

WikiSplit++: Easy Data Refinement for Split and Rephrase

Hayato Tsukagoshi[◇], Tsutomu Hirao[♣], Makoto Morishita[♣]
Katsuki Chousa[♣], Ryohei Sasano[◇], Koichi Takeda[◇]

[◇] Graduate School of Informatics, Nagoya University

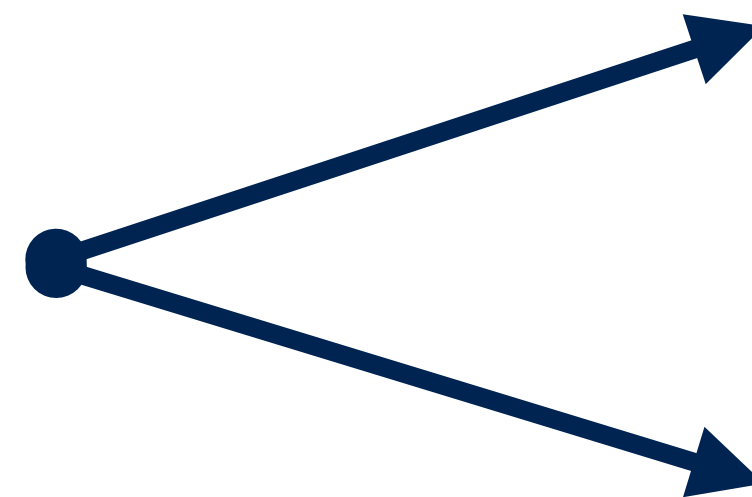
[♣] NTT Communication Science Laboratories

Split and Rephrase

- Breaks down a complicated sentence into shorter, simpler ones
 - Simple sentences maintain the meaning of the original complex sentence

Complex Sentence

**He lives in Brooklyn,
New York and is
married to Melissa Block.**



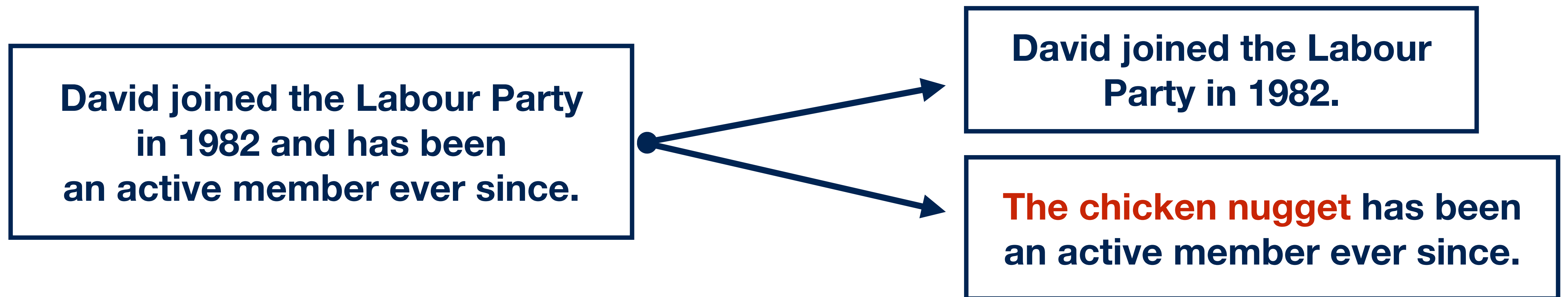
Simple Sentences

**He lives in Brooklyn,
New York.**

**He is married to
Melissa Block.**

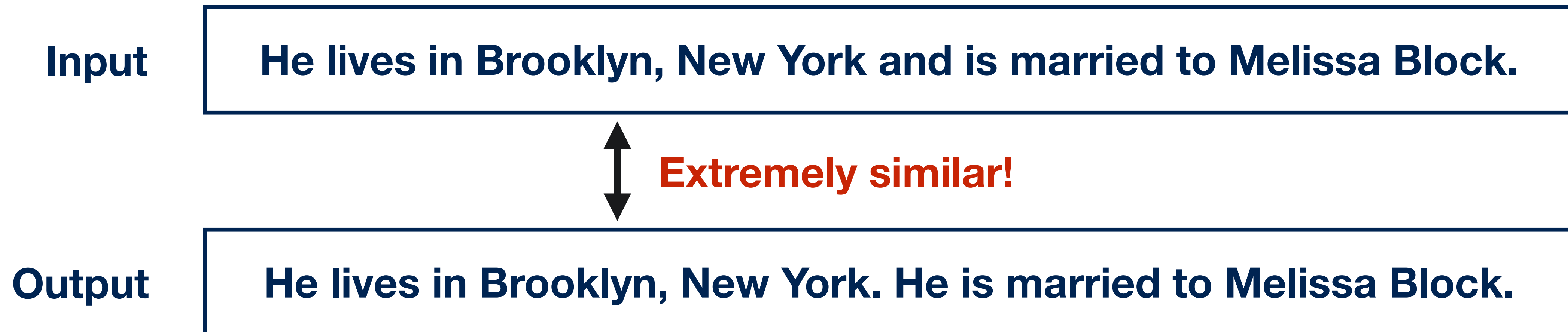
Problems of Split and Rephrase: Data Quality

- Existing studies use automatically generated large-scale datasets.
 - WikiSplit: created from the edit history of Wikipedia
- However, these datasets contain a large number of instances that are not correct complex-simple sentence pairs.
- Noisy training dataset negatively affect model training and can cause **hallucinations**.



Problems of Split and Rephrase: Training Difficulties

- The input and output in the Split and Rephrase task are very similar
 - Training with a standard Seq2Seq model may result in low loss even if the model learns to simply reproduce the input as output
 - The training loss might decrease, but it renders the model meaningless



WikiSplit++: Easy Data Refinement for Split and Rephrase

- In this study, we improve Split and Rephrase models w/o model modifications
- Easy data refinement can improve performance and reduce hallucinations
 - **Filtering by Natural Language Inference Classifier**
 - **Sentence-Order Reversing**
- Refined WikiSplit: **WikiSplit++**

- WikiSplit++ demonstrates superior performance to WikiSplit while reducing the data by about 37%.
- WikiSplit++ has been shown to reduce hallucinations in the generated sentences compared to models trained on WikiSplit and other existing models.

Preliminaries: Natural Language Inference

NLI (Natural Language Inference)

- Classify the relationship between the given premise and hypothesis
- Representative datasets: Stanford NLI (SNLI) , Multi-Genre NLI (MNLI)



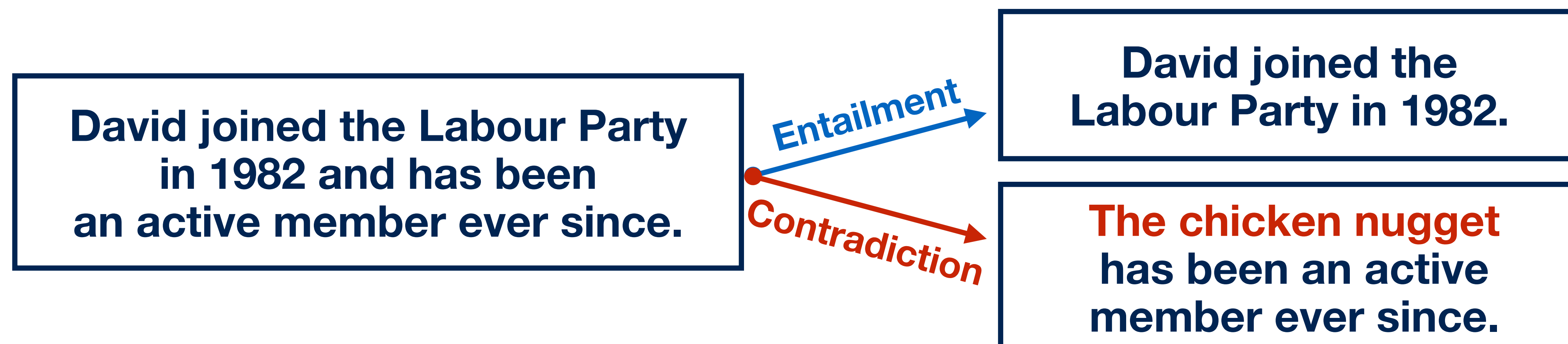
Refinement: Natural Language Inference Filtering

In The Split and Rephrase task...

- A complex sentence should entail each simple sentence
- Simple sentences not implied by complex sentences can be noise

NLI Filtering

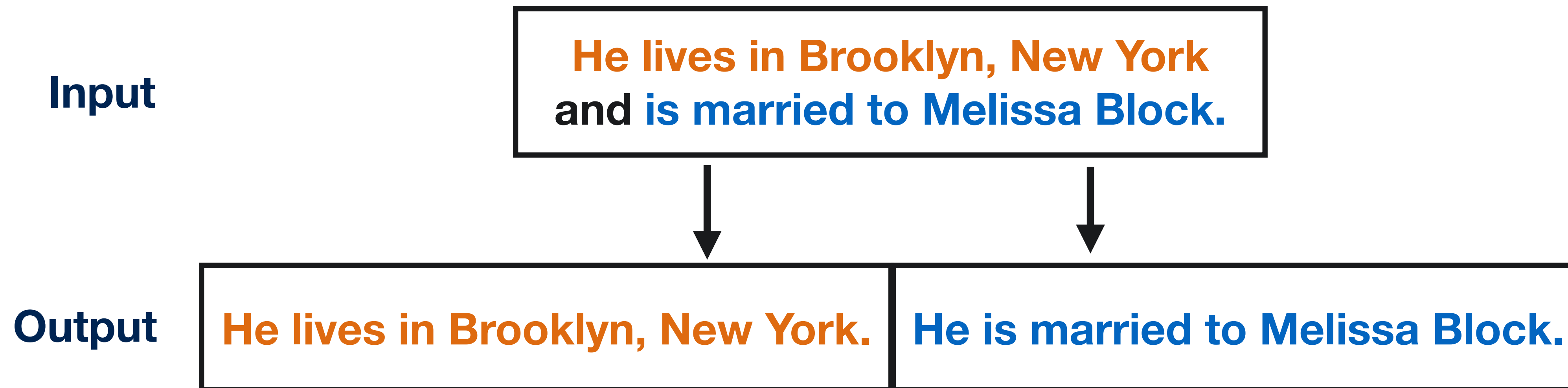
- We utilize NLI classification models to remove inappropriate examples



Refinement: Sentence-Order Reversing

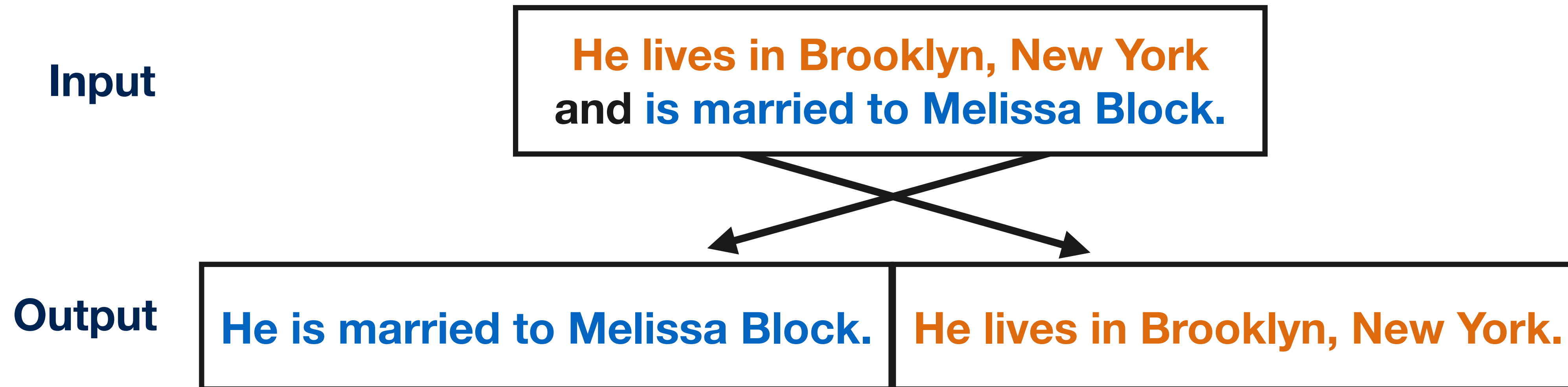
In The Split and Rephrase task...

- The input and output are very similar
 - Resulting in a trained model that is unable to split sentences properly



Refinement: Sentence-Order Reversing

- To solve this, **reverse the order of reference sentences** in advance
- Reversed simple sentences do not resemble the input complex sentence
- Additionally, it is difficult to reduce the loss without learning to split the sentences at appropriate places.



Experiments

- Training dataset
- Baseline methods
- Evaluation datasets
- Evaluation metrics

Experiments: Training Dataset



Base Dataset: WikiSplit

- The only large dataset built upon sentences written by human
- Apply NLI filtering to WikiSplit
 - About 36.6% are removed
- sentence-order reversing dose not reduce the number of examples
- Use PySBD for sentence splitting

NLI Classifier for Filtering

- DeBERTa-xxl fine-tuned with MNLI

	WikiSplit	WikiSplit++
overall	994,481	630,433
train	795,585	504,375
dev	99,448	63,065
test	99,448	62,993

Experiments: Baseline Methods

DisSim

- A rule-based discourse-aware sentence-splitting framework

BiSECT Model

- SotA Split and Rephrase model based on BERT-Initialized Transformer

GPT-3 (0/3-shot)

- few-shot learning with large language models

Experiments: Evaluation Details

Metrics

- BLUE, BERTScore, BLEURT, SARI, FKGL
- **Entailment Ratio**
 - The ratio of complex sentences that entail corresponding simple sentences
- #Sent., Copy

Datasets

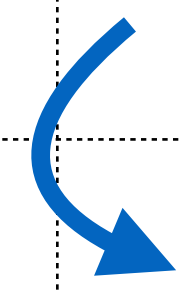
- HSplit: consists of Wikipedia sentences
- WikiBM: consists of more difficult Wikipedia sentences
- ContBM: consists of publicly available legal procurement contracts

Experimental Results: HSplit



- T5 fine-tuned with WikiSplit++ overall outperformed the baseline methods
 - Particularly in the **entailment ratio**

System	Train Dataset	BLEU	BERTScore	BLEURT	SARI	Entailment	FKGL	#Sent.	Copy
DisSim		63.71	94.93	75.54	66.74	92.20	7.85	2.96	21.73
BiSECT Model	BiSECT +WikiSplit	86.98	96.54	81.35	57.65	95.26	8.58	1.98	2.23
GPT-3 zero-shot		54.39	94.25	76.79	74.34	94.15	8.90	2.14	5.29
T5-small	WikiSplit	87.95	96.65	82.06	57.17	95.49	8.63	1.98	2.48
T5-small	WikiSplit++	88.06	96.57	81.71	56.79	98.02	8.59	2.00	0.72



Ablation Study: NLI Filtering & Sentence-Order Reversing

- Investigate the impact of NLI filtering and sentence-order reversing for each
- **NLI only**: improve entailment ratio, worsening the num. of splits
- **Rev. only**: slightly improve entailment ratio and decrease the num of copying

NLI	Rev.	BLEU	BERTScore	BLEURT	SARI	Entailment	FKGL	#Sent.	Copy
		87.95	96.65	82.06	57.17	95.49	8.63	1.98	2.48
✓		88.81	96.80	82.79	57.23	97.74	8.71	1.93	7.35
	✓	88.21	96.58	81.68	56.68	96.74	8.57	2.00	0.33
✓	✓	88.06	96.57	81.71	56.79	98.02	8.59	2.00	0.72

Discussion: Data Refinement for Other Datasets

- Apply our proposed refinement method to MinWikiSplit and BiSECT
 - MinWikiSplit: dataset created by applying DisSim to WikiSplit
 - BiSECT: translated 1-to-2 sentences from parallel corpora
- Consistent improvements in the entailment ratio

Dataset	BLEU	BERTScore	BLEURT	SARI	Entailment	FKGL	#Sent.	Copy
WikiSplit	87.95	96.65	82.06	57.17	95.49	8.63	1.98	2.48
WikiSplit++	88.06	96.57	81.71	56.79	98.02	8.59	2.00	0.72
MinWikiSplit	77.98	95.77	76.38	65.45	83.54	8.47	2.11	25.88
MinWikiSplit++	77.67	95.71	76.87	65.93	90.11	8.49	2.13	27.24
BiSECT	73.57	96.10	79.13	67.41	90.72	8.73	1.98	2.79
BiSECT++	73.54	96.01	79.44	68.13	95.88	8.57	1.99	1.20

Discussion: Impact of NLI Classifier

- Investigating potential biases introduced by the NLI classifiers
- Obtained consistent performance improvements regardless of the classifier

Classifier	BLEU	BERTScore	BLEURT	SARI	Entailment	FKGL	#Sent.	Copy
	87.95	96.65	82.06	57.17	95.49	8.63	1.98	2.48
DeBERTa	88.06	96.57	81.71	56.79	98.02	8.59	2.00	0.72
RoBERTa	88.08	96.57	81.74	56.73	98.25	8.60	2.00	0.67
TRUE	88.20	96.58	81.79	56.73	98.25	8.61	2.00	0.45

Conclusion

- Proposed easy data refinement methods
 - **NLI Filtering**
 - **Sentence-Order Reversing**
- Proposed WikiSplit++ by applying the data refinement to WikiSplit
- WikiSplit++ improved overall model performance & reduced hallucinations

Future work

- Enhancing the diversity of examples in datasets
- Examination of methods to further improve quality