Controllable Text Simplification with Explicit Paraphrasing

Mounica Maddela , Fernando Alva-Manchego, Wei Xu

NAACL 2021

Text Simplification

- Improve the readability of texts with simpler grammar and word choices while preserving meaning
- Applications
  - Reading assistance to children
  - Non-native speakers
  - People with reading disabilities
- Tasks
  - Machine Translation (reduces post-edit efforts, improves fluency)
  - Information extraction
  - SRL, Parsing

Example and Issues

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.*

- Text simplification – sophisticated combination of
  - Deletion
  - Paraphrasing
  - Sentence splitting
- Issues
  - Mostly do deletion
  - Tends to generate very short outputs
  - At the cost of meaning preservation

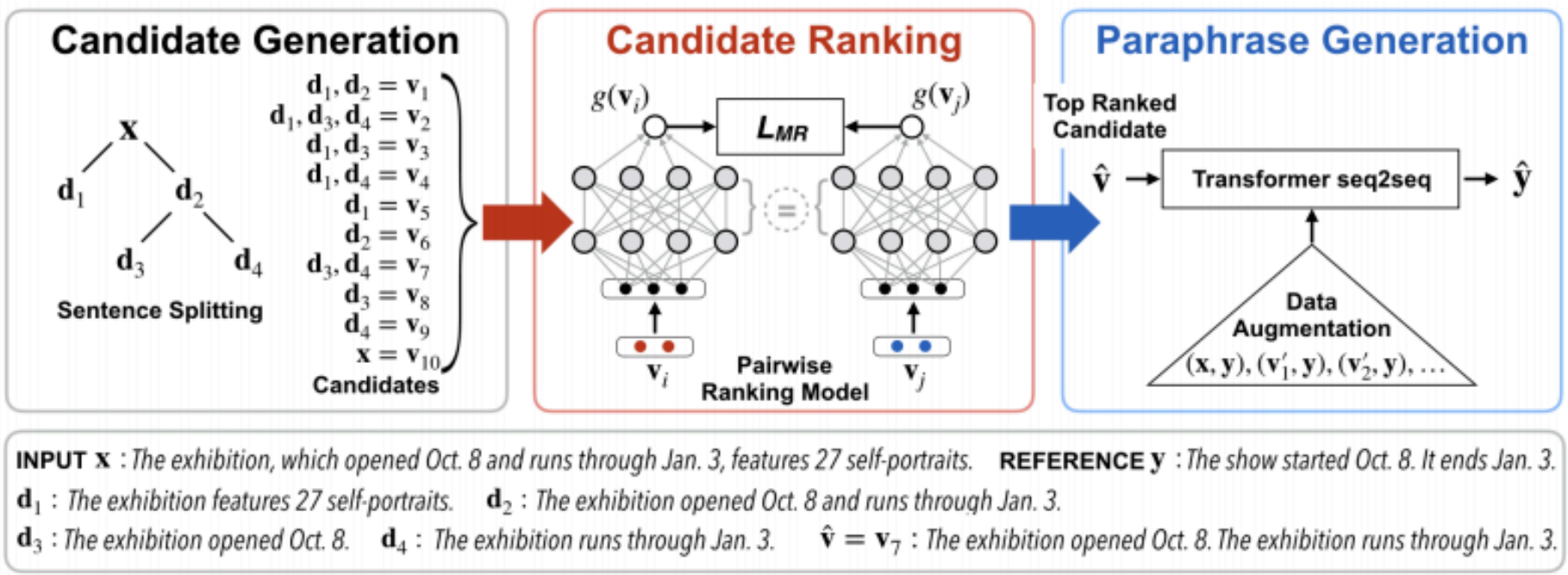
	OLen	**%new**	**%eq**	**%split**
Complex (input)	20.7	0.0	100.0	0.0
Narayan and Gardent (2014)†	10.4	0.7	0.8	0.4
Zhang and Lapata (2017)†	13.8	8.1	16.8	0.0
Dong et al. (2019)†	10.9	8.4	4.6	0.0
Kriz et al. (2019)†	10.8	11.2	1.2	0.0
LSTM	17.0	6.1	28.4	1.2
Our Model	17.1	17.0	3.0	31.8
Simple (reference)	17.9	29.0	0.0	30.0

Table 1: Output statistics of 500 random sentences from the Newsela test set. Existing systems rely on deletion and do not paraphrase well. **OLen**, **%new**, **%eq** and **%split** denote the average output length, percentage of new words added, percentage of system outputs that are identical to the inputs, and percentage of sentence splits, respectively. †We used the system outputs shared by their authors.

Goal of this paper

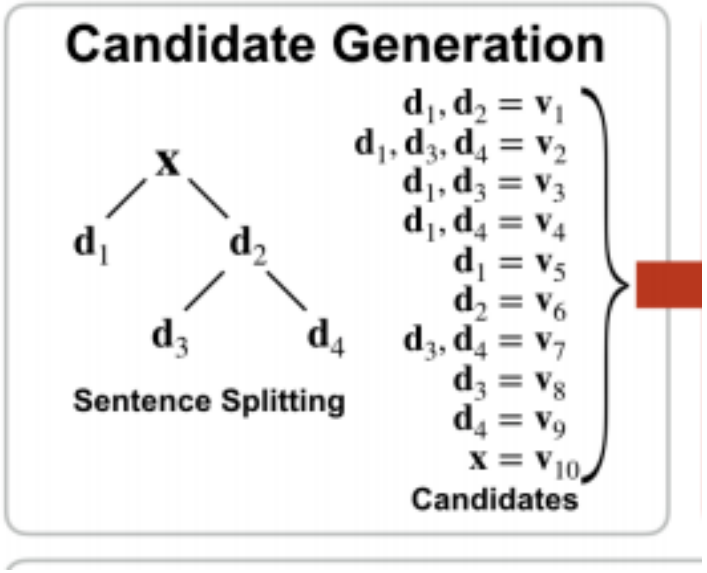
- Propose a sentence simplification model
  - Deletion
  - Paraphrasing
  - Sentence splitting
  - Controllability
    - What constitutes simplified text for one group of users may not be acceptable for other groups?
    - Control for
      - Paraphrasing (soft constraint on % of words copied from input)
      - Select candidates that underwent a desired amount of splitting and/or deletion
- Better than existing models
- New test dataset with multiple human references to specifically evaluate lexical paraphrasing

Overview


Candidate Generation
x
d1
d2
d3
d4
Sentence Splitting
d1, d2 = v1
d1, d3, d4 = v2
d1, d3 = v3
d1, d4 = v4
d1 = v5
d2 = v6
d3, d4 = v7
d3 = v8
d4 = v9
x = v10
Candidates
Candidate Ranking
g(vi)
g(vj)
LMR
vi
vj
Pairwise
Ranking Model
Paraphrase Generation
Top Ranked
Candidate
v̂
Transformer seq2seq
ŷ
Data
Augmentation
(x, y), (v'1, y), (v'2, y), ...
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.
d1 : The exhibition features 27 self-portraits.
d2 : The exhibition opened Oct. 8 and runs through Jan. 3.
d3 : The exhibition opened Oct. 8.
d4 : The exhibition runs through Jan. 3.
v̂ = v7 : The exhibition opened Oct. 8. The exhibition runs through Jan. 3.

Candidate Generation

- Splitting and Deletion
  - Generate intermediate simplifications that have undergone splitting and deletion
  - SOTA structural simplification, DisSim
    - 35 hand-crafted rules to break down complex sentence into a set of hierarchically organized sub-sentences
- Paraphrasing
  - Neural Deletion and Split module trained on text simplification corpus
  - Transformer seq2seq model + copy
  - Constrained beam search
    - 10 outputs with splitting
    - 10 outputs without splitting



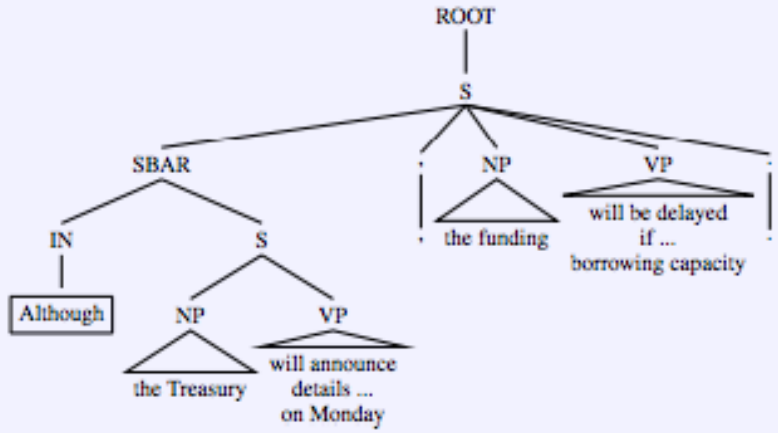
INPUT x : *The exhibition, which opened Oct. 8 and runs through Jan. 3, features 27 self-portraits.* REFERENCE y : *The show started Oct. 8. It ends Jan. 3.*
d_1 : *The exhibition features 27 self-portraits.* d_2 : *The exhibition opened Oct. 8 and runs through Jan. 3.*
d_3 : *The exhibition opened Oct. 8.* d_4 : *The exhibition runs through Jan. 3.* $\hat{v} = v_7$: *The exhibition opened Oct. 8. The exhibition runs through Jan. 3.*

DISSIM

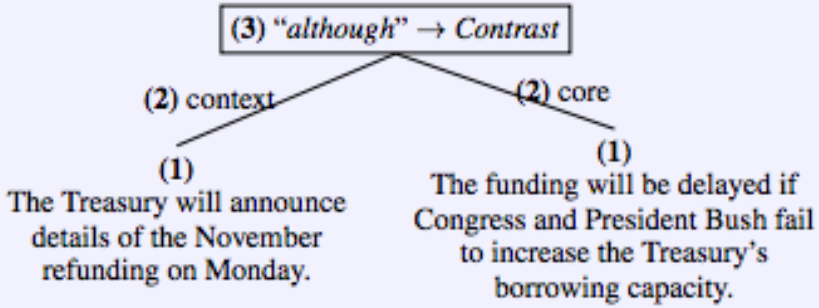
Example: SUBORDINATIONPREEXTRACTOR

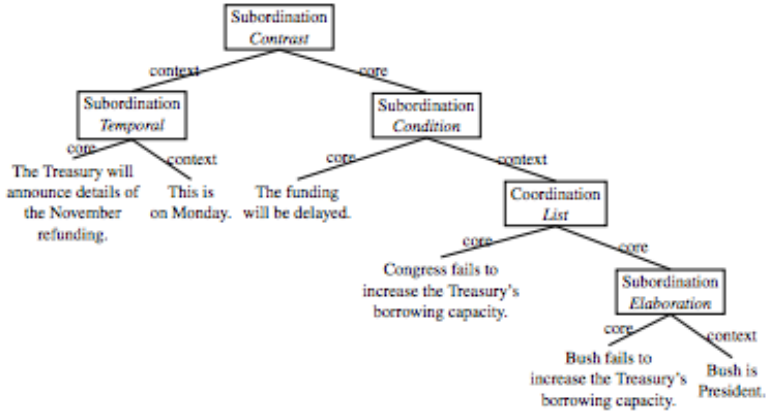
Input: *"Although the Treasury will announce details of the November refunding on Monday, the funding will be delayed if Congress and President Bush fail to increase the Treasury's borrowing capacity."*

Matched Pattern:



Extraction:



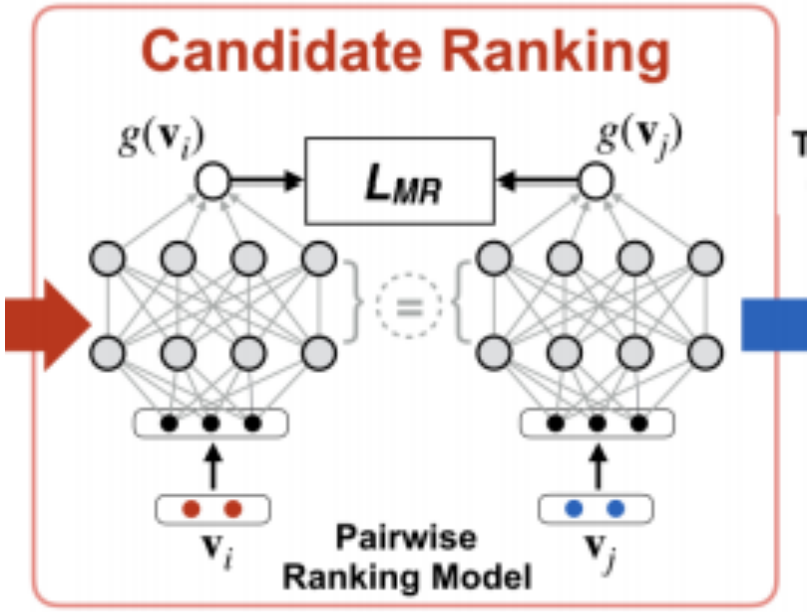


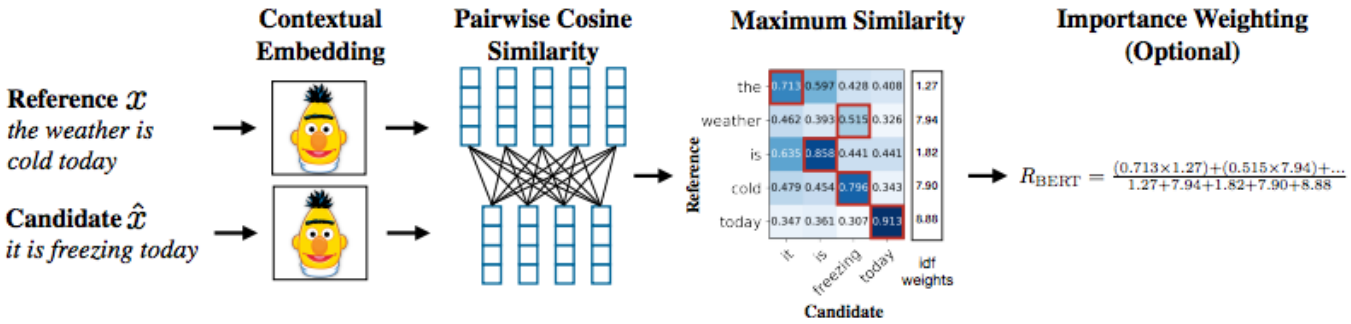
Candidate Ranking

- Neural ranking model to score all candidates and feed top-ranked one to lexical paraphrasing model for final output
- Scoring function – candidate "goodness"

$$g^*(\mathbf{v}_i, \mathbf{y}) = e^{-\lambda||\phi_{\mathbf{v}_i} - \phi_{\mathbf{y}}||} \times BERTScore(\mathbf{v}_i, \mathbf{y}) \quad (1)$$

  - Compression ratio (v_i) - # words in v_i / # words in x





Candidate Ranking

- Pairwise ranking model

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

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

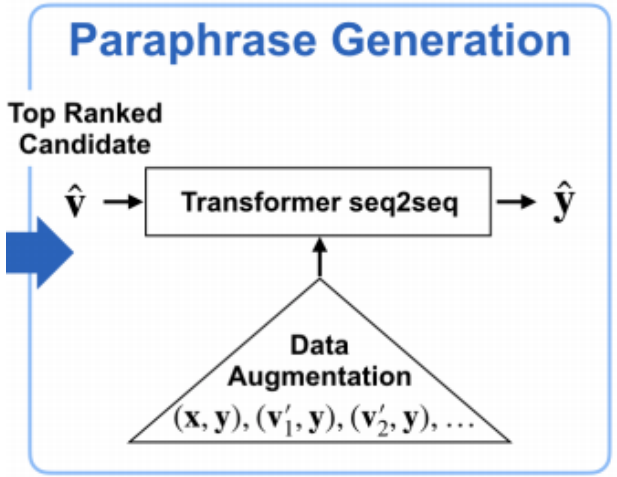
$$l_{ij}^k = sign\left(g^*(\mathbf{v}_i^k, \mathbf{y}^k) - g^*(\mathbf{v}_j^k, \mathbf{y}^k)\right) \quad (2)$$

  - g(.) FFN
    - # words in v_i and x
    - Compression ratio of (v_i,x)
    - Rules applied on x to obtain v_i
    - Number of rule applications
  - Score each candidate v_i separately and rank them in decreasing order of g(v_i)

Paraphrase generation

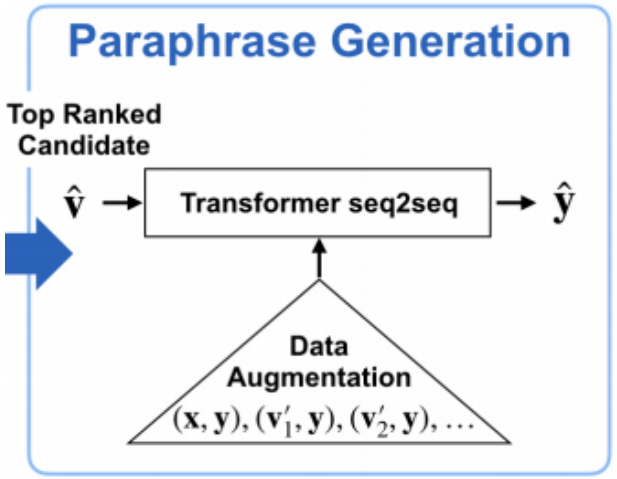
- Paraphrase top-ranked candidate to generate final simplification
- Paraphrase model can explicitly control the extent of lexical paraphrasing by specifying the % of words to be copied from the input sentence as a soft constraint
- Base model = Seq2Seq + Copy (Enc=BERT)
- Copy control
  - cp - % of words to be copied from v_i

$$(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_l) = encoder([cp; \hat{v}_1, \hat{v}_2, \ldots, \hat{v}_l])$$
$$\bar{\mathbf{h}}_i = \mathbf{h_i} + p_i \cdot \mathbf{u},$$
$$\bar{\mathbf{H}} = (\bar{\mathbf{h}}_1, \bar{\mathbf{h}}_2, \ldots, \bar{\mathbf{h}}_l) \quad (3)$$



Paraphrase generation

- Multi-task training
  - Predicting whether a word should be copied
  - Gold labels – each word in the input sentence also appears in the human reference
    - JacanaAlign to determine which occurrence is the one that gets copied
- Data Augmentation
  - Select a subset of candidates using g*(v_i,y)>=0.5 and discard candidates that have different number of split sentences w.r.t reference
- Controllable generation
  - Select candidates with desirable length or # splits during candidate generation step



Experiments

- Newsela, Wikipedia
- Evaluation
  - Standard evaluation
  - Split-focused subset
  - Compression ratio < 0.7 and no splits
  - Lexical paraphrasing
- Methods
  - Transformer Seq2Seq (Enc=BERT)
  - EditNTS (neural programmer-interpreter)
  - LSTM Seq2Seq
  - Hybrid-NG (Splitting/deleting using probabilistic model and lexical substitution using MT)
- Metrics
  - SARI (Averages F1/precision of n-grams inserted, deleted, kept)
    - F1-score for n-grams that are added (add)
    - F1-score for n-grams that are kept (keep), Precision for those deleted (del)
    - BLEU (s-BL)

Results

	System Outputs
Complex	*Since 2010, project researchers have uncovered documents in Portugal that have revealed who owned the ship.*
Simple	*Since 2010, experts have been figuring out who owned the ship.*
Hybrid-NG	*since 2010, the project **scientists** have uncovered documents in portugal that have **about** who **owns** the ship.*
LSTM	*since 2010, **scientists** have uncovered documents in portugal that have revealed who owned the ship.*
Transformer$_{bert}$	***they discovered** that the ship **had been important**.*
EditNTS	*since 2010, project researchers have uncovered documents in portugal**. have** revealed who owned the ship.*
Our Model ($cp = 0.6$)	***scientists** have **found** a **secret deal**. **they have discovered** who owned the ship.*
Our Model ($cp = 0.7$)	***scientists** have **found** documents in portugal. **they have also found out** who owned the ship.*
Our Model ($cp = 0.8$)	***scientists** have **found** documents in portugal. **they have discovered** who owned the ship.*
Complex	*Experts say China's air pollution exacts a tremendous toll on human health.*
Simple	*China's air pollution is very unhealthy.*
Hybrid-NG	*experts say the **government's** air pollution exacts a toll on human health.*
LSTM	*experts say china's air pollution exacts a tremendous toll on human health.*
Transformer$_{bert}$	*experts say china's pollution has a tremendous **effect** on human health.*
EditNTS	*experts say china's air pollution **can cause** human health.*
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.*

Table 7: Examples of system outputs. **Red** marks the errors; **blue** marks good paraphrases. *cp* is a soft constraint that denotes the percentage of words that can be copied from the input.

Results

Models	SARI	add	keep	del	FK	SLen	OLen	CR	%split	s-BL	%new	%eq
Complex (input)	15.9	0.0	47.6	0.0	12.0	23.7	23.8	1.0	0.0	100.0	0.0	100.0
Simple (reference)	90.5	86.8	86.6	98.2	7.4	14.4	19.0	0.83	28.0	35.5	33.0	0.0
LSTM	35.0	1.6	**45.5**	57.8	8.9	17.6	17.9	0.8	1.9	66.5	5.0	20.2
Hybrid-NG	35.8	1.9	41.8	63.7	9.9	21.2	23.7	1.0	11.6	59.7	8.8	5.1
Transformer$_{bert}$	37.0	3.1	43.6	64.4	8.1	**15.6**	20.2	0.87	**24.1**	58.8	12.8	10.2
EditNTS	37.9	1.3	45.2	67.1	8.8	16.7	20.3	0.88	20.3	68.7	5.2	2.6
Our Model	**38.7**	**3.3**	42.9	**70.0**	**7.9**	15.8	**20.1**	**0.86**	23.9	**48.7**	**16.2**	**0.4**

Table 2: Automatic evaluation results on NEWSELA-AUTO test set. We report **SARI, the main automatic metric** for simplification, and its three edit scores namely precision for delete (**del**) and F1 scores for **add** and **keep** operations. We also report FKGL (**FK**), average sentence length (**SLen**), output length (**OLen**), compression ratio (**CR**), self-BLEU (**s-BL**), percentage of sentence splits (**%split**), average percentage of new words added to the output (**%new**), and percentage of sentences identical to the input (**%eq**). **Bold** typeface denotes the best performances (i.e., closest to the reference).

- Best w.r.t SARI
- LSTM, EditNTS focus on deletion (high self-BLEU, similar FK)
- Transformer model is conservative (copies 10.2% of sentences)
- Lowest self-BLEU, FK, %eq, least conservative, good paraphrases, mimics human references better