A Multi-grained Dataset for News Event Triggered Knowledge Update

Yu-Ting Lee, Ying-Jhe Tang, Yu-Chung Cheng, Pai-Lin Chen, Tsai-Yen Li, Hen-Hsen Huang

National Chengchi University, Taipei, Taiwan Academia Sinica, Taipei, Taiwan https://github.com/hhhuang/NetKu

event e.

event e .

文A Language

E Contents

^ Characteristics

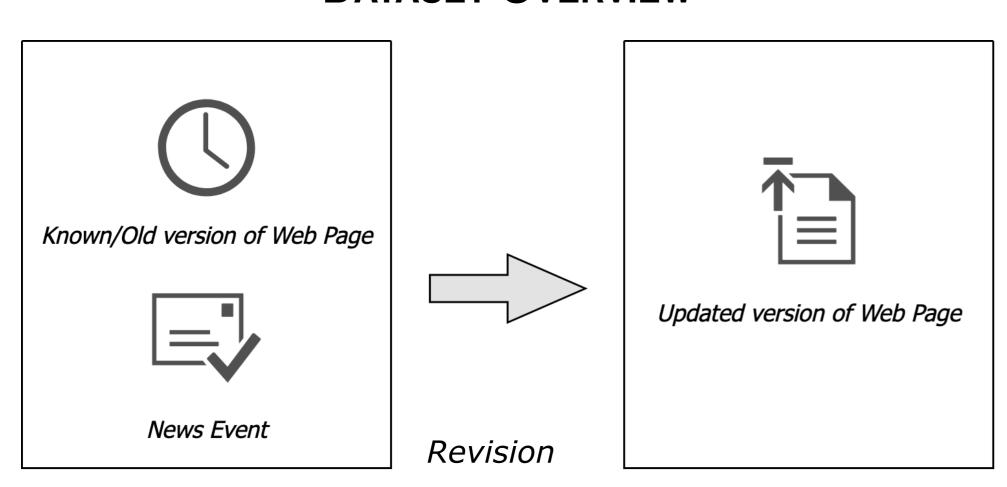
Sri Lankan leopard

oologist Paules Edward Pieris Deraniyagala.^{[2}

ABSTRACT

- The world rapidly changes and new information emerges every second.
- Keep the knowledge up-to-date is not a trivial task
 - Articles from Wikipedia are collected and aligned with multiple language units.
 - Given an existing article about a topic with a news event about the topic, aims to generate an updated article according to the known information and new occurred event.
- We create the news-triggered dataset and baselines for knowledge update generation task with three finegrained levels.

DATASET OVERVIEW



DATASET CONSTRUCTION

- Parsing contents from the Wikipedia and crawling raw news from external website:
 - Daily news and the topic related to news ranging from 2006 to 2022 in the portal of Wikipedia Current Events.
 - Collect relevant topics linked from Wikipedia-Current Events: Retrieve and parse raw news from Internet Archive Wayback Machine by given cited news urls from Wikipedia Current Events.
- Data Alignment $(x_i \oplus e_i \rightarrow y_i)$
 - Align news event e_i listed on the portal of Wikipedia Current Events at time *i* with the version of the Wikipedia page before and after time t_i , representing in x_i and y_i .
 - Irrelevant edits problems between x_i and y_i : We restrict the time window of one week between (e_i, y_i)
- Data Filtering: Reduce the sentence-level bias
 - Remove the duplicated instances and sentences appeared in WCEP-10 dataset.
 - Keep the instances with existing x_i and English language summary.

LEVELS OF KNOWLEDGE FACTS

and proclaimed intentions to re-establish the Islamic Emirate of Afghanistan after the August 2021 fall of Kabul, there have been two known drone strikes against the country. As a result of the airstrikes, there has been at least 12 fatalities and one injury in the Islamic Emirate of Afghanistan.

The summary of the previous version

aid worker for U.S. based Nutrition and Education International, who was bringing home water for his family. At least 10 people are killed during the strike, including seven

The triggered news event

fter the Taliban returned to power in Afghanistan and proclaimed intentions to re-establish the Islamic Emirate of Afghanistan after the August 2021 fall of Kabul, there have been two known drone strikes against the country As a result of the airstrikes, there has been at least 12 fatalities and one injury in the Islamic Emirate of

Match The generated summary

After the Taliban returned to power in Afghanistan and proclaimed intentions to re-establish the Islamic Emirat of Afghanistan after the August 2021 fall of Kabul, there have been two known drone strikes against the country. As a result of the airstrikes, there has been at least 12 fatalities and one injury in the Islamic Emirate of

Crawl and parse the web-page contents, further merged with the portal of Wikipedia Current Events.

IMPLEMENTATION

Full Content

Summary: Generate the updated summary y given the

summary of the previous version and the triggered news

Full content: Generate the updated full content y given the

full content of the previous version and the triggered news

First Paragraph

☆ Watch 🎤 Edit

Sri Lankan leopard

Conservation status

Summary

- Align the news triggered event and Wikipedia page contents between versions.
- Running data filtering to reduce bias.

The **Sri Lankan leopard** (*Panthera pardus kotiya*) is a leopard subspecies native to

Since 2020, the Sri Lankan leopard has been listed as vulnerable on the IUCN Red List, as the

population is estimated at less than 800 mature individuals, and is probably declining.^[1]

The Sri Lankan leopard has a tawny or rusty yellow coat with dark spots and close-set

rosettes, which are smaller than in Indian leopards. Seven females measured in the early 20th century averaged a weight of 64 lb (29 kg) and had a mean head-to-body-length of 3 ft 5 in

(1.04 m) with a 2 ft 6.5 in (77.5 cm) long tail, the largest being 3 ft 9 in (1.14 m) with a 2 ft 9 in

(84 cm) long tail; 11 males averaged 124 lb (56 kg), the largest being 170 lb (77 kg), and measured 4 ft 2 in (1.27 m) with a 2 ft 10 in (86 cm) long tail, the largest being 4 ft 8 in

- Apply lexical overlapping calculation to obtain basic baseline.
- We conduct the SOTA pre-trained model PRIMERA to produce inference and fine-tune with our data.
- Evaluate the experiments results with multiple metrics (recallbased, precision-based, sentence similarity).
- Given a news event e and a related knowledge fact x, aims to generate an update knowledge fact y.

$$y' = \arg \max_{u} Pr(y|x, e)$$

Method	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	BS
Generating the First Paragraph									
Original x	0.9560	0.9520	0.9559	0.5536	0.5936	0.5743	0.5478	0.7193	0.988
Concatenation of x and e	0.9613	0.9537	0.9608	0.5540	0.5941	0.5747	0.5482	0.7144	0.956
PRIMERA w/o fine-tuning	0.4534	0.3713	0.4422	0.5705	0.4679	0.4237	0.3992	0.5517	0.910
PRIMERA (fine-tuned)	0.9601	0.9585	0.9599	0.8454	0.9098	0.8905	0.8600	0.8519	0.990
Generating the Summary									
Original x	0.9467	0.9371	0.9466	0.7786	0.8758	0.8557	0.8186	0.7629	0.985
Concatenation of x and e	0.9505	0.9391	0.9499	0.5608	0.6398	0.6230	0.5920	0.6592	0.962
PRIMERA w/o fine-tuning	0.3554	0.2790	0.3469	0.6383	0.5343	0.4783	0.4467	0.5689	0.902
PRIMERA (fine-tuned)	0.9565	0.9525	0.9564	0.7854	0.8904	0.8780	0.8480	0.7743	0.989
Generating the Full Content									
Original x	0.9117	0.8883	0.9115	0.4496	0.7172	0.7634	0.7351	0.3447	0.979
Concatenation of x and e	0.9173	0.8918	0.9166	0.4184	0.6800	0.7263	0.6996	0.3259	0.979
PRIMERA w/o fine-tuning	0.0695	0.0372	0.0676	0.8256	0.6961	0.6007	0.5411	0.6101	0.870
PRIMERA (fine-tuned)	0.5185	0.4705	0.5179	0.5999	0.7785	0.7899	0.7641	0.4942	0.966

Table 3: Results of the experiments at the levels of the first paragraph, the summary, and the full content.

Model learns what remains unchanged in the updated

After the Taliban returned to power in Afghanistan A U.S. airstrike targets a vehicle driven by an

first paragraph generation.

First Paragraph: Generate the updated first paragraph ygiven the first paragraph of the previous version and the triggered news event e.

Reference Summary