Fact-based Text Editing

Hayate Iso, Chao Qiao, Hang Li

NAIST ByteDance

The status quo of Text Editing

- **Model, *p*(y | x), learns how to edit the input, *x* into the desired output, *y*.**

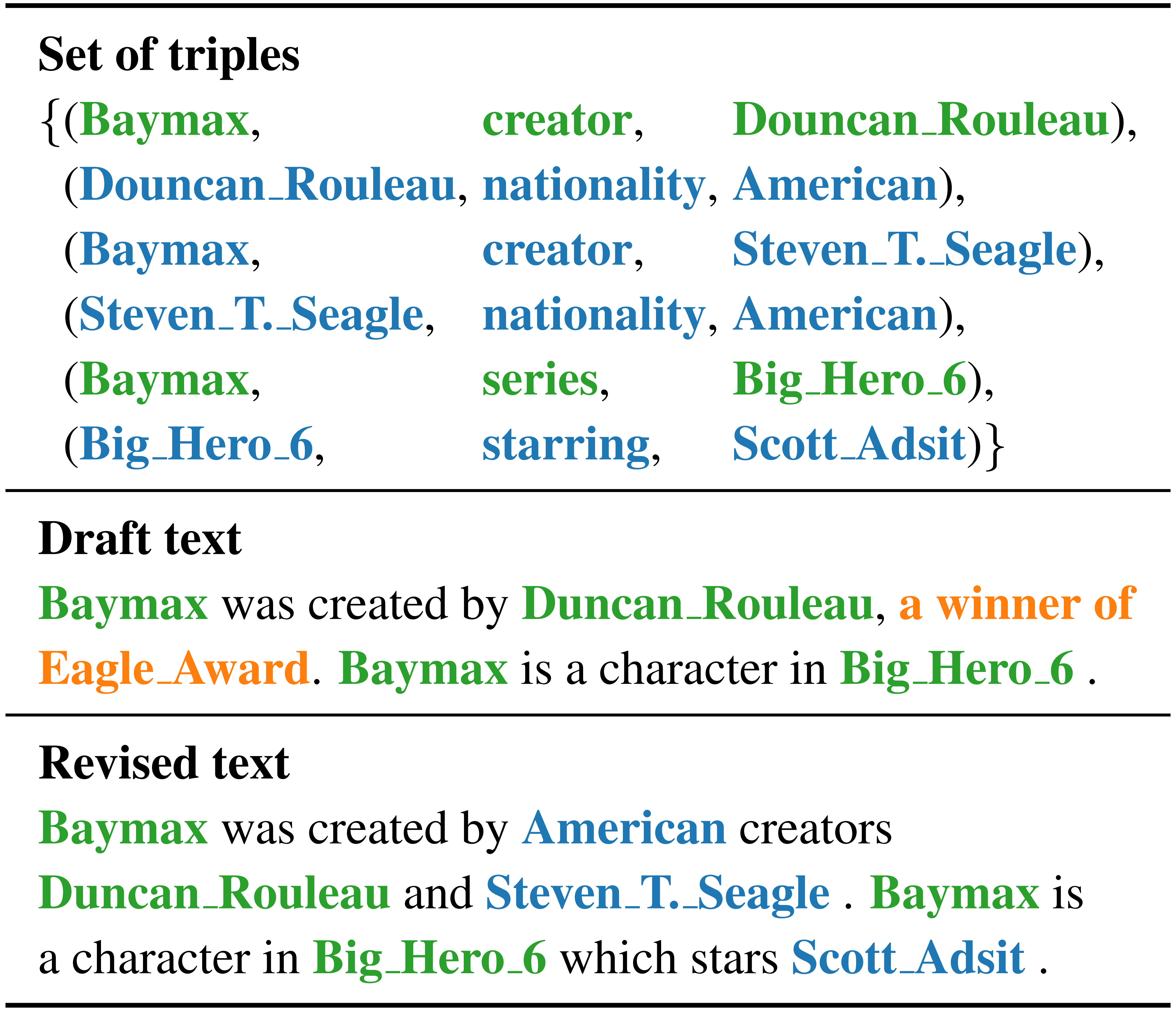
x = “This is the worst game!” $\xrightarrow{\text{Style Transfer}}$ *y* = “This is the best game!”

x = “Last year, I read the book that is authored by Jane” $\xrightarrow{\text{Simplification}}$ *y* = “Jane wrote a book. I read it last year”

x = “Fish firming uses the lots of specials” $\xrightarrow{\text{Grammatical Error Correction}}$ *y* = “Fish firming uses a lot of specials”

What is Fact-based Text Editing?

- The goal of *fact-based text editing* is to ***revise*** a given document to better describe the facts in a knowledge base.
    - e.g., several triples



Set of triples

{(**Baymax**, **creator**, **Douncan_Rouleau**),
(**Douncan_Rouleau**, **nationality**, **American**),
(**Baymax**, **creator**, **Steven_T._Seagle**),
(**Steven_T._Seagle**, **nationality**, **American**),
(**Baymax**, **series**, **Big_Hero_6**),
(**Big_Hero_6**, **starring**, **Scott_Adsit**)}

Draft text

Baymax was created by **Duncan_Rouleau**, **a winner of Eagle_Award**. **Baymax** is a character in **Big_Hero_6** .

Revised text

Baymax was created by **American** creators **Duncan_Rouleau** and **Steven_T._Seagle** . **Baymax** is a character in **Big_Hero_6** which stars **Scott_Adsit** .

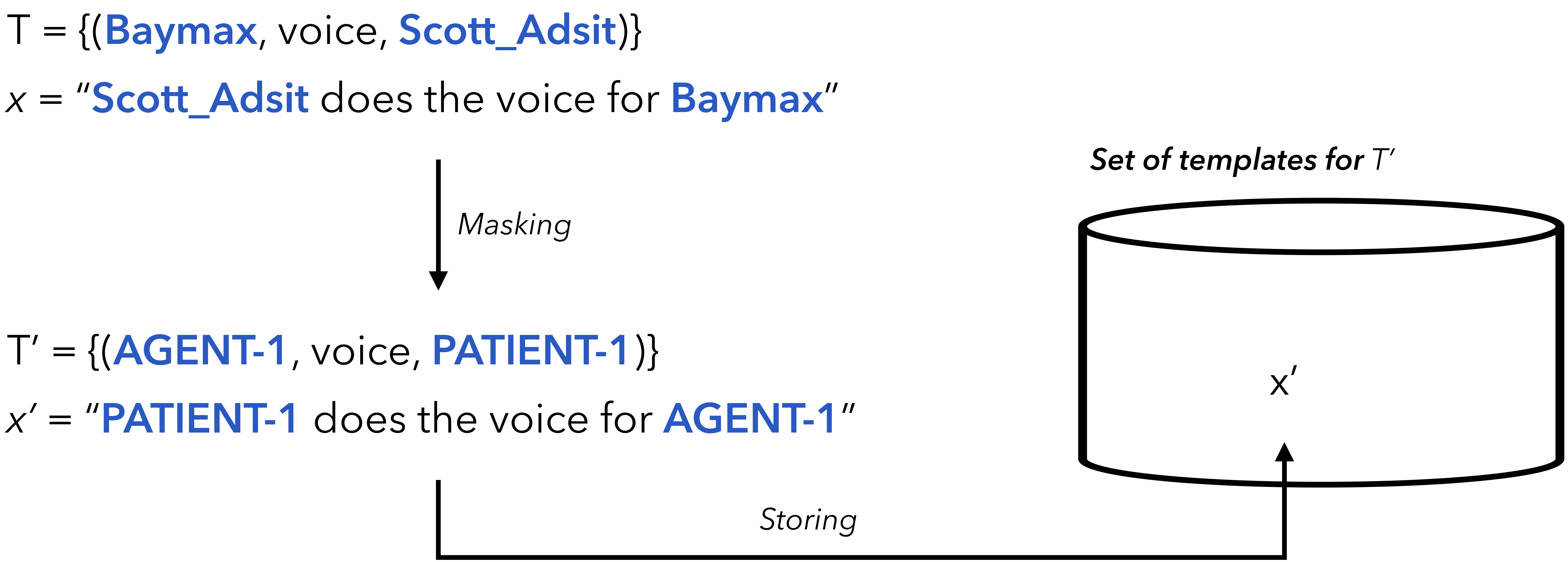
Overview of this research

- **Data Creation:**
  - We have proposed a data construction method for fact-based text editing and created two datasets.

- **Fact-based Text Editing model:**
  - We have proposed a model for fact-based text editing, which performs the task by generating a sequence of actions, instead of words.

Data Creation:Factual Masking

- For all of table-to-text pairs in the training data, we create the template by factual masking.



T = {(**Baymax**, voice, **Scott_Adsit**)}

x = “**Scott_Adsit** does the voice for **Baymax**”

Masking

T′ = {(**AGENT-1**, voice, **PATIENT-1**)}

x′ = “**PATIENT-1** does the voice for **AGENT-1**”

Storing

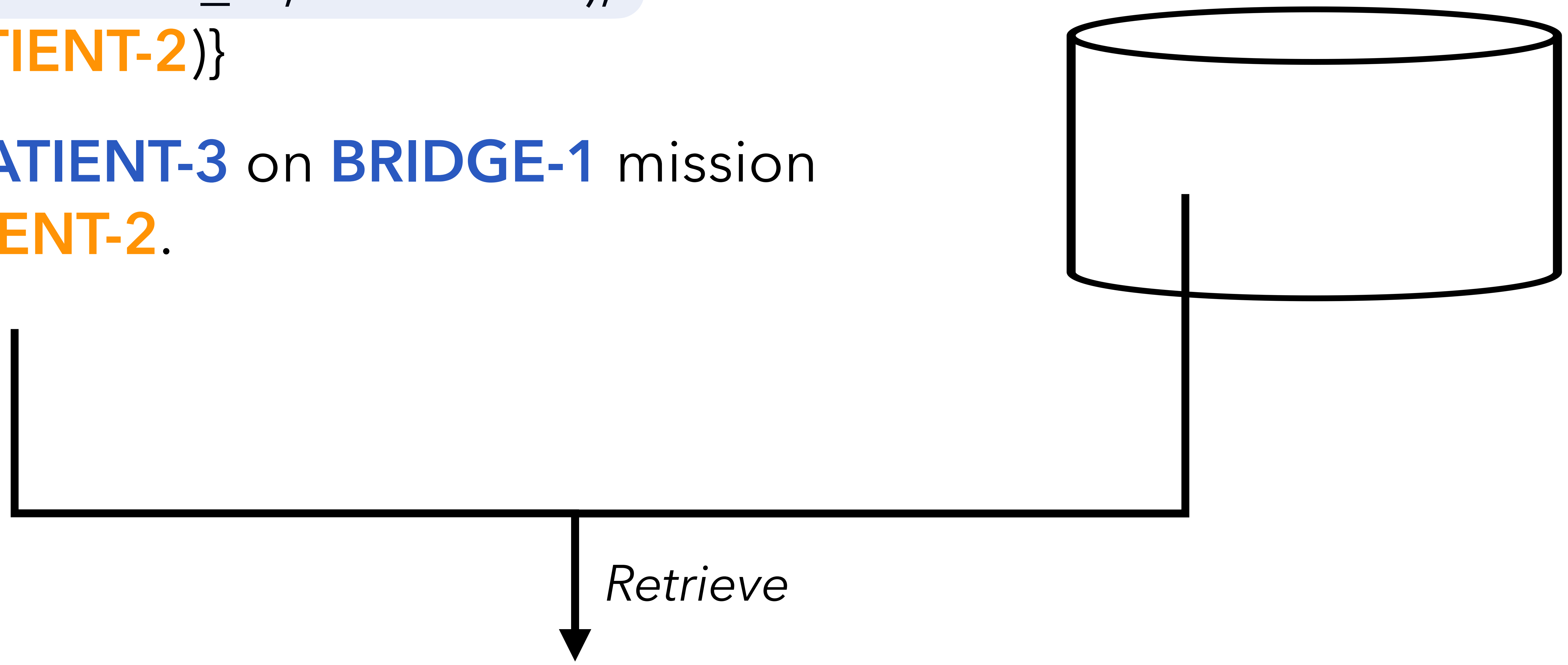
***Set of templates for** T′*

Data Creation: Retrieve LCS matched template

T' = {(AGENT-1, occupation, PATIENT-3),
(AGENT-1, was_a_crew_member_of, BRIDGE-1),
(BRIDGE-1, operator, PATIENT-2)}

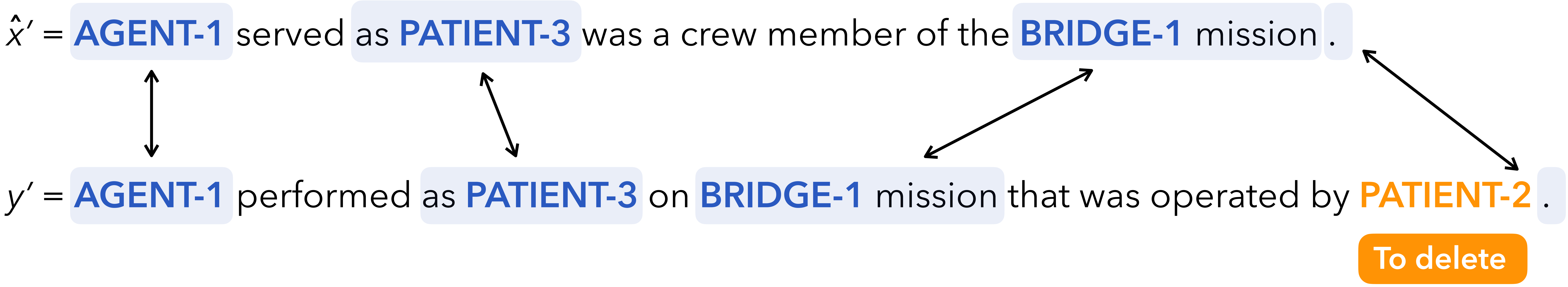
y' = AGENT-1 performed as PATIENT-3 on BRIDGE-1 mission that was operated by PATIENT-2.

Set of templates for
{(AGENT-1, occupation, PATIENT-3),
(AGENT-1, was_a_crew_member_of, BRIDGE-1)}



$\hat{x}'$ = AGENT-1 served as PATIENT-3 was a crew member of the BRIDGE-1 mission.

Data Creation: Token Alignment

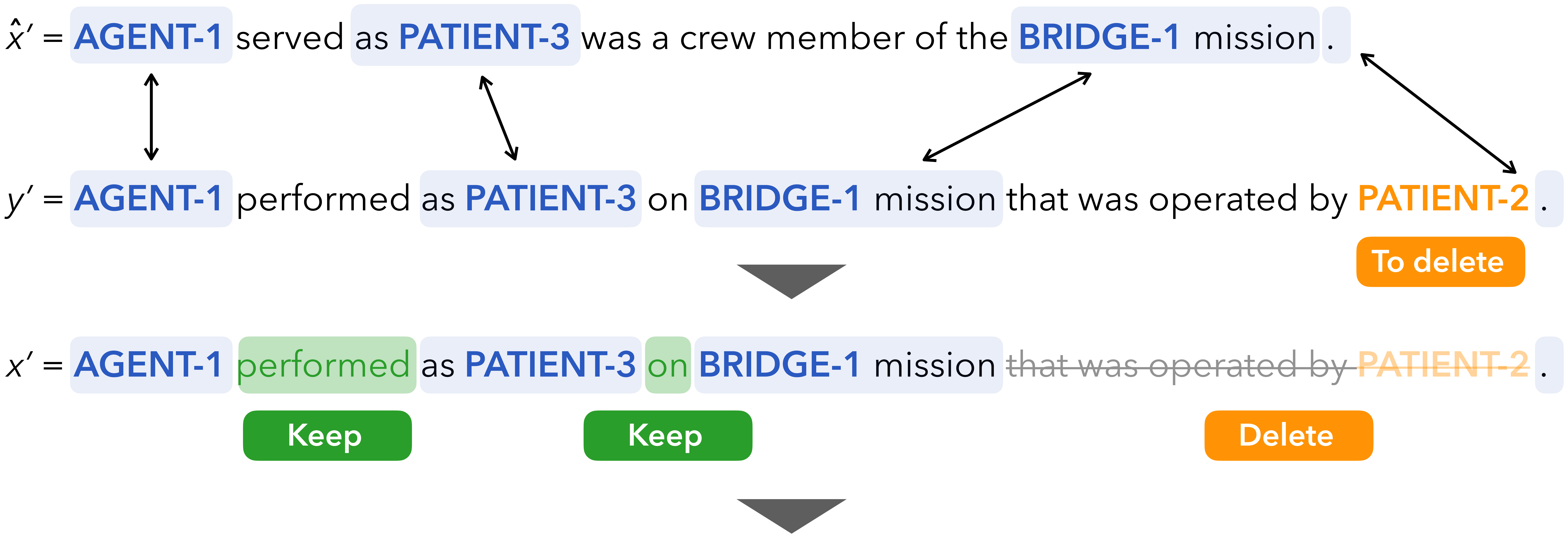


$\hat{x}' =$ AGENT-1 served as PATIENT-3 was a crew member of the BRIDGE-1 mission .

$y' =$ AGENT-1 performed as PATIENT-3 on BRIDGE-1 mission that was operated by PATIENT-2 .

To delete

Data Creation: Delete Substring



$\hat{x}' =$ AGENT-1 served as PATIENT-3 was a crew member of the BRIDGE-1 mission .

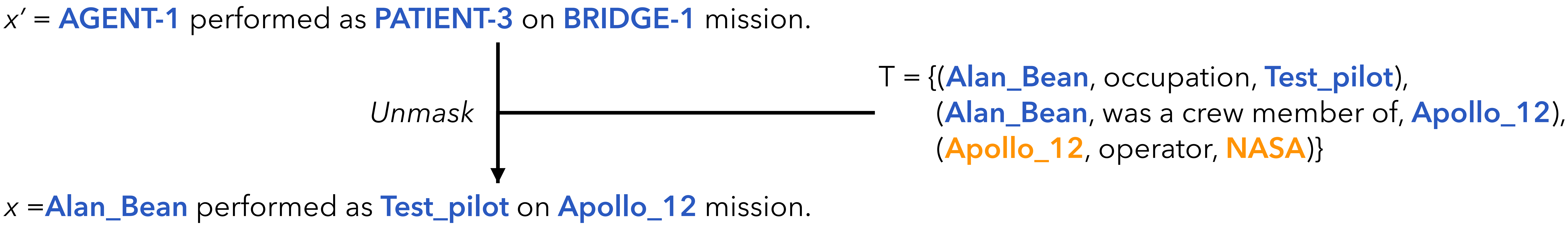
$y' =$ AGENT-1 performed as PATIENT-3 on BRIDGE-1 mission that was operated by PATIENT-2 .

$x' =$ AGENT-1 performed as PATIENT-3 on BRIDGE-1 mission ~~that was operated by PATIENT-2~~ .

$x' =$ AGENT-1 performed as PATIENT-3 on BRIDGE-1 mission.

Data Creation: Fact Unmasking

- Recovering the factual information by original facts, T.



x' = **AGENT-1** performed as **PATIENT-3** on **BRIDGE-1** mission.

Unmask

T = {(**Alan_Bean**, occupation, **Test_pilot**),
(**Alan_Bean**, was a crew member of, **Apollo_12**),
(**Apollo_12**, operator, **NASA**)}

x =**Alan_Bean** performed as **Test_pilot** on **Apollo_12** mission.

Fact-based Text Editing instance

T = {(**Alan_Bean**, occupation, **Test_pilot**), (**Alan_Bean**, was a crew member of, **Apollo_12**),
(**Apollo_12**, operator, **NASA**)}

x =**Alan_Bean** performed as **Test_pilot** on **Apollo_12** mission.

y =**Alan_Bean** performed as **Test_pilot** on **Apollo_12** mission that was operated by **NASA**.

Data Creation: Statistics

- We applied our data creation method for two publicly available datasets, **WebNLG** (Gardent et al., 2017) and **RotoWire** (Wiseman et al., 2017), to create fact-based text editing datasets, **WebEdit** and **RotoEdit**.

	WebEdit			RotoEdit		
	Train	Valid	Test	Train	Valid	Test
$\#\mathcal{D}$	181k	23k	29k	27k	5.3k	4.9k
$\#\mathcal{W}_{\mathrm{d}}$	4.1M	495k	624k	4.7M	904k	839k
$\#\mathcal{W}_{\mathrm{r}}$	4.2M	525k	649k	5.6M	1.1M	1.0M
$\#\mathcal{S}$	403k	49k	62k	209k	40k	36k

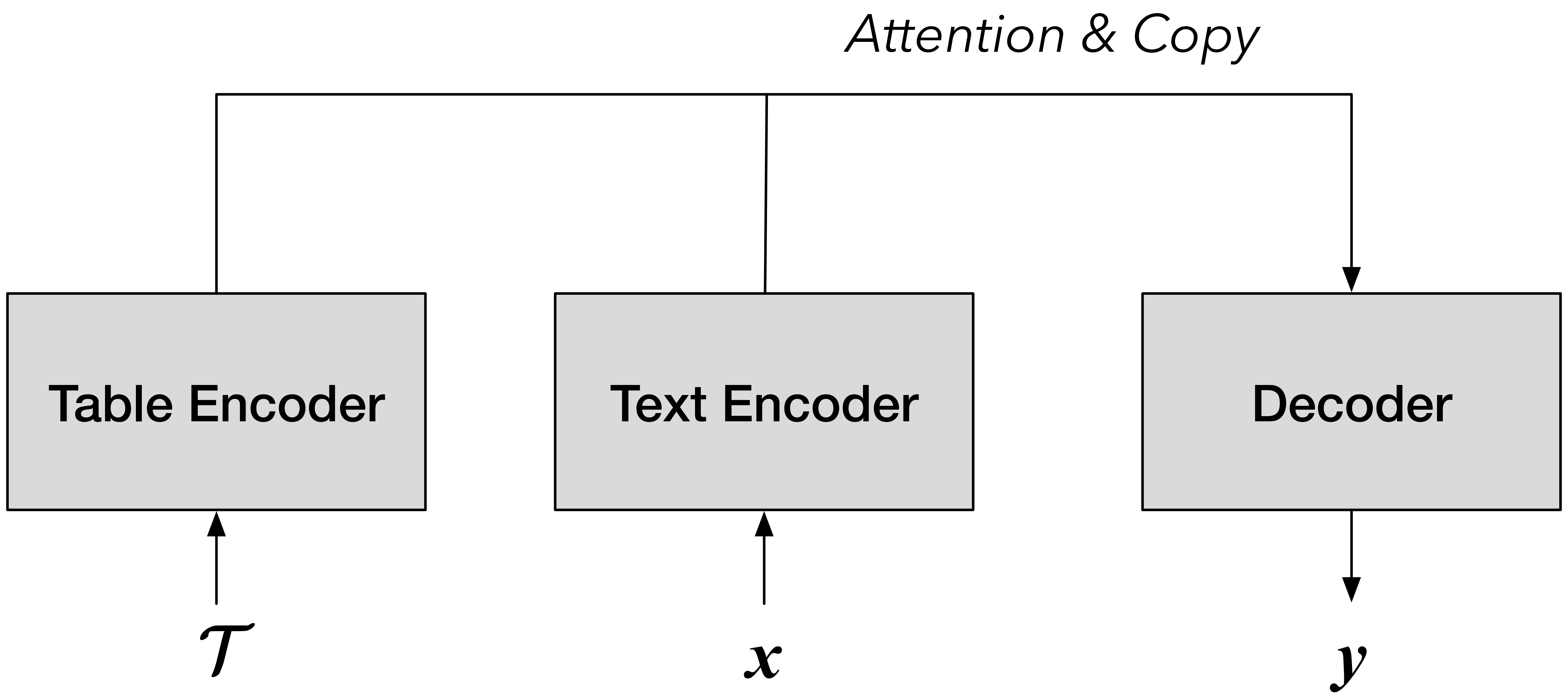
https://github.com/isomap/factedit

How to model the Fact-based Text Editing?

- A natural choice is an encoder-decoder model with attention & copy to generate the revised text from scratch.

✘ Unnecessary word replacement could happen.

✘ Inefficient for the long input & output.

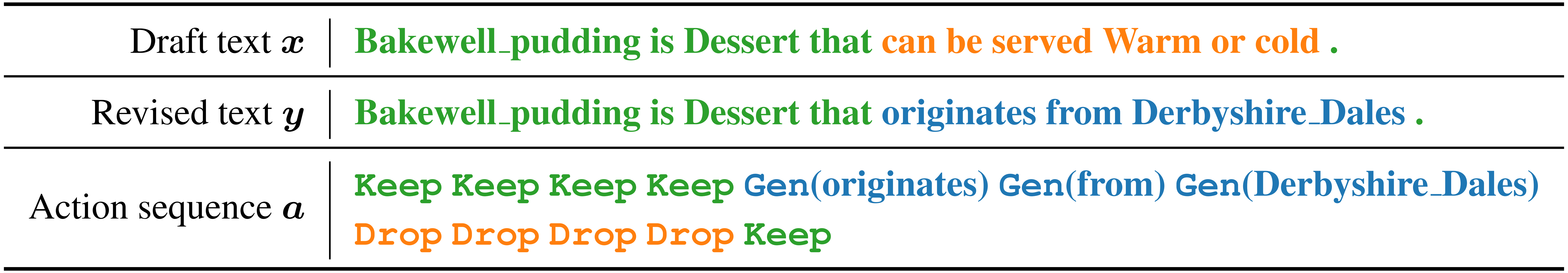


Approach: Editing through Tagging

- Instead of generating ***words*** from scratch, the model just predicts predefined ***actions***.

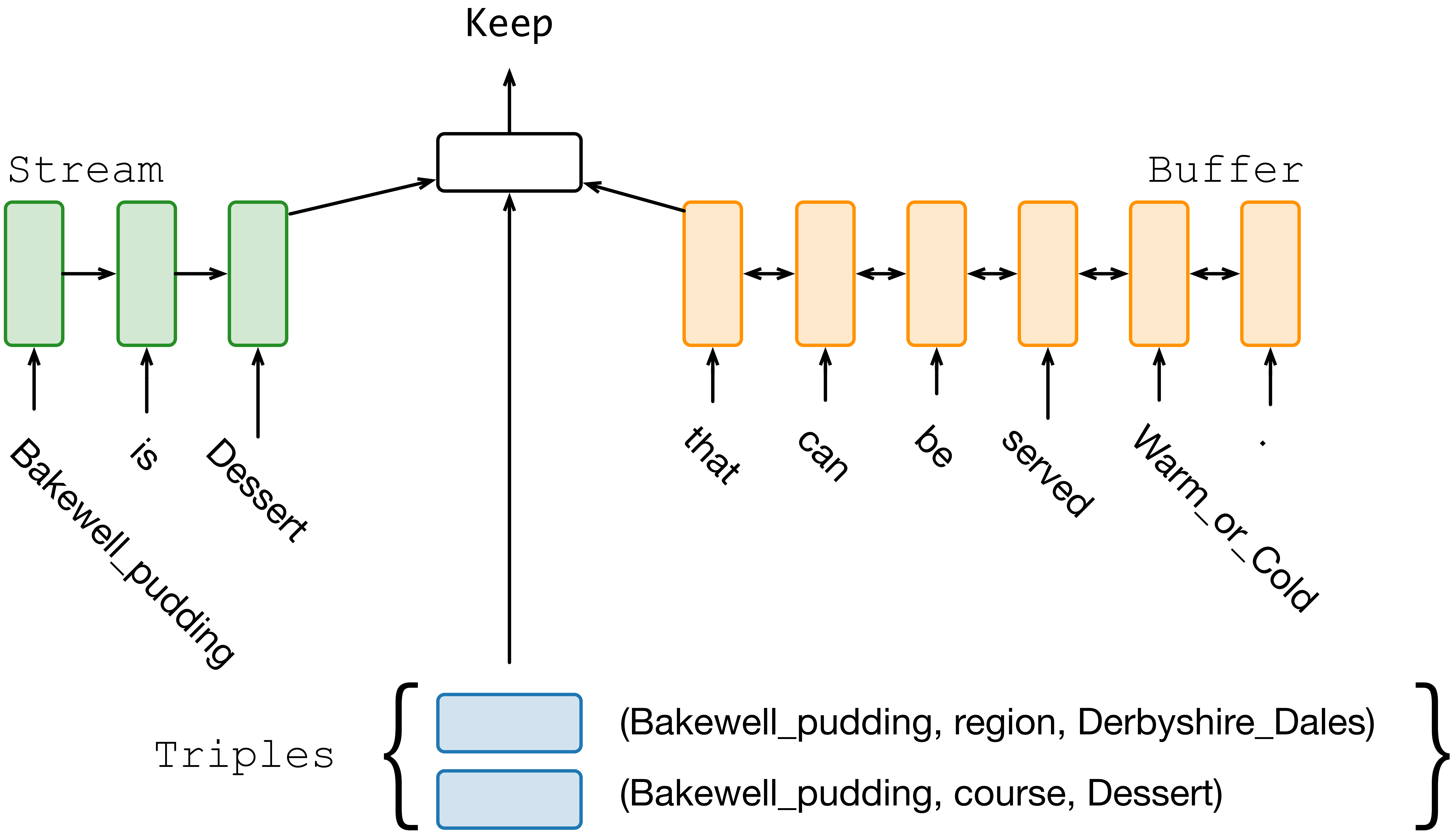
✓ Model only focuses on the explicit editing

✓ Robust to the length of input & output



Draft text $\boldsymbol{x}$	**Bakewell_pudding is Dessert that can be served Warm or cold .**
Revised text $\boldsymbol{y}$	**Bakewell_pudding is Dessert that originates from Derbyshire_Dales .**
Action sequence $\boldsymbol{a}$	**Keep Keep Keep Keep Gen(originates) Gen(from) Gen(Derbyshire_Dales) Drop Drop Drop Drop Keep**

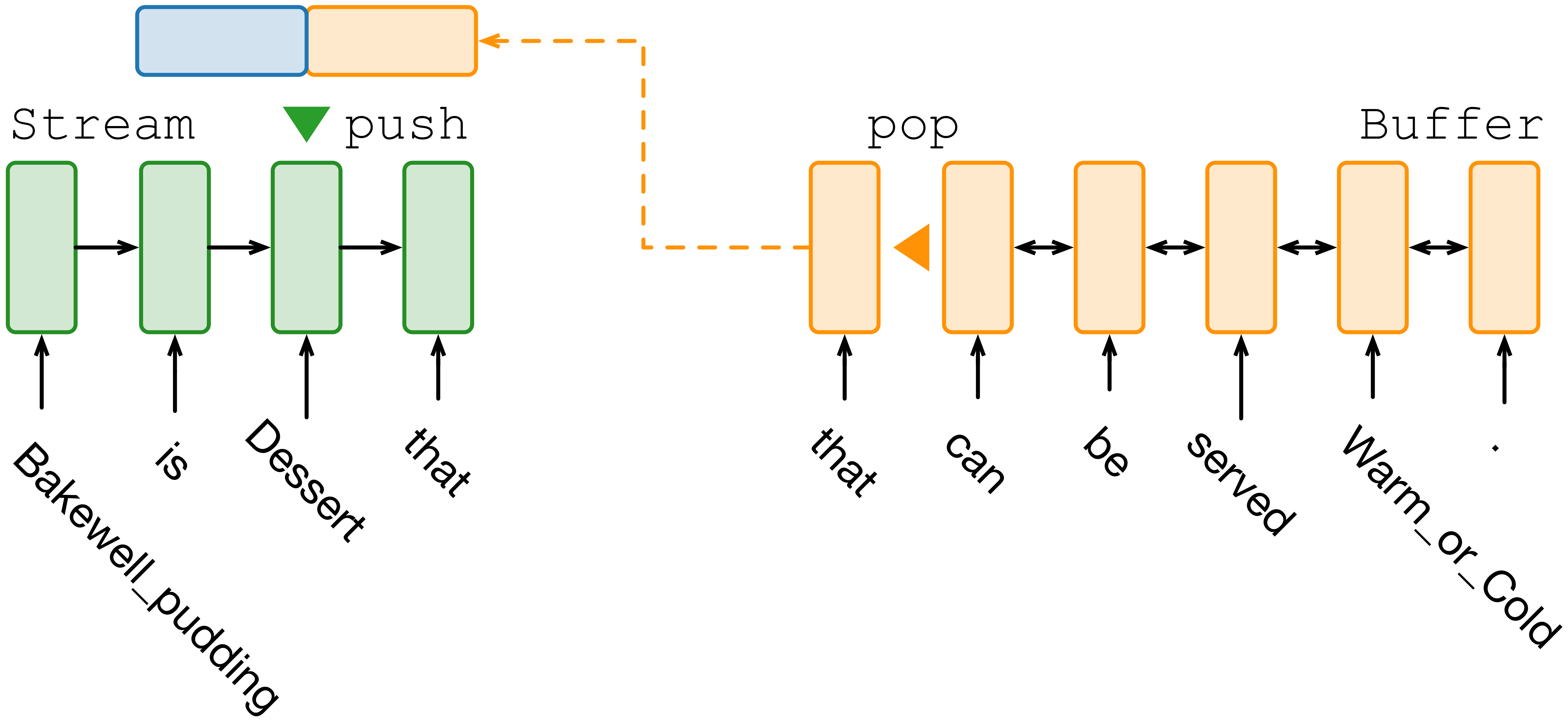
A running example: Keep



(Bakewell_pudding, region, Derbyshire_Dales)

(Bakewell_pudding, course, Dessert)

A running example: Keep

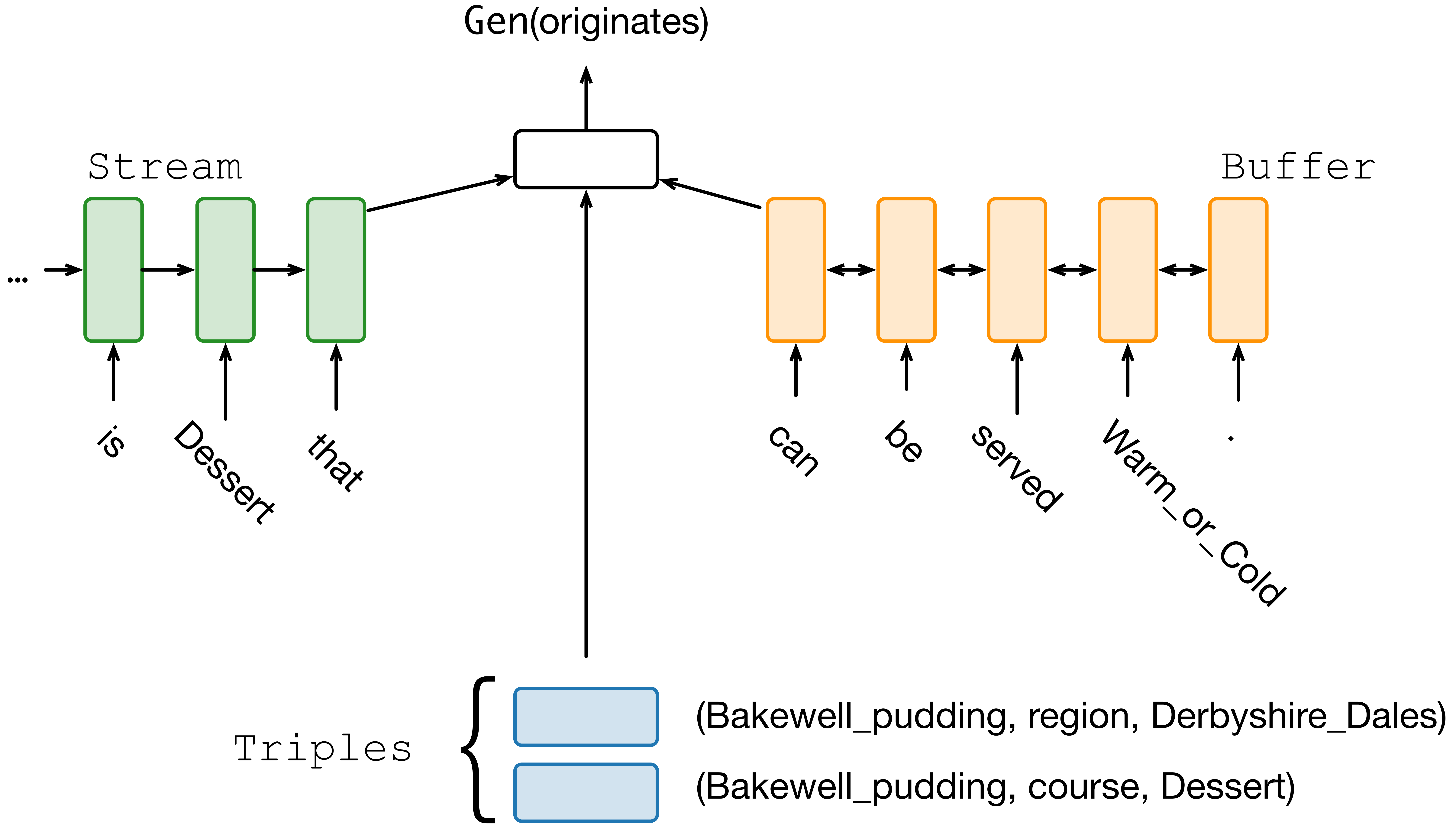


Triples

(Bakewell_pudding, region, Derbyshire_Dales)

(Bakewell_pudding, course, Dessert)

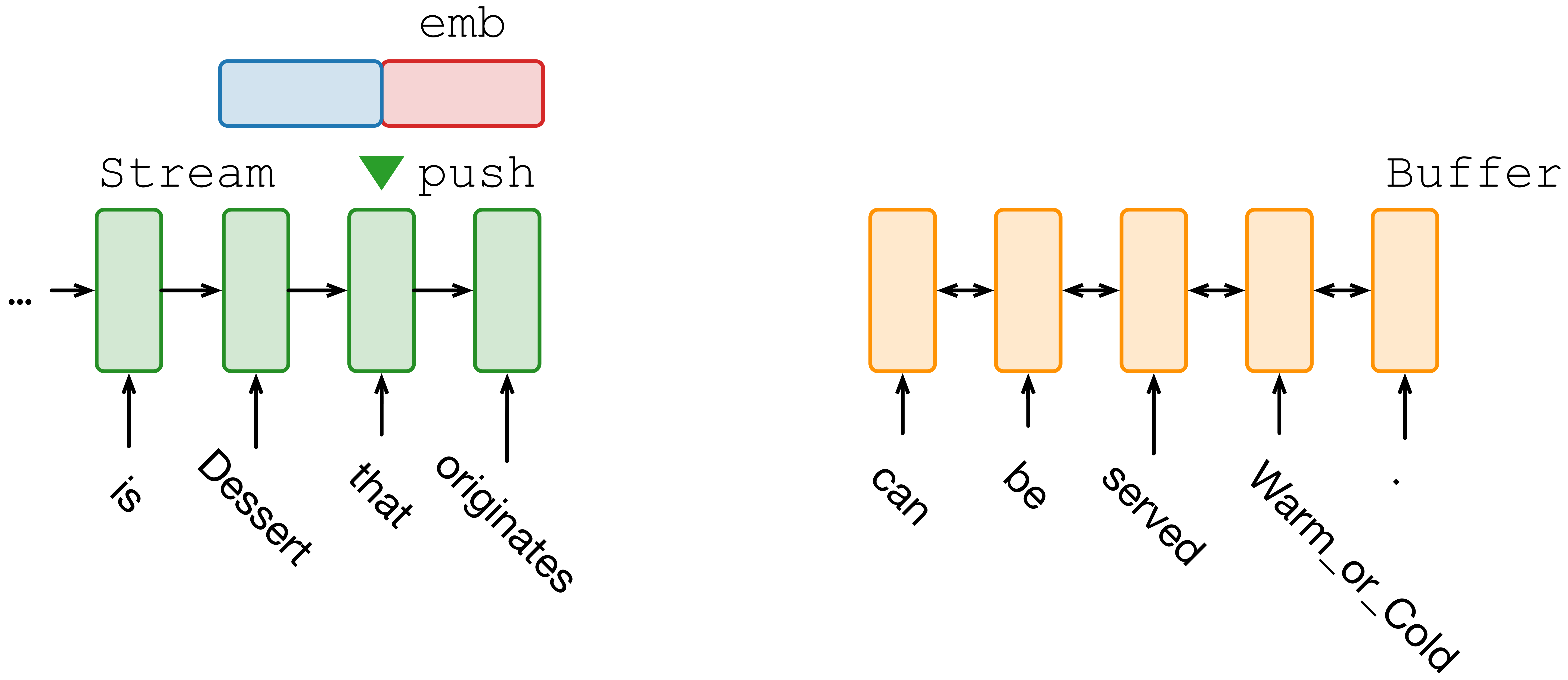
A running example: Gen



(Bakewell_pudding, region, Derbyshire_Dales)

(Bakewell_pudding, course, Dessert)

A running example: Gen

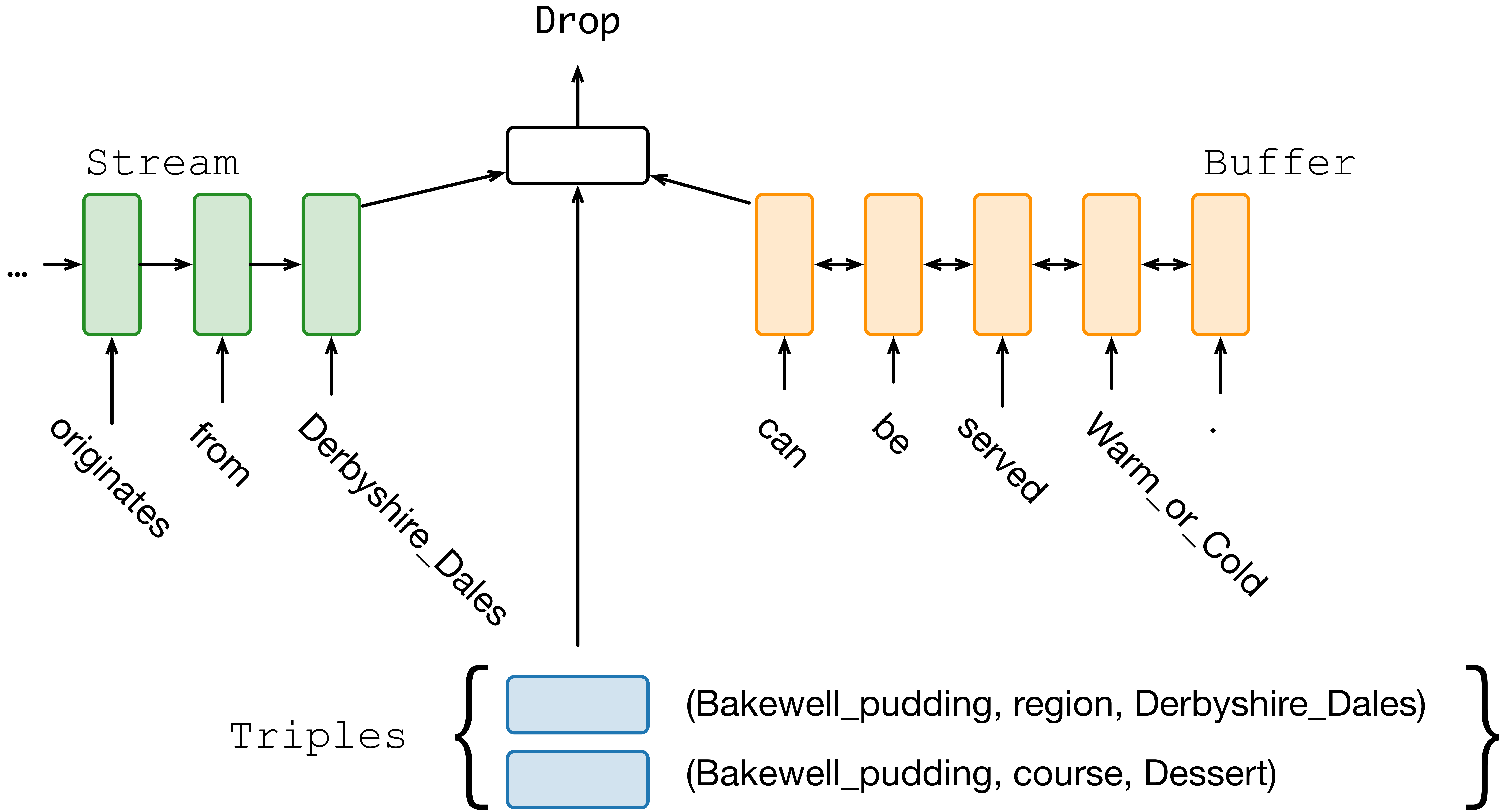


Triples

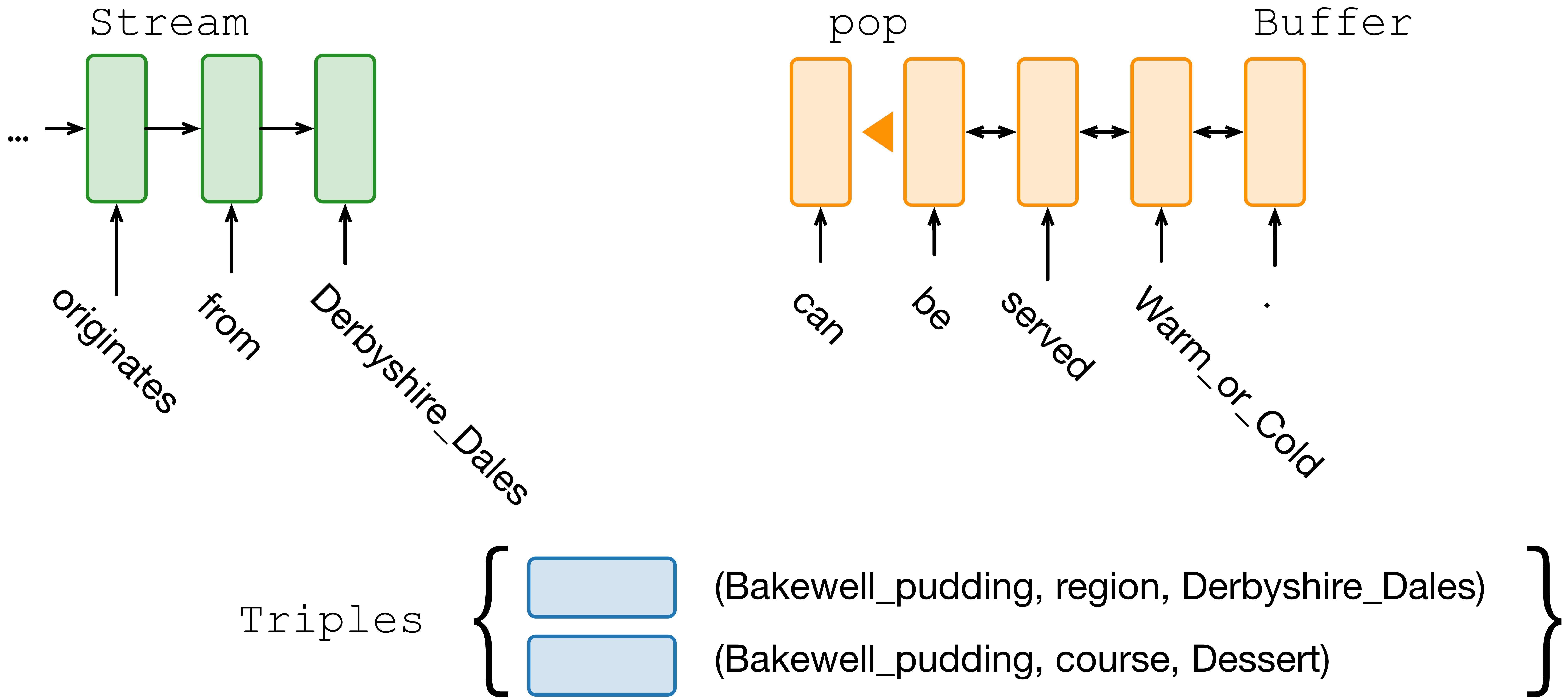
(Bakewell_pudding, region, Derbyshire_Dales)

(Bakewell_pudding, course, Dessert)

A running example: Drop



A running example: Drop



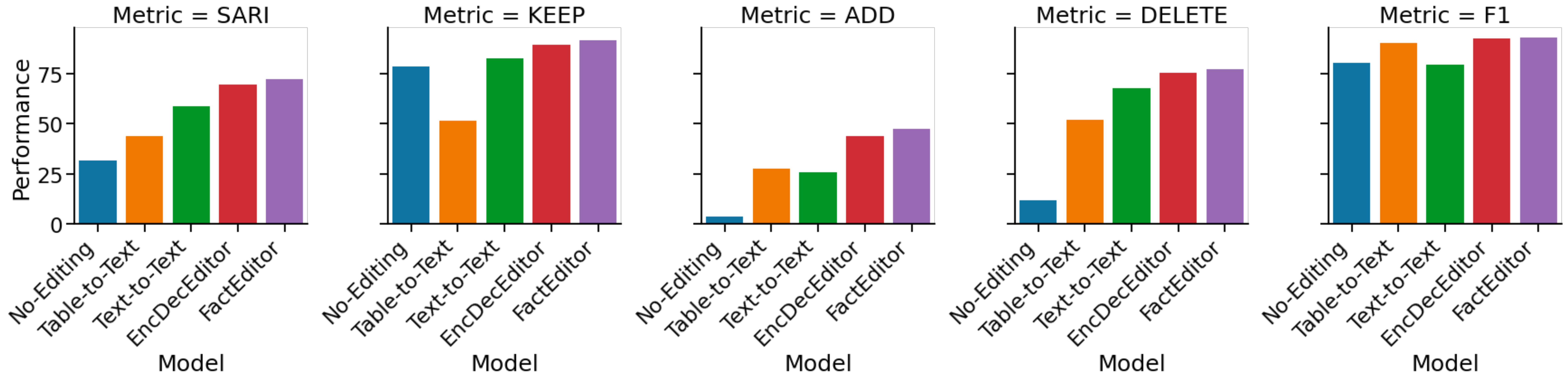
(Bakewell_pudding, region, Derbyshire_Dales)

(Bakewell_pudding, course, Dessert)

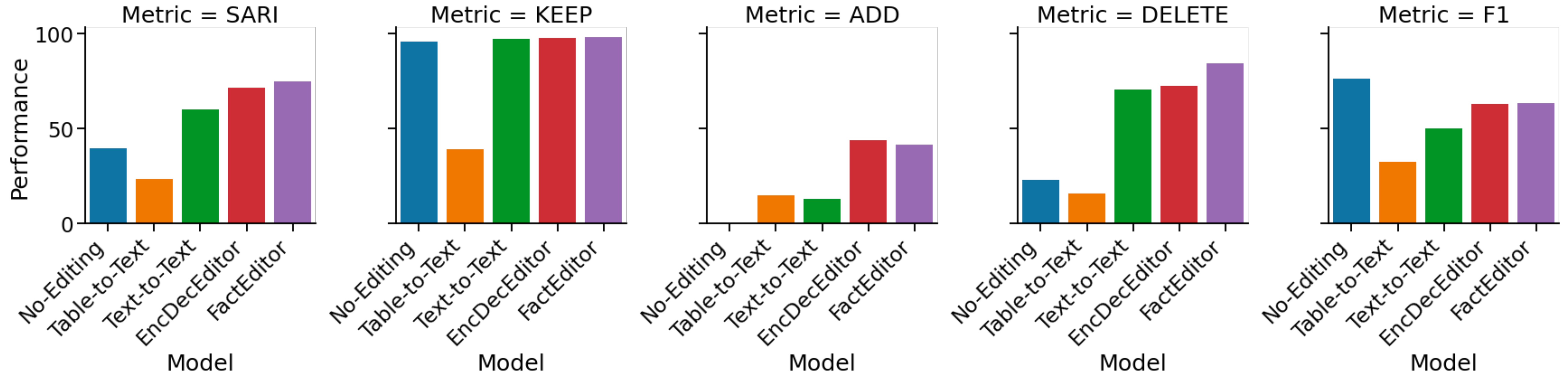
Experimental Results

- The proposed model, *FactEditor*, shows generally better performance.

WebEdit

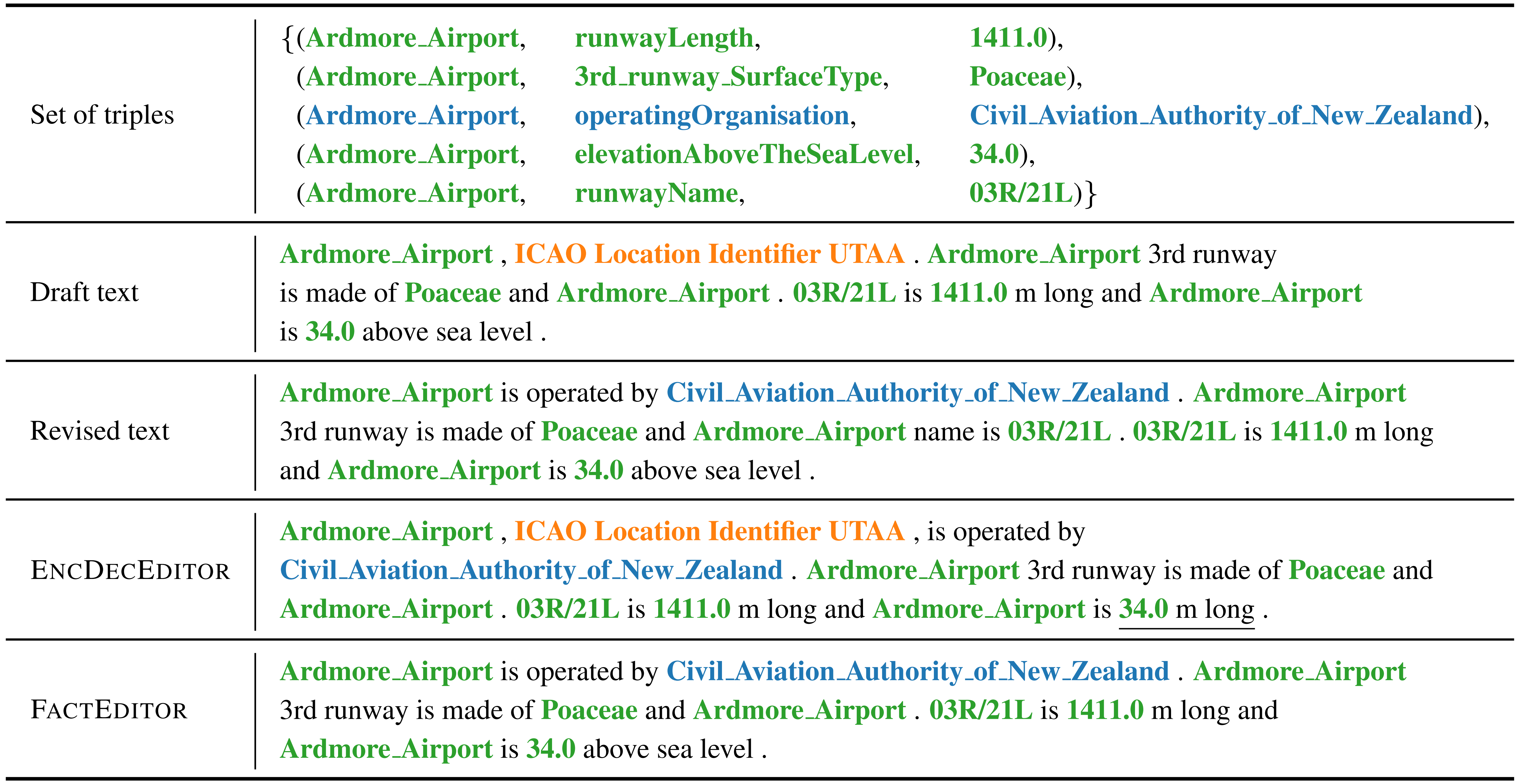


RotoEdit



Further results are in the paper

Examples

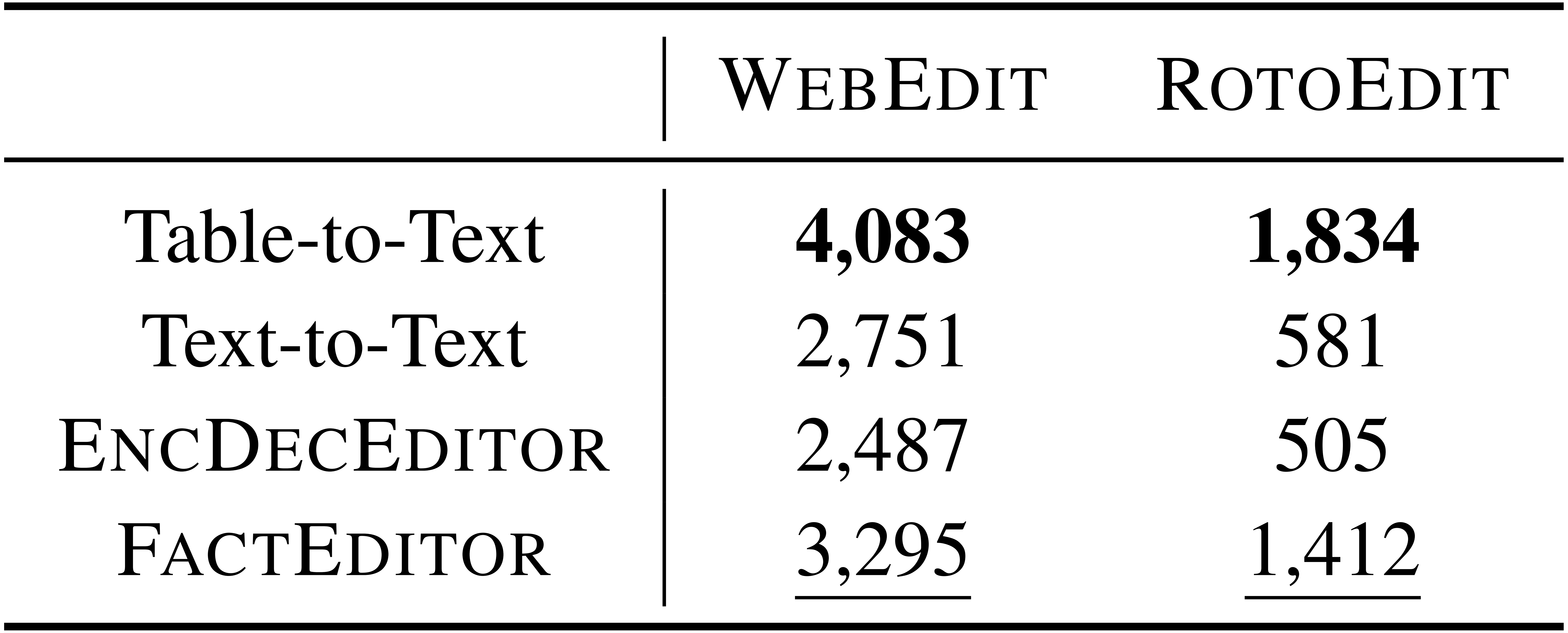


Set of triples	{(Ardmore_Airport, runwayLength, 1411.0), (Ardmore_Airport, 3rd_runway_SurfaceType, Poaceae), (Ardmore_Airport, operatingOrganisation, Civil_Aviation_Authority_of_New_Zealand), (Ardmore_Airport, elevationAboveTheSeaLevel, 34.0), (Ardmore_Airport, runwayName, 03R/21L)}
Draft text	Ardmore_Airport , ICAO Location Identifier UTAA . Ardmore_Airport 3rd runway is made of Poaceae and Ardmore_Airport . 03R/21L is 1411.0 m long and Ardmore_Airport is 34.0 above sea level .
Revised text	Ardmore_Airport is operated by Civil_Aviation_Authority_of_New_Zealand . Ardmore_Airport 3rd runway is made of Poaceae and Ardmore_Airport name is 03R/21L . 03R/21L is 1411.0 m long and Ardmore_Airport is 34.0 above sea level .
ENCDECEDITOR	Ardmore_Airport , ICAO Location Identifier UTAA , is operated by Civil_Aviation_Authority_of_New_Zealand . Ardmore_Airport 3rd runway is made of Poaceae and Ardmore_Airport . 03R/21L is 1411.0 m long and Ardmore_Airport is 34.0 m long .
FACTEDITOR	Ardmore_Airport is operated by Civil_Aviation_Authority_of_New_Zealand . Ardmore_Airport 3rd runway is made of Poaceae and Ardmore_Airport . 03R/21L is 1411.0 m long and Ardmore_Airport is 34.0 above sea level .

	EncDecEditor	FactEditor
Fluency	☺	☺
Adequecy	☹	☺
Unnecessary paraphrasing	☹	☺

Runtime analysis

- FactEditor shows the 2nd fastest inference performance.
  - It processes three times faster than EncDecEditor on RotoEdit dataset.



	WebEdit	RotoEdit
Table-to-Text	**4,083**	**1,834**
Text-to-Text	2,751	581
EncDecEditor	2,487	505
FactEditor	3,295	1,412

Summary

- We introduced the new task, *Fact-based Text Editing*.
- We have proposed a data construction method for fact-based text editing and created two datasets.
- We have proposed a model for fact-based text editing, which performs the task by generating a sequence of actions.

Code & Data available at https://github.com/isomap/factedit