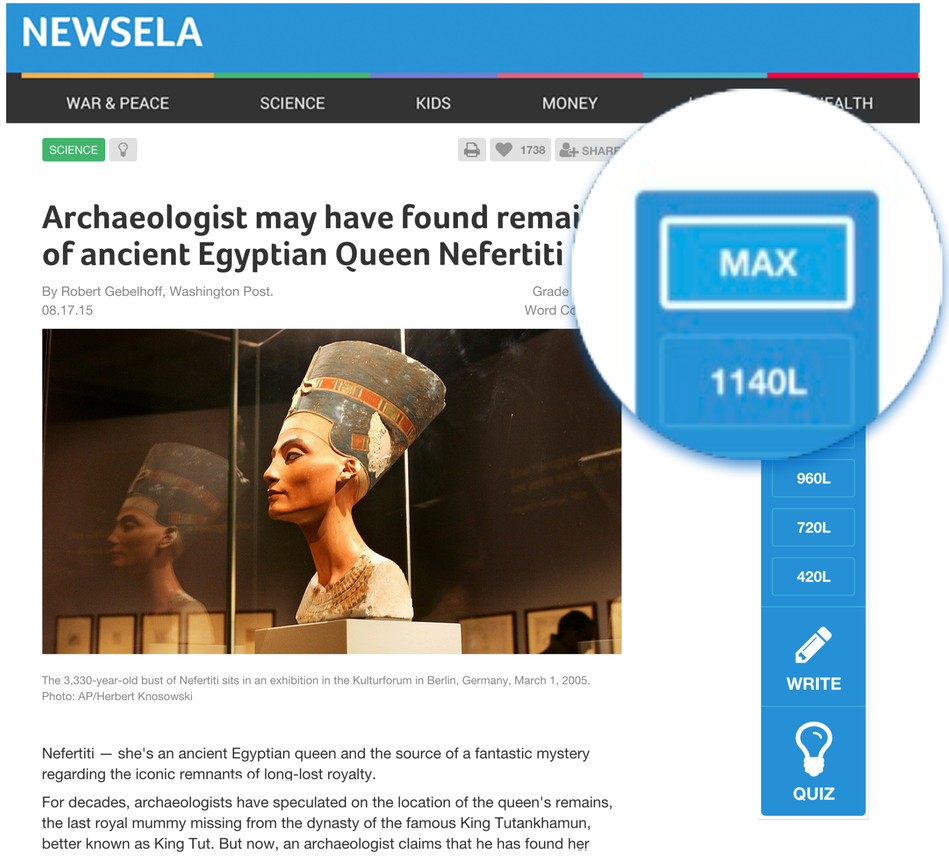
And when we die, these ~~dense, calcified~~ micro-fossils remain intact, even ~~as most of the rest of us decomposes~~.

paraphrase

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

Human Text Simplification

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



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)

Why Text Simplification?

It can help a lot of people!

- Children (Leonardo et al., 2018)
- Second language learners (Housel et al., 2020)
- Deaf and hard-of-hearing students (Alonzo et al., 2020)
- People with dyslexia (Rello at al., 2013)
- People with autism spectrum disorder (González-Navarro et al., 2014)

research on education using Newsela data

our collaborators at RIT

- 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)

Automatic Text Simplification

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

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.

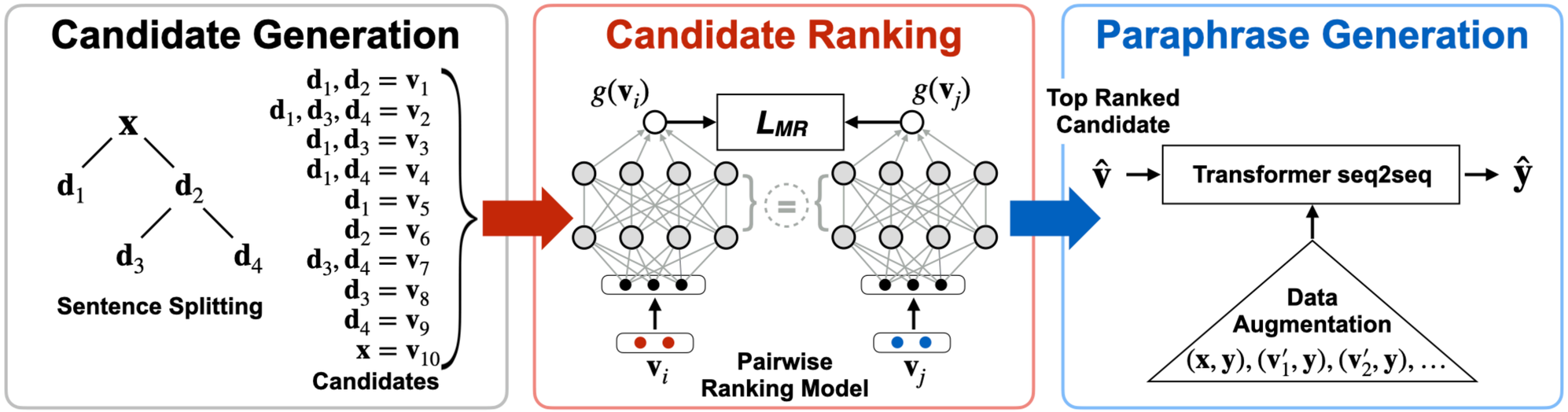
Part 1 — Controllable Generation Model

Controllable Text Simplification with Explicit Paraphrasing

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

Controllable Text Generation

We incorporate linguistic knowledge into neural seq-to-seq models to improve controllability.



INPUT x : *The exhibition, which opened Oct. 8 and runs through Jan. 3, features 27 self-portraits.* **REFERENCE y** : *The show started Oct. 8. It ends Jan. 3.*

$\mathbf{d}_1$: *The exhibition features 27 self-portraits.* $\mathbf{d}_2$: *The exhibition opened Oct. 8 and runs through Jan. 3.*

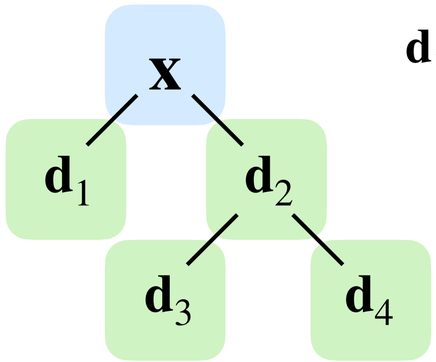
$\mathbf{d}_3$: *The exhibition opened Oct. 8.* $\mathbf{d}_4$: *The exhibition runs through Jan. 3.* $\hat{\mathbf{v}} = \mathbf{v}_7$: *The exhibition opened Oct. 8. The exhibition runs through Jan. 3.*

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



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.

Split sentences:

The exhibition features 27 portraits.

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

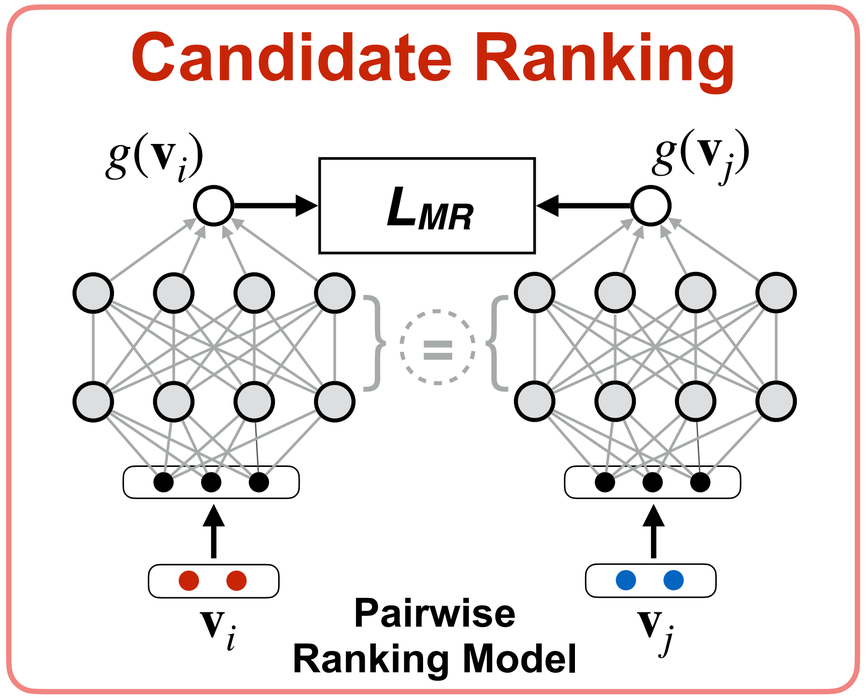
The exhibition opened Oct. 8.

The exhibition runs through Jan. 3.

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

Step 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)

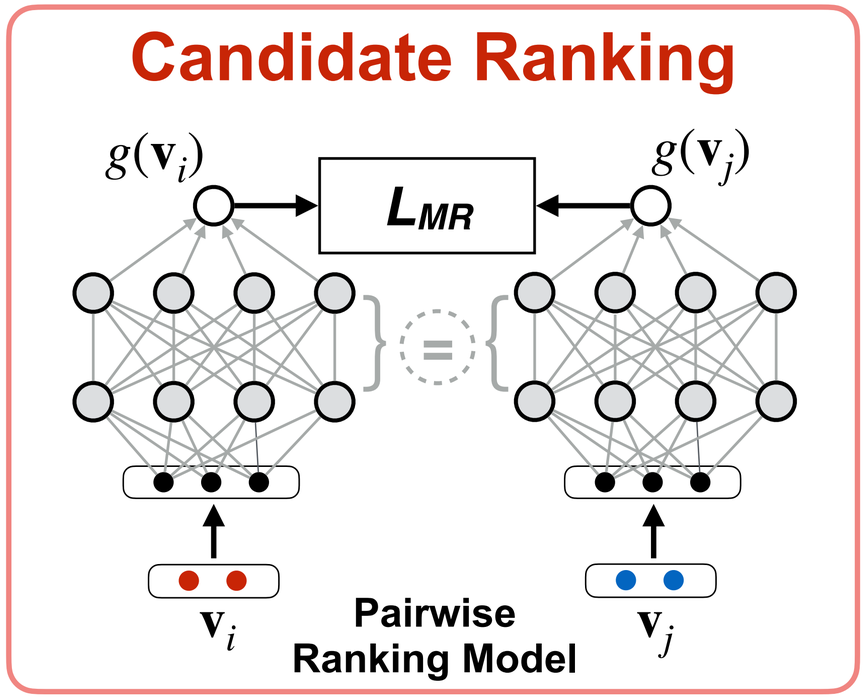
Human reference:

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

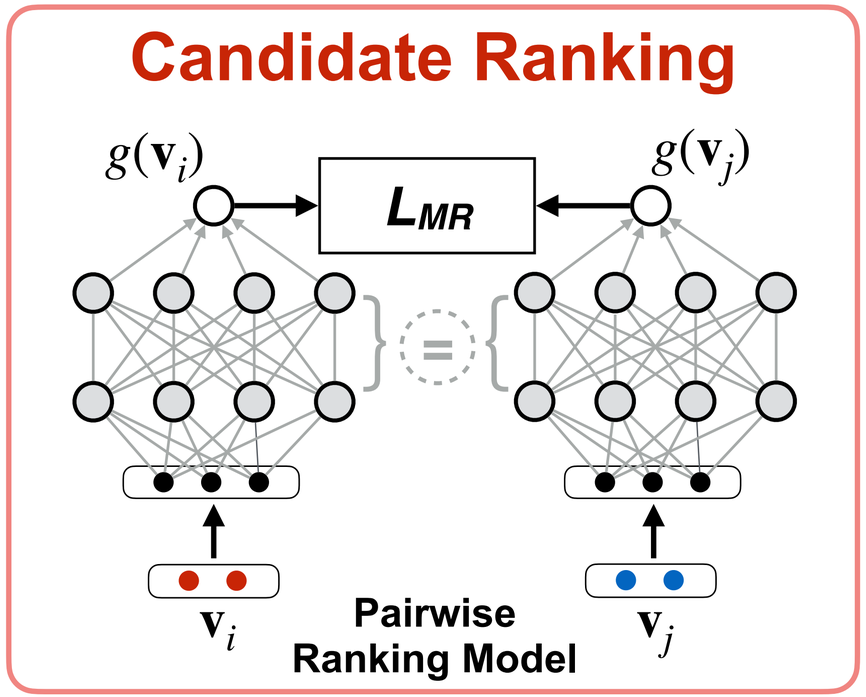
$$g^*(\mathbf{v}_i, \mathbf{y}) = e^{-\lambda||\phi_{\mathbf{v}_i} - \phi_{\mathbf{y}}||} \times BERTScore(\mathbf{v}_i, \mathbf{y})$$

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:

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

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

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

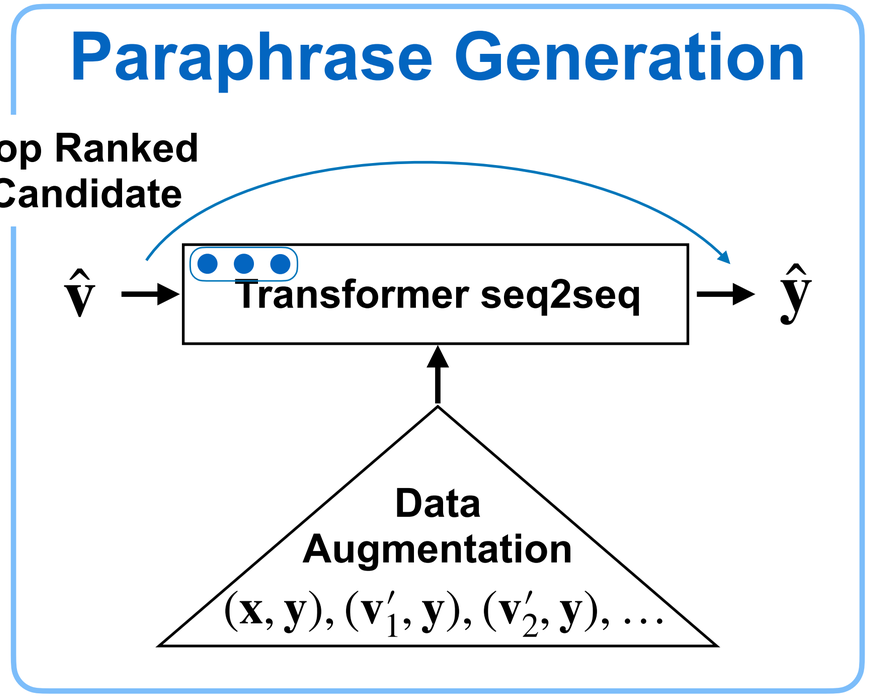
Length-penalized BERTScore

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)



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)

Controllable Text Generation

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

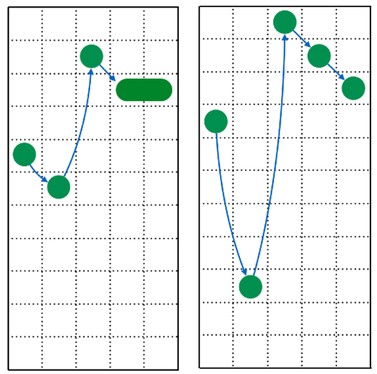
Input: Experts say China's air pollution exacts a tremendous toll on human health.

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.

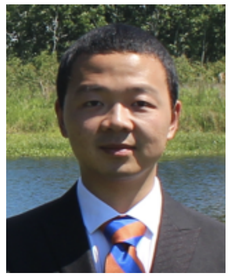
Mounica Maddela, Fernando Alva-Manchego, Wei Xu. "Controllable Text Simplification with Explicit Paraphrasing" (NAACL 2021)

Part 2 — Monolingual Word Alignment



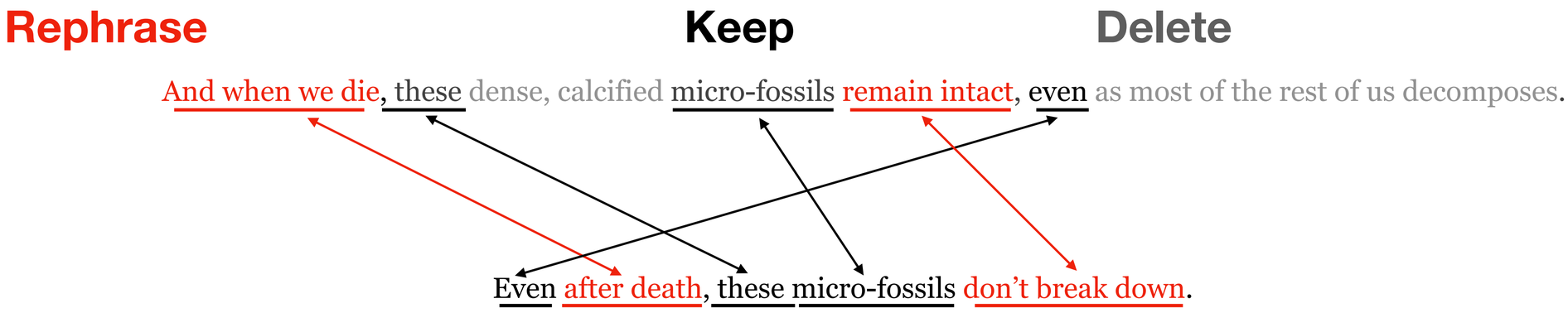
Neural semi-Markov CRF for Monolingual Word Alignment

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



Monolingual ~~Word~~ Span Alignment

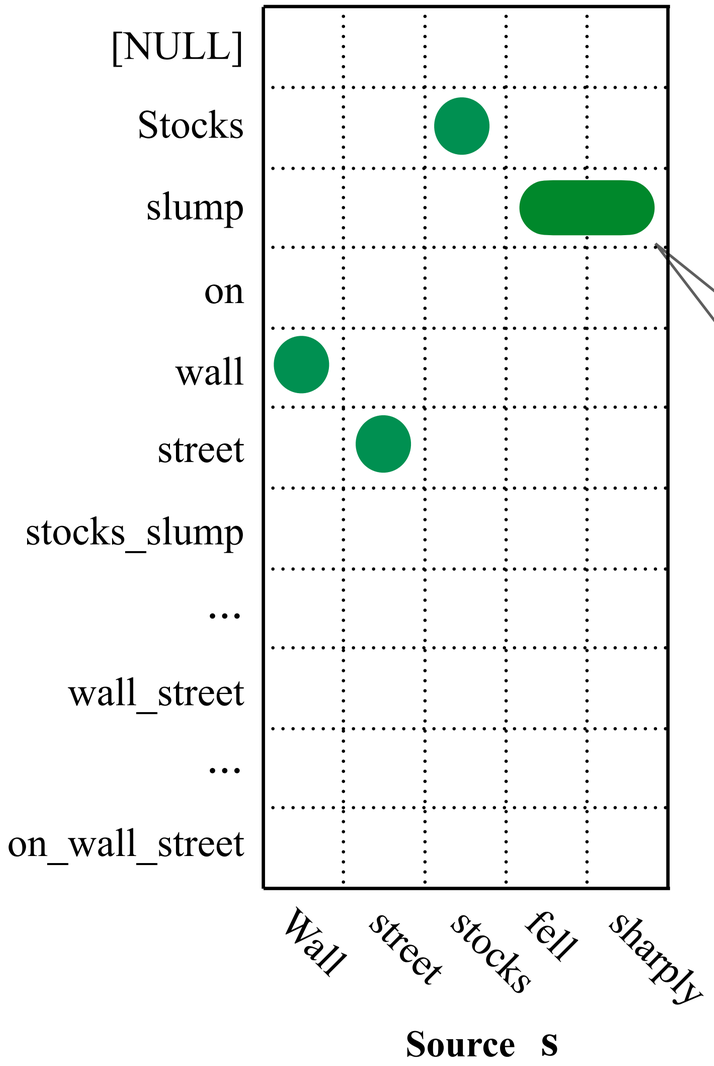
Can support not only text-to-text generation tasks, but also natural language understanding tasks.



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

Semi-CRF Word Alignment Model

Span Interaction Matrix



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

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

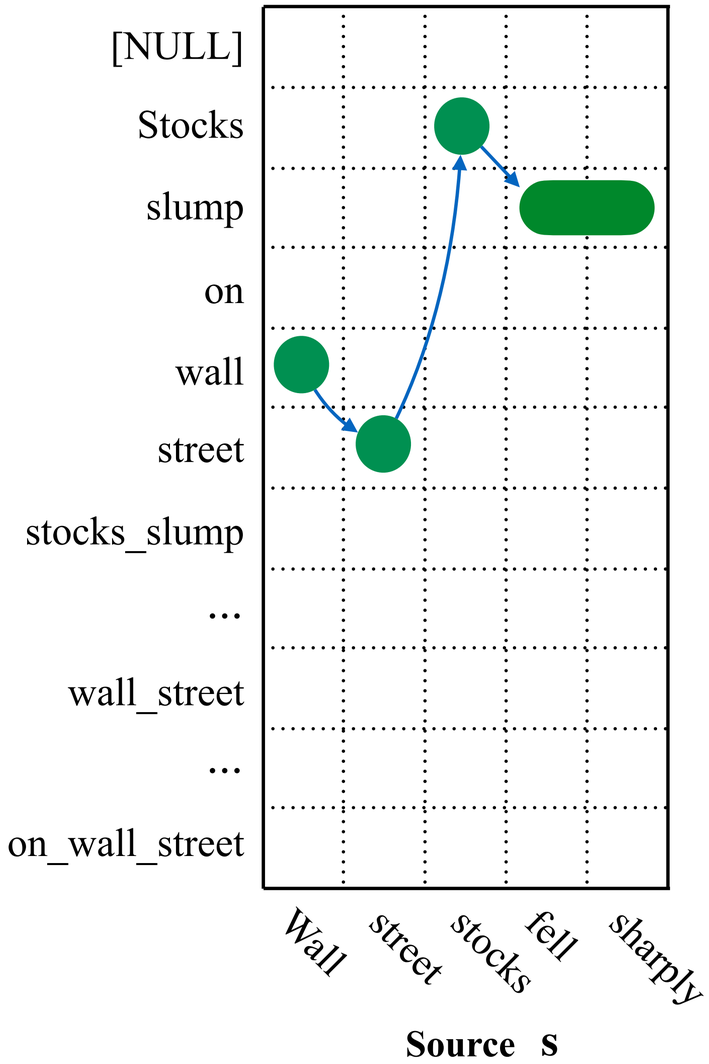
$$score(s_i, t_j) = \text{FFNN}(h_i^s; h_j^t; |h_i^s - h_j^t|; h_i^s \circ h_j^t)$$

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**

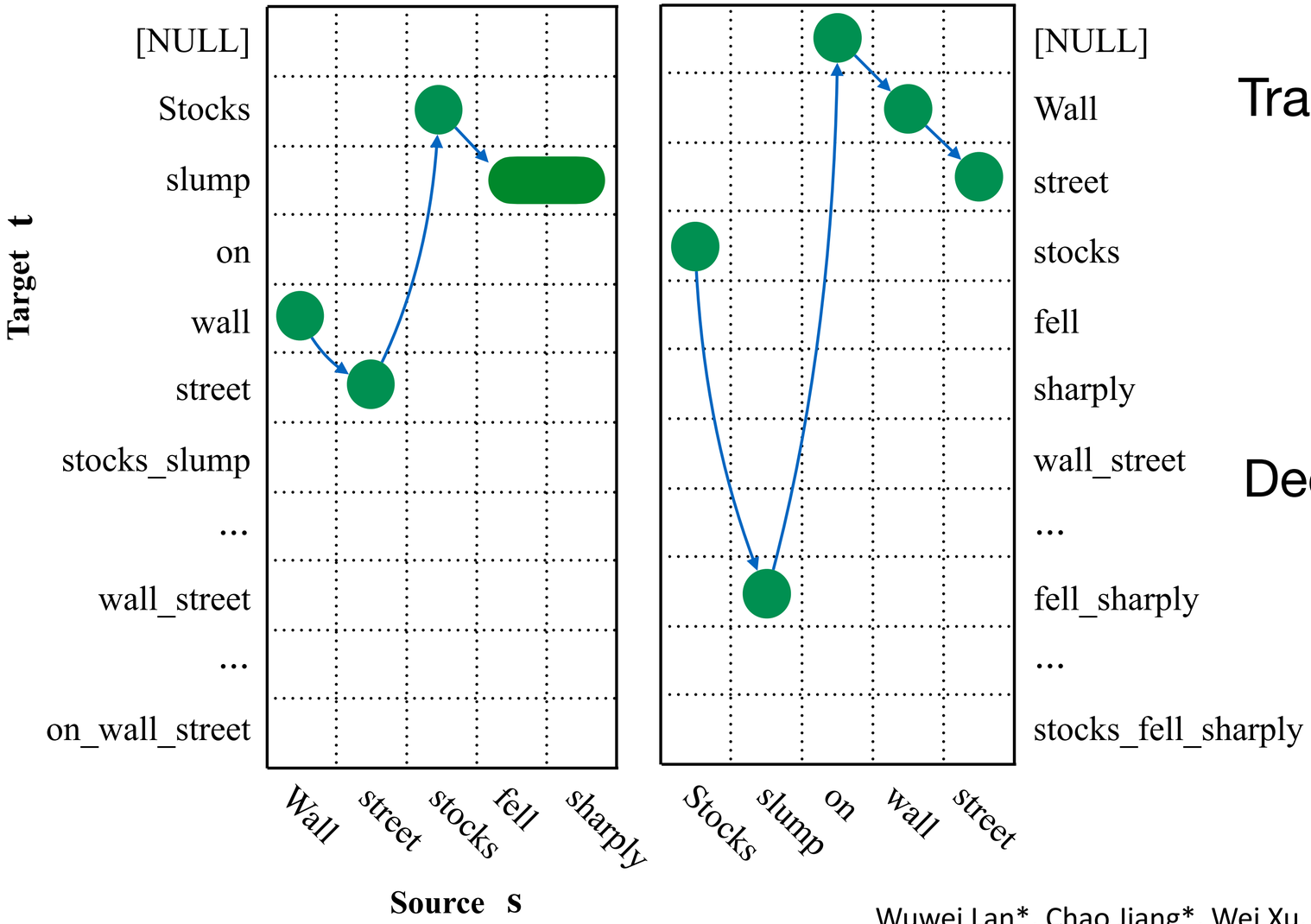
$$P(\mathbf{a} \mid \mathbf{s}, \mathbf{t}) = \frac{exp\left(\Psi(\mathbf{a}, \mathbf{s}, \mathbf{t})\right)}{\sum_{\mathbf{a} \in \mathbf{A}} exp\left(\Psi(\mathbf{a}, \mathbf{s}, \mathbf{t})\right)}$$

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)

Our Work — Automatic Text Simplification

- **Controllable Generation Model**
  - Neural semi-Markov CRF for Monolingual Word Alignment (Lan*, Jiang* & Xu, ACL 2021)
  - Controllable Text Simplification with Explicit Paraphrasing (Maddela, Alva-Manchego & Xu, NAACL 2021)

Also useful for semantics and natural language understanding.

How to incorporate linguistic rules with neural networks?

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

Performance gains from better data are huge!

- BiSECT: Learning to Split and Rephrase Sentences with Bitexts (Kim*, Maddela*, Kriz, Xu, Callison-Burch, EMNLP 2021)
- 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)

SimplePPDB++

A database of 14.1 million paraphrase rules with improved complexity ranking scores.

Paraphrase Rule	Score
→ *self-supporting*	0.93
self-reliant → *self-sufficient*	0.48
→ *self-sustainable*	-0.60
→ *possible*	0.94
viable → *realistic*	0.15
→ *plausible*	-0.91
→ *in-depth review*	0.89
detailed assessement → *careful examination*	0.28
→ *comprehensive evaluation*	-0.87

All our code/data/models are open-source! You can find them at: https://cocoxu.github.io/

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

NLP X Research Lab

Georgia Tech School of Interactive Computing
College of Computing

We design machine learning algorithms to help computer to understand and generate human languages.
X = Machine Learning, Education, Accessibility, Social Media, …

Natural Language Processing

- Text generation and evaluation
- Semantics and structured prediction
- Misinformation, social media
- Offensive/biased language
- Interactive NLP systems

Machine Learning

- Sequence-to-sequence models
- Controllability of neural networks
- Cross-lingual transfer learning
- Learning from noisy data
- Data augmentation

Wei Xu
Assistant Professor

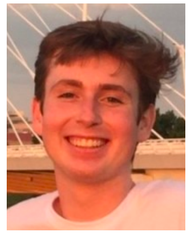
Mounica Maddela
PhD student

Chao Jiang
PhD student

Yao Dou
PhD student

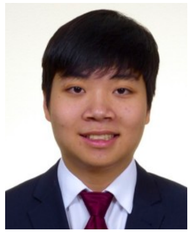
Jagriti Sikka
MS student

Angana Borah
MS student



David Heineman
Undergrad

Ema Goh
Undergrad



Jonathan Zheng
Undergrad

Michael Ryan
Undergrad

Alex Kelso
Undergrad

Dylan Small
Undergrad

Srushti Nandu
Intern