Developing a Dataset of Overridden Information in Wikipedia

Masatoshi Tsuchiya
Yasutaka Yokoi
Toyohashi University of Technology

Example of overridden information

s_t	As of the year 2010, Oshida Station is an unmanned station in Asagishi Aza Oshida, Morioka City, Iwate Prefecture, Japan, with three train stops per day.
s_r	Oshida Station was an unmanned station in Asagishi Aza Oshida, Morioka City, Iwate Prefecture, Japan, which was discontinued in the year 2016, and is not in use as of the year 2018.

- The operation status of Oshida Station described in sentence s_t has been overridden by sentence s_r.
  - Sentence s_t was taken from Japanese Wikipedia in 2014.
  - Sentence s_r was taken from Japanese Wikipedia in 2018.
  - In this research, sentence s_t is called as a target sentence, and sentence s_r is called as a reference sentence.
- Since all information on the Web is not updated in a timely manner, detection of overridden information is necessary.

Two possible task settings

- A binary classification problem taking **only** a target sentence as an input.
- A binary classification problem taking **both** a target sentence and a reference sentence as an input.
- The latter setting is employed, although the former setting looks better, because the subtask to discover a suitable reference sentence is expected to make the whole task too difficult.

s_t	As of the year 2010, Oshida Station is an unmanned station in Asagishi Aza Oshida, Morioka City, Iwate Prefecture, Japan, with three train stops per day.

s_t	As of the year 2010, Oshida Station is an unmanned station in Asagishi Aza Oshida, Morioka City, Iwate Prefecture, Japan, with three train stops per day.
s_r	Oshida Station was an unmanned station in Asagishi Aza Oshida, Morioka City, Iwate Prefecture, Japan, which was discontinued in the year 2016, and is not in use as of the year 2018.

Existence of neutral sentence

s_t	As of the year 2010, Oshida Station is an unmanned station in Asagishi Aza Oshida, Morioka City, Iwate Prefecture, Japan, with three train stops per day.
s_r	Oshida Station was an unmanned station in Asagishi Aza Oshida, Morioka City, Iwate Prefecture, Japan, which was discontinued in the year 2016, and is not in use as of the year 2018.
s_n	Morioka City constructed a 13-kilometer sidewalk from Oshida Station to Asagishi Station, which became available for use on May 15, 2015.

- We **can** suppose that the operation status of Oshida Station described in sentence s_t has been overridden by sentence s_r.
- We **cannot** suppose that the operation status of Oshida Station described in sentence s_t has been overridden by sentence s_n.
- The definition of information override is necessary.

Decompose the sentence into a tuple

s_1	The tallest building in Japan is Yokohama Landmark Tower as of the year 2010.
s_2	The tallest building in Japan is Abeno Harukas as of the year 2015.
s_3	The tallest building in the world is Burj Khalifa as of the year 2010.

$s_1 = \langle x_1, t_1 \rangle$
x_1 = The tallest building in Japan is Yokohama Landmark Tower.
t_1 = as of the year 2010

$s_2 = \langle x_2, t_2 \rangle$
x_2 = The tallest building in Japan is Abeno Harukas.
t_2 = as of the year 2015

- Suppose the three sentences on the left side.
  - Reference sentence s_2 overrides target sentence s_1.
  - Neutral sentence s_3 does not override target sentence s_1.
- Decompose the sentence s into a tuple of the content x that is valid regardless of time and the time t that the sentence s is focused on.

$$s = \langle x, t \rangle$$

Contradiction occurs when replacing the time information of the target sentence

s_1	The tallest building in Japan is Yokohama Landmark Tower as of the year 2010.
s_2	The tallest building in Japan is Abeno Harukas as of the year 2015.
s_3	The tallest building in the world is Burj Khalifa as of the year 2010.

$s_1 = \langle x_1, t_1 \rangle$
x_1 = The tallest building in Japan is Yokohama Landmark Tower.

$s_2 = \langle x_2, t_2 \rangle$
t_2 = as of the year 2015

$s_{1,2} = \langle x_1, t_2 \rangle$
= The tallest building in Japan is Yokohama Landmark Tower as of the year 2015.

- Generate sentence $s_{1,2}$ from sentence s_1 by replacing its time only with the time of sentence s_2.
- Note that sentence $s_{1,2}$ **contradicts** sentence s_2.

Contradiction DOES NOT occur when replacing the time information of the target sentence

s_1	The tallest building in Japan is Yokohama Landmark Tower as of the year 2010.
s_2	The tallest building in Japan is Abeno Harukas as of the year 2015.
s_3	The tallest building in the world is Burj Khalifa as of the year 2010.

$s_1 = \langle x_1, t_1 \rangle$
x_1 = The tallest building in Japan is Yokohama Landmark Tower.

$s_3 = \langle x_3, t_3 \rangle$
t_3 = as of the year 2010

$s_{1,3} = \langle x_1, t_3 \rangle$
= The tallest building in Japan is Yokohama Landmark Tower as of the year 2010.

- Generate sentence $s_{1,3}$ from sentence s_1 by replacing its time only with the time of sentence s_3.
- Note that sentence $s_{1,3}$ **does not contradict** sentence s_3.
- This observation suggests that contradiction may be a starting point of the definition of information override.

Definition of information override

- Formal definition of textual entailment recognition

$$f(s_h, s_p) = \begin{cases} \text{if } s_h \rightarrow s_p \ \vee \ s_p \rightarrow s_h \\ \quad \text{entailment} \\ \text{if } s_h \cap s_p = \phi \\ \quad \text{contradiction} \\ \text{otherwise} \\ \quad \text{independence} \end{cases}$$

- Formal definition of detection of overridden information based on textual entailment

$$g(s_t, s_r) = \begin{cases} \text{if } f(s_t, s_r) = \text{contradiction} \\ \quad \vee f(\langle x_t, t_r \rangle, s_r) = \text{contradiction} \\ \quad \vee f(s_t, \langle x_r, t_t \rangle) = \text{contradiction} \\ \quad \text{override} \\ \text{otherwise} \\ \quad \text{neutral} \end{cases}$$

This definition means that the detection of overridden information is considered as the combined inference of textual entailment and temporal relation.

Construction procedure of the dataset

1. Sentence level changes between two versions of Japanese Wikipedia were collected as annotation targets.
2. Remove changes that do not contain either reference to updated articles or time and date expressions.
3. Remove changes caused by minor editing
4. Remove changes from low-quality articles
5. Human annotation!

Statistics of the dataset

	Override		Neutral			
Entailment	430	(4.5%)	3765	(39.2%)	4195	(43.7%)
Contradiction	1359	(14.2%)	0	(0.0%)	1359	(14.2%)
Independence	1301	(13.6%)	2745	(28.6%)	4046	(42.1%)
	3090	(32.2%)	6510	(67.8%)	9600	

- 32.2% of target pairs were judged to be overridden information, and it was revealed that the old version of Wikipedia contained much overridden information.
- The inter-annotator agreement ratio for textual entailment labels was 83.7%, and the ratio for information override labels was 88.0%. This suggests that human annotation for overridden information is reliable.

Article Title	Target Sentence	Reference Sentence	Textual Entailment Label	Information Override Label
Excesior Cafe	As of March 2014, it operates three stores in Tokyo and Saitama.	As of October 2016, it operates only one store in Saitama.	independence	override
Machinori (rental bicycle)	As of November 2012, it operates 19 service stations.	As of May 2017, it operates 22 service stations, including its office.	independence	override
President of Italy	The current holder is Giorgio Napolitano.	The current holder is Sergio Mattarella.	contradiction	override
Okinawa Urban Monorail	It is scheduled to open in the spring of 2019.	It was scheduled to open in the spring of 2019, but it was announced in May 2018 that it would be in the summer of 2019 at the earliest.	contradiction	override
Kunitachi Station	The ticket gates were consolidated to one under the elevated tracks as of January 13, 2013.	The ticket gates were consolidated to one under the elevated tracks as of January 13, 2013, but Nonowa Gate was built on the west side on April 24, 2016, bringing the number of ticket gates to two.	entailment	Override
Komeri Co.	As of August 2011, it operates stores in all prefectures except Okinawa.	As of July 2018, it operates stores in all prefectures except Okinawa.	independence	neutral
Nihon University Itabashi Hospital	It is operated by Nihon University Educational Corporation.	It is an affiliated hospital of Nihon University School of Medicine.	entailment	neutral

Conclusion

- We propose a new task of detecting overridden information and formalize it as a binary classification problem to determine whether a reference sentence has overridden a target sentence.
- We offer a formal definition of information override between two sentences while relating it to textual entailment.
- We propose a procedure to construct a dataset of overridden information by collecting sentence pairs from the difference between two versions of Wikipedia.