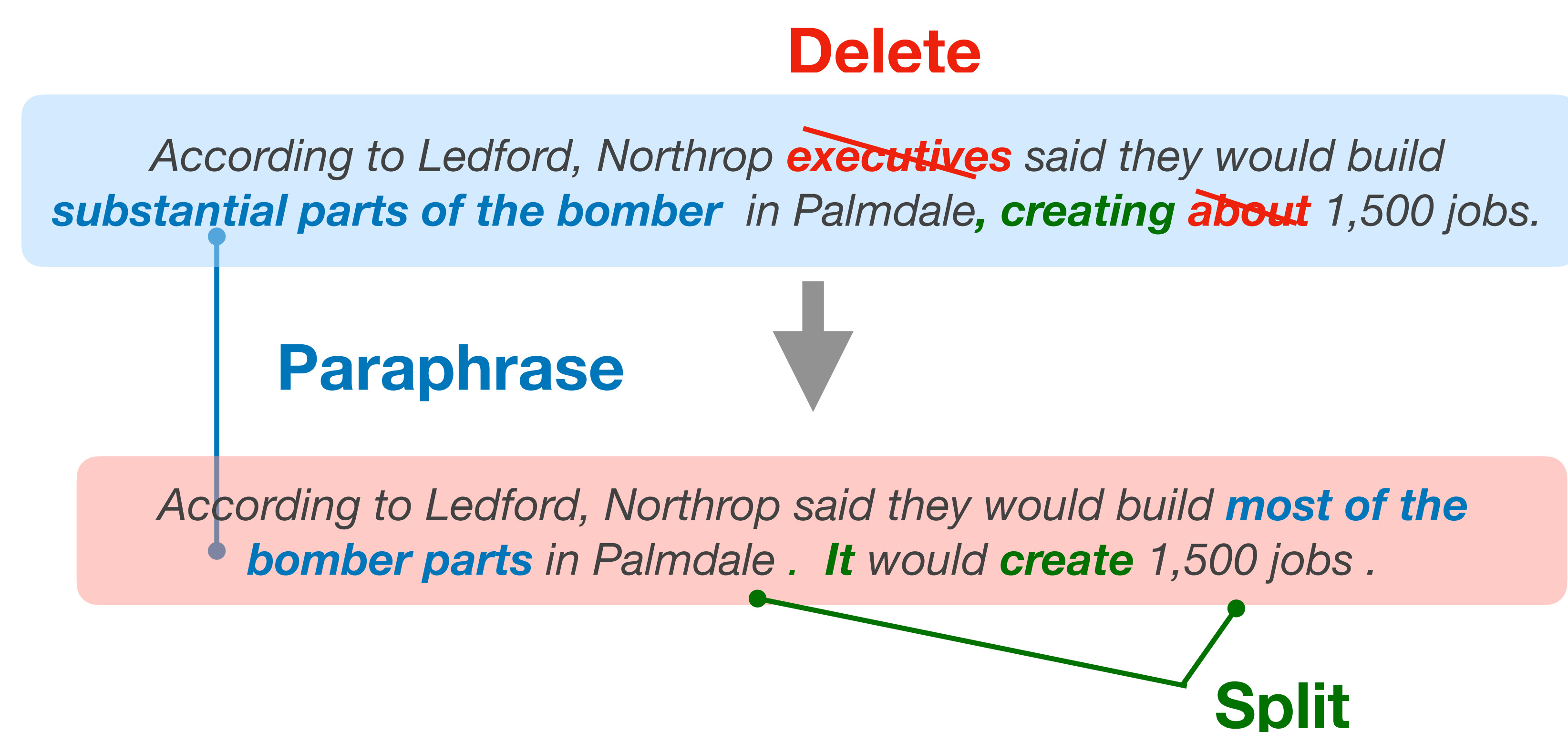


1. Automatic Text Simplification

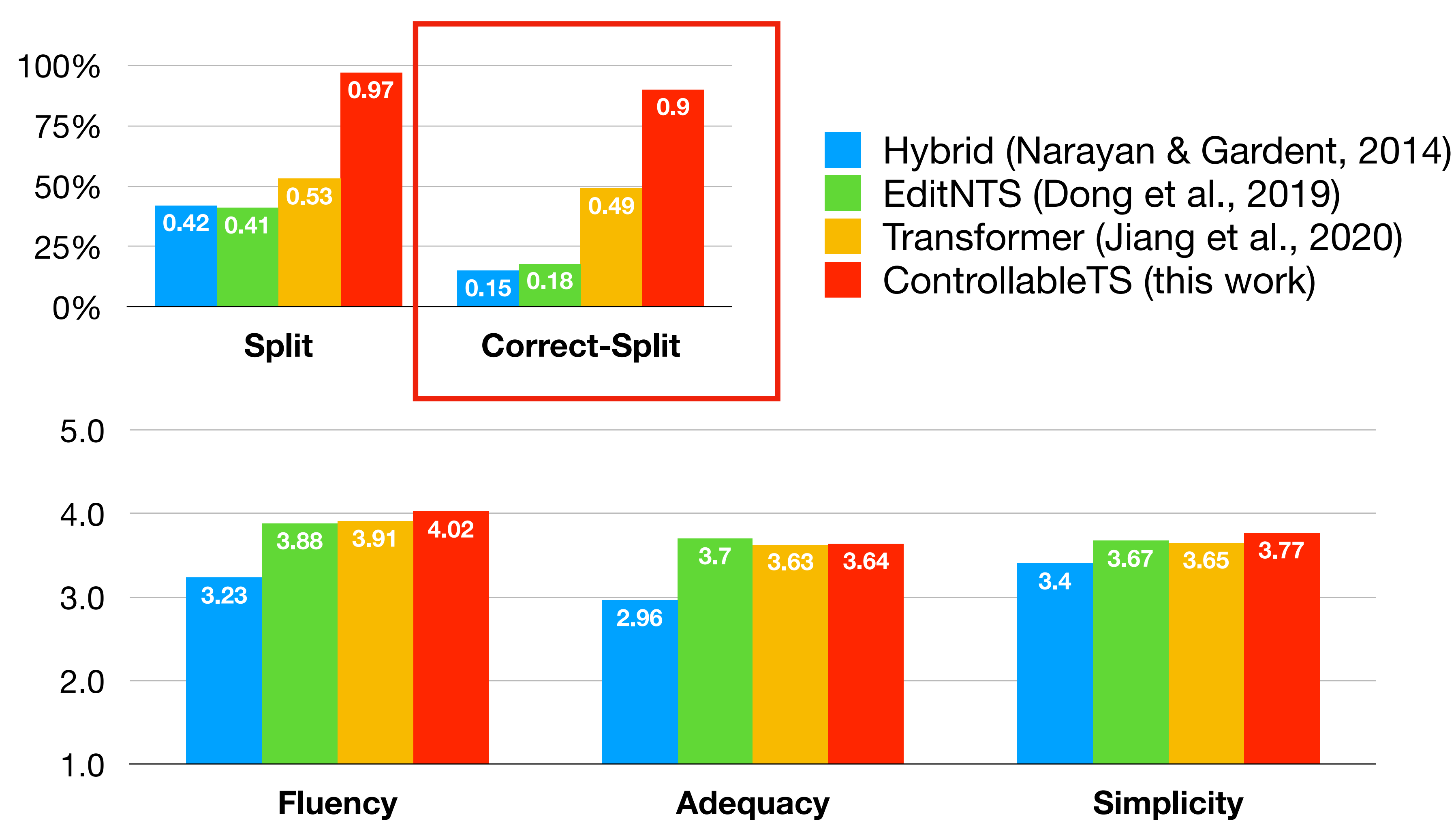
- ▶ Rewrite complex text into simpler language while retaining the original meaning.
- ▶ Treated as a **text-to-text generation** task.
- ▶ Involves **3 edit operations** - delete, split, and paraphrase.



- ▶ Problems with SOTA simplification systems:
 - ▶ **Lack controllability** in terms of edit operations.
 - ▶ Perform only **deletion**.

3. Experiments

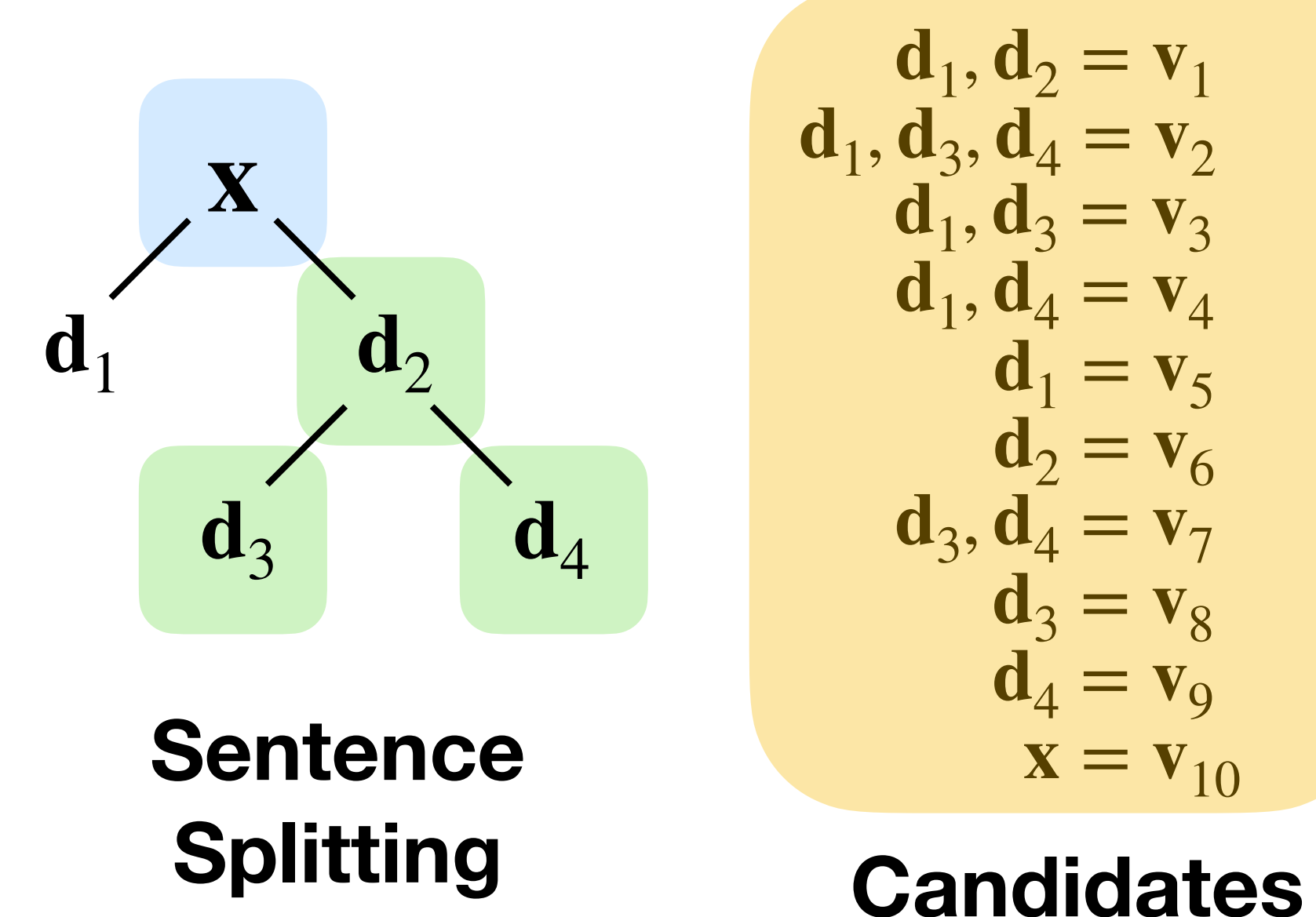
- ▶ Trained on **Newsela-Auto** (Jiang et al. 2020).
- ▶ 259,778 train + 32,689 valid + 33,391 test pairs.
- ▶ Human evaluation on 100 random simplifications.



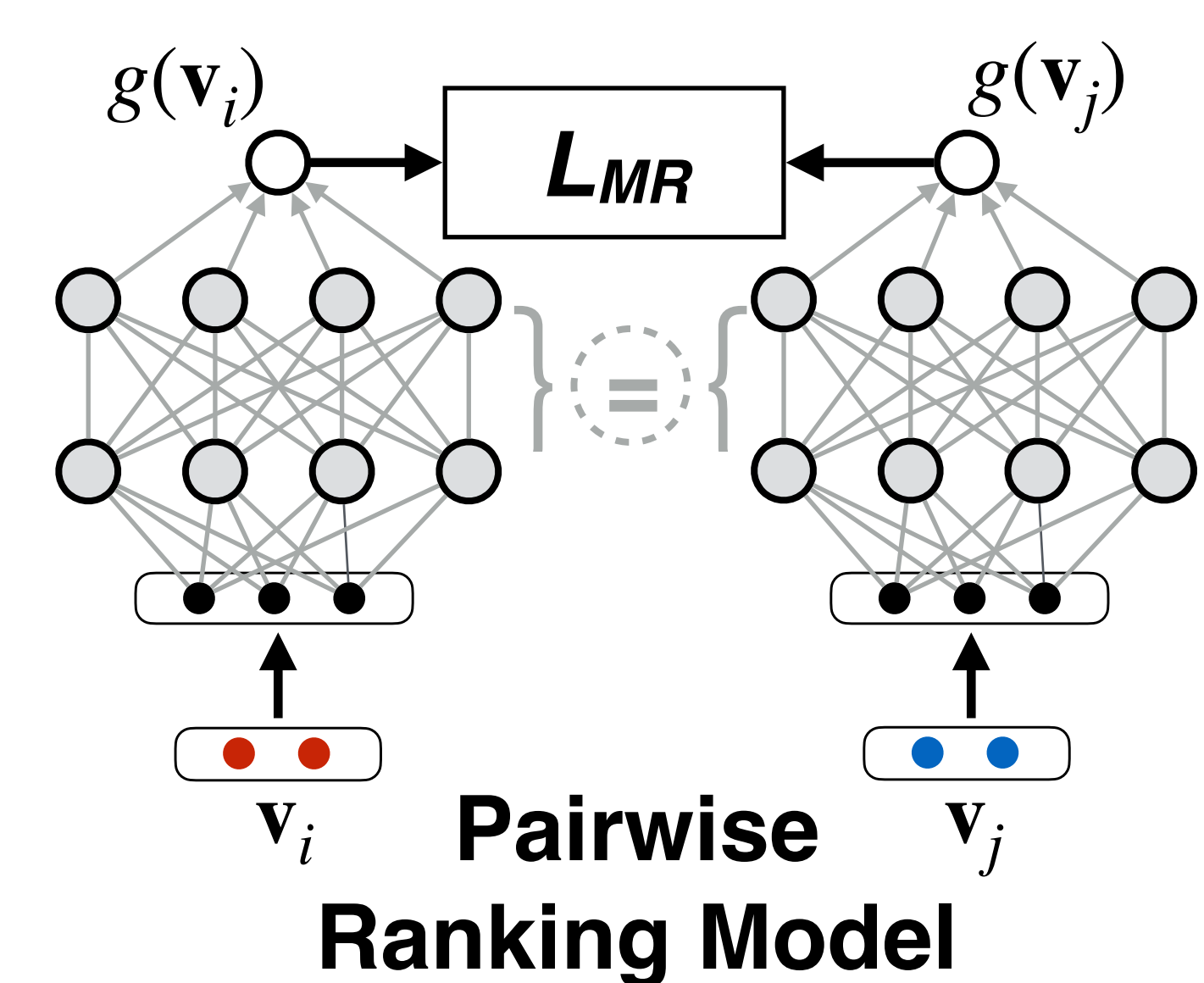
- ▶ Our model outperforms the SOTA systems, especially when splitting is involved.

2. Our Approach: Controllable Text Generation

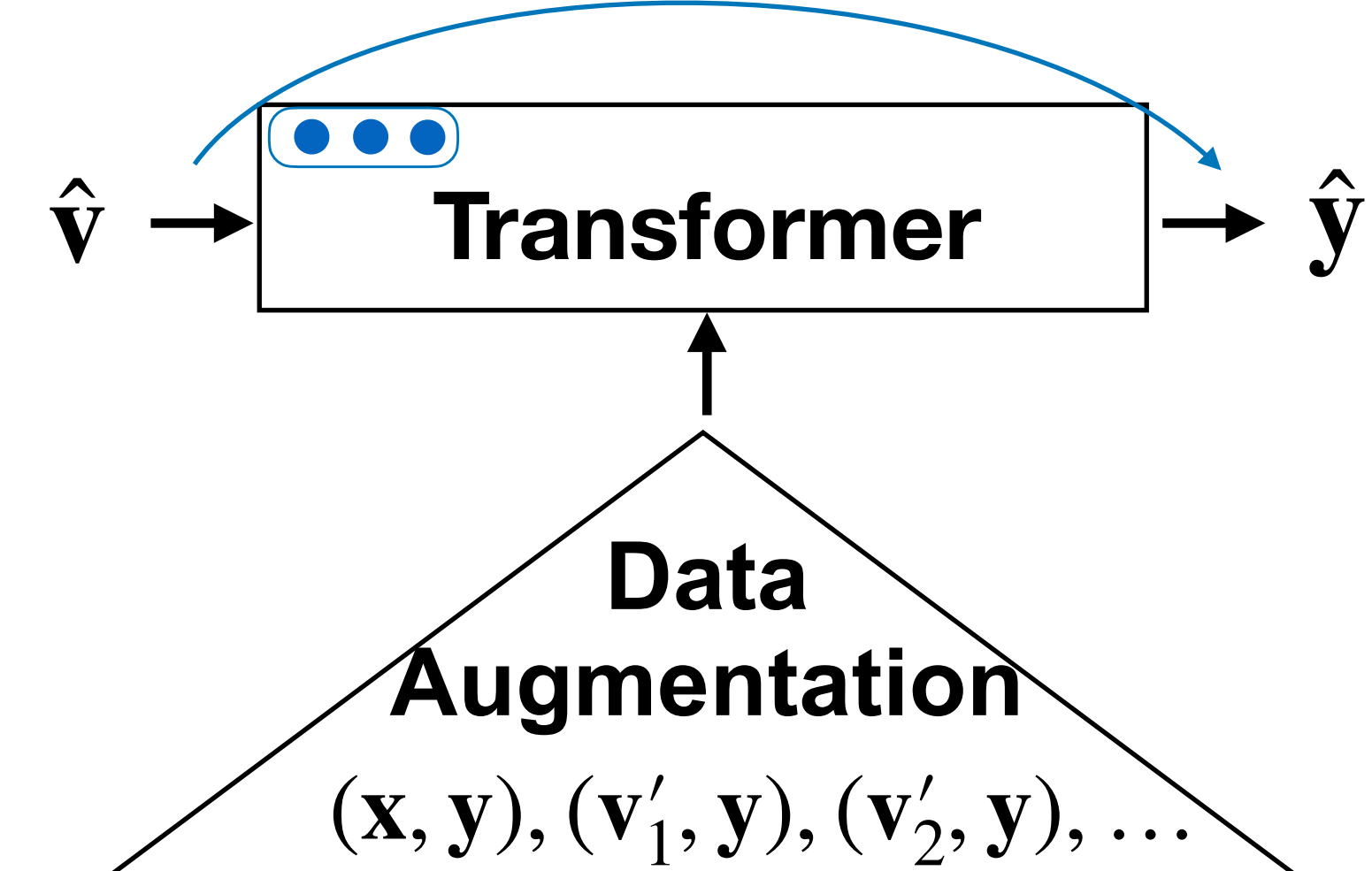
Candidate Generation



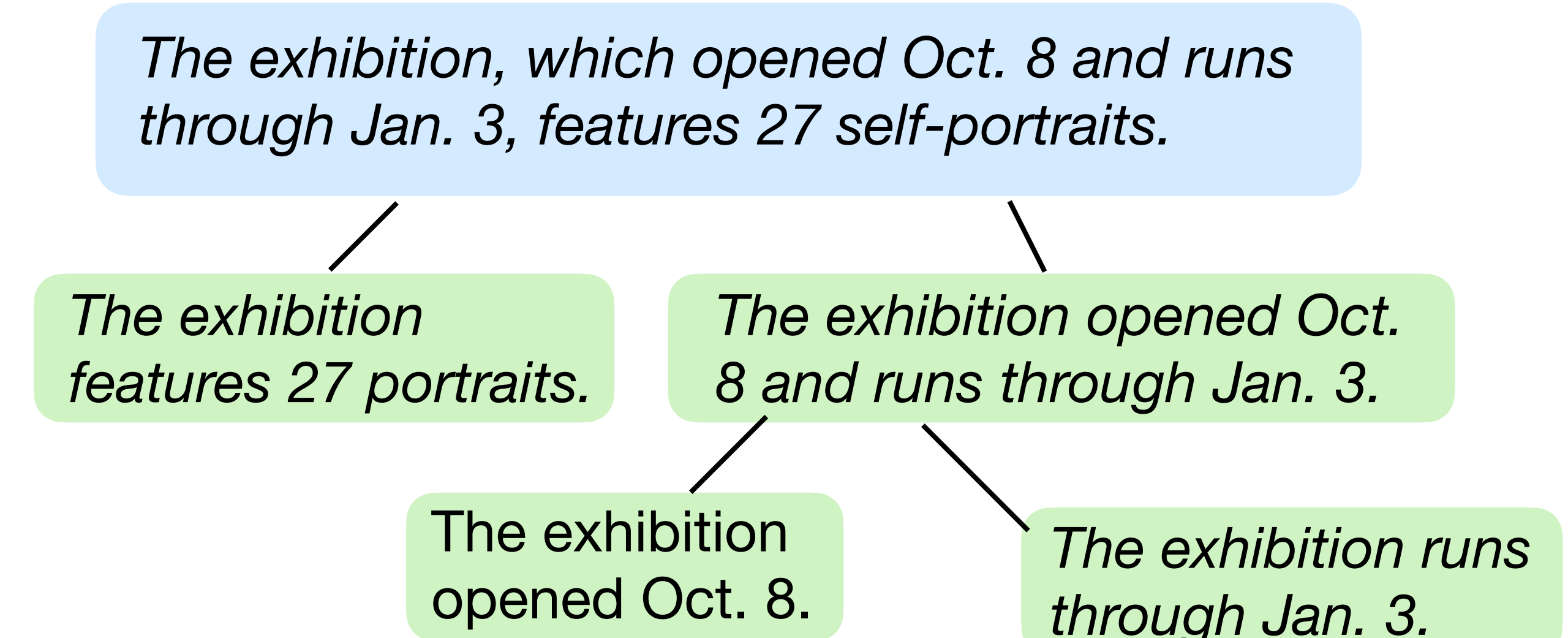
Candidate Ranking



Paraphrase Generation



Split the input sentence using 35 linguistic rules (Niklaus et al., 2019) + a seq2seq model.



Combine splits as 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. .. (and more)

Rank all intermediate candidates.

Gold Scoring Function:

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

target compression ratio candidate reference

Loss function:

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

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

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

Ranker score Length-penalized BERTScore

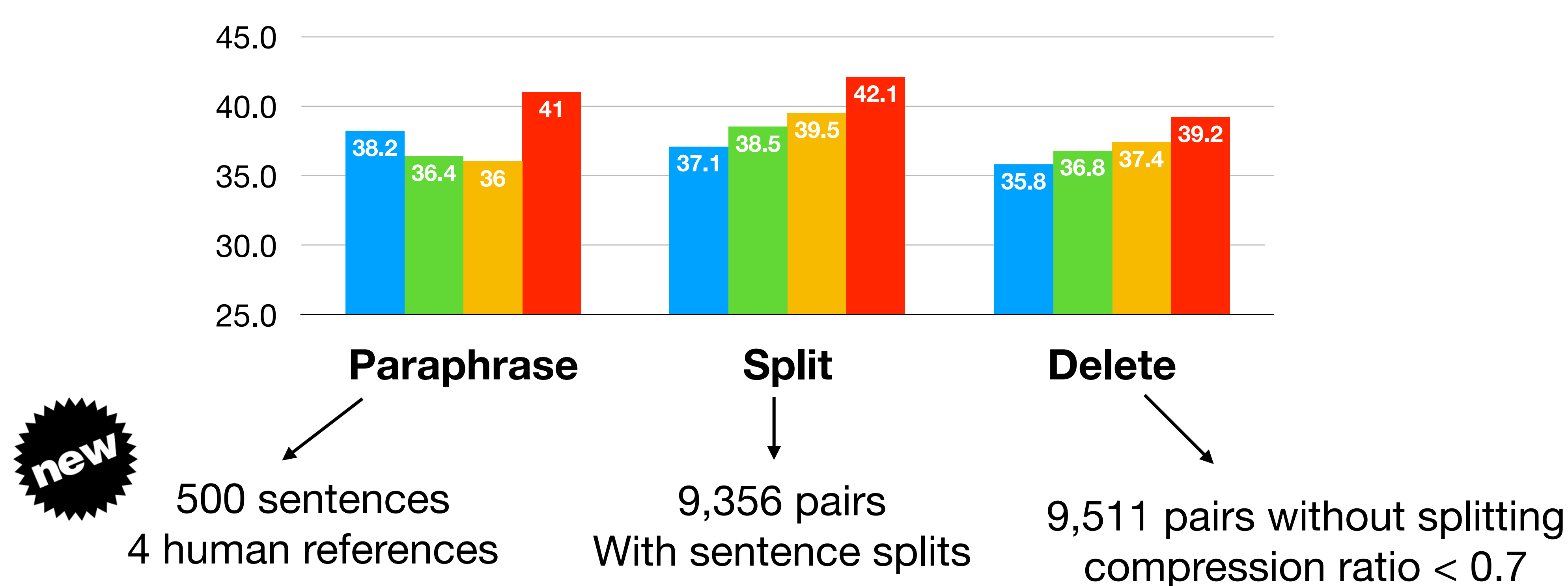
- ▶ Paraphrase **top-ranked** candidate.
- ▶ **Data Augmentation** with additional training data that focuses on lexical paraphrasing.
- ▶ **Copy-control** token as a soft constraint to control paraphrasing.
- ▶ **Auxiliary task** (if a word should be copied).

Controllability of edit operations:

- ▶ Split / Delete **Candidate Selection**
- ▶ Paraphrase **Copy-control token**

4. Edit-focused Evaluation

- ▶ Evaluation on sections of Newsela-Auto test set.
- ▶ We report **SARI**, the main automatic metric of simplification.



5. Takeaways

- ▶ **Control over 3 types** of edit operations.
- ▶ Inject **linguistic knowledge** into neural models.
- ▶ **New dataset** to evaluate lexical paraphrasing.
- ▶ **Evaluation setup** for edit operations.

Code/data at https://github.com/mounicam/controllable_simplification

