Puzzle

- A bad king has a cellar of 1000 bottles of delightful and very expensive wine. A neighbour queen wants to kill the bad king and sends a servant to poison the wine.

- Fortunately (or say unfortunately) the bad king's guards catch the servant after he could poison only one bottle. The poison takes **one day** to have an effect and make people die.

- The bad king decides to get some of the prisoners to drink the wine in order to help him find the bottle of wine that was poisoned **in one day**.

- Despite being a bad king, he still wants to reduce the death rate. Do you know how many prisoners needed **at least**? Please explain~

Puzzle Answer

- The answer: **no more than 10 prisoners**
- The number of the bottles are 1 to 1000. Now, write the number in binary format:
  - bottle 1 = 0000000001 (10 digit binary)
  - bottle 2 = 0000000010
  - …
  - bottle 500 = 0111110100
  - bottle 1000 = 1111101000
- Now, take 10 prisoners and number them 1 to 10. Let prisoner i take a sip from every bottle that has a 1 in its i-th bit. And, this process will continue for every prisoner until the last prisoner is reached. For example:
  - Prisoner = 10 9 8 7 6 5 4 3 2 1
  - Bottle 924 = 1 1 1 0 0 1 1 1 0 0
- For instance, bottle no. 924 would be sipped by 10,9,8,5,4 and 3. That way if bottle no. 924 was the poisoned one, only those prisoners would die.
- After one day, line the prisoners up in their bit order and read each living prisoner as a 0 bit and each dead prisoner as a 1 bit. The number that you get is the poisoned wine.
- We know, 1000 is less than 1024 (2^10). Therefore, if there were 1024 or more bottles of wine it would take more than 10 prisoners.

Generating topic-oriented summaries using neural attention
[NAACL 2018]

Group Presentation

WANG, Yue
01/21/2019

Outline

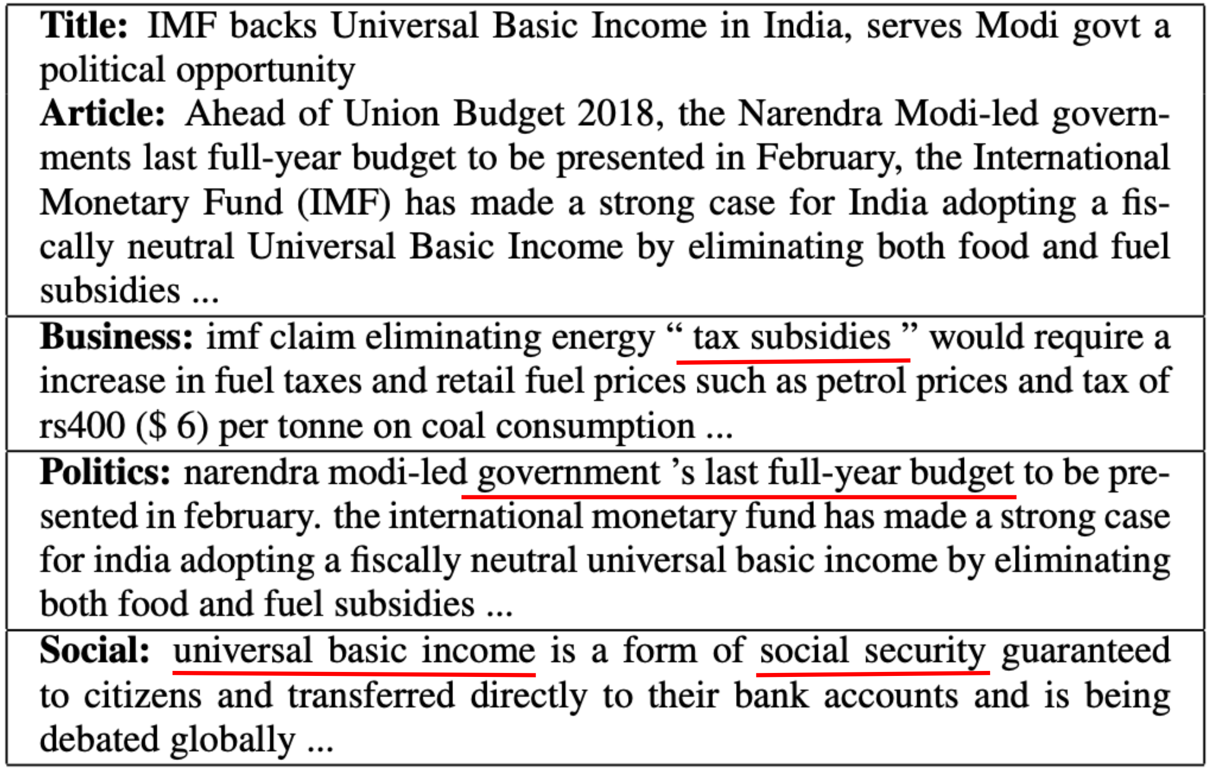
- Background
- Method
- Experiment
- Conclusion

Background

- Document summarization
    - Identifying the **important parts** of the document to provide a quick overview to a reader
- New task
    - Summarize a **single** document into **multiple** summaries with **different topics** of interest
- Motivation
    1. A long article can span several topics, which cannot be entirely cover by a single summary
    2. The interests of readers can vary and the notion of importance can change across different readers

Background

- Example:



Title: IMF backs Universal Basic Income in India, serves Modi govt a political opportunity
Article: Ahead of Union Budget 2018, the Narendra Modi-led governments last full-year budget to be presented in February, the International Monetary Fund (IMF) has made a strong case for India adopting a fiscally neutral Universal Basic Income by eliminating both food and fuel subsidies ...

Business: imf claim eliminating energy “ tax subsidies ” would require a increase in fuel taxes and retail fuel prices such as petrol prices and tax of rs400 ($ 6) per tonne on coal consumption ...

Politics: narendra modi-led government ’s last full-year budget to be presented in february. the international monetary fund has made a strong case for india adopting a fiscally neutral universal basic income by eliminating both food and fuel subsidies ...

Social: universal basic income is a form of social security guaranteed to citizens and transferred directly to their bank accounts and is being debated globally ...

Table 1: Topic oriented summaries generated by our method for an article (from LiveMint) touching multiple topics

Method

- **A novel approach** to artificially create a topic-centric training corpus key part!!

- **Topic aware** pointer-generator network

Method

- **Ideal dataset:** each article can have multiple summaries, each annotated with a topic

- **Challenge:** no existing suitable datasets
  - Usually one article with one summary
  - Summaries without topic annotation

- How to create such a dataset?
  - Too expensive for human labelling
  - Create an artificial dataset!

Method

- **Source dataset:** CNN/Dailymail dataset
  - One article with one summary without topic annotation

- **Two questions** before creating our target dataset
  1. How to annotate a summary with a topic
  2. How to convert the (article, summary) pair into {(article, summary1), (article, summary2), … }

Method

- Summary topic annotation

1. Learn topic representation by leveraging external dataset
   - News dataset tagged with topics like politics, sports, education etc. *(from 2017 KDD Data Science + Journalism Workshop)*
   - For each topic, group its articles and compute the normalized bag-of-words (BOW) representation as the topic vector $\boldsymbol{e_t}$:
     - $e_t = normalize([n_1, n_2, \ldots, n_v])$
     - v: vocabulary size
     - n_i: occurrence count of $\boldsymbol{i}$-th word

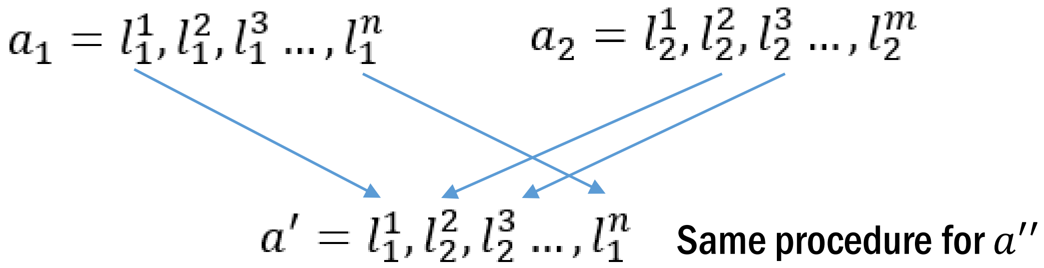
Method

- Summary topic annotation

2. Assign topic by comparing summary's BOW with $\boldsymbol{e_t}$
    - Corpus: article and summary pairs (a, s)
    - For each topic t_i: $sim(s, t_i) =< v_s, e_{t_i} >$
        - v_s: BOW of summary s
        - e_{t_i}: topic vector of i-th topic
    - For the topic i and j with highest and second highest similarity:
        - Set confidence $c = \frac{sim(s,t_i)}{sim(s,t_j)}$
        - If $c > threshold$(1.2),add (a, u_{t_i}, s) into **intermediate** dataset
            - u_{t_i}: a one-hot vector with i-th entry storing confidence score c

Method

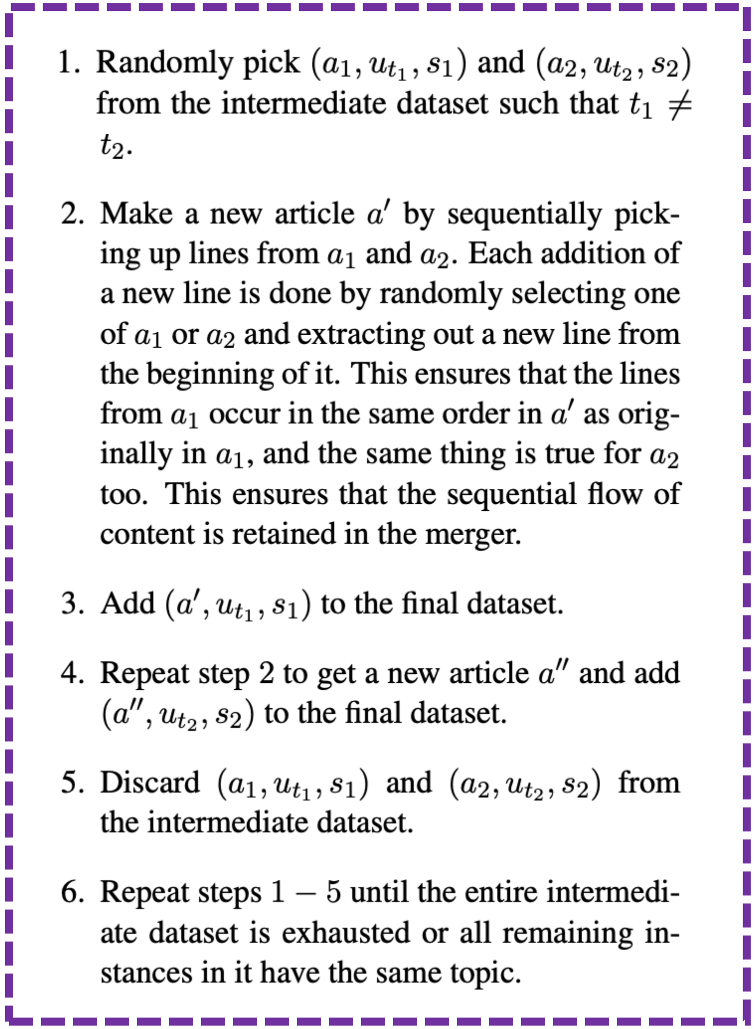
- Start create our artificial datasets
  - For (a_1, u_{t_1}, s_1) and (a_2, u_{t_2}, s_2)
    - $a_1 = l_1^1, l_1^2, l_1^3 \dots, l_1^n$
    - $a_2 = l_2^1, l_2^2, l_2^3 \dots, l_2^m$
  - Randomly pick lines from $\boldsymbol{a_1}$ or $\boldsymbol{a_2}$



$a_1 = l_1^1, l_1^2, l_1^3 \dots, l_1^n$ $\quad$ $a_2 = l_2^1, l_2^2, l_2^3 \dots, l_2^m$

$a' = l_1^1, l_2^2, l_2^3 \dots, l_1^n$ **Same procedure for a''**

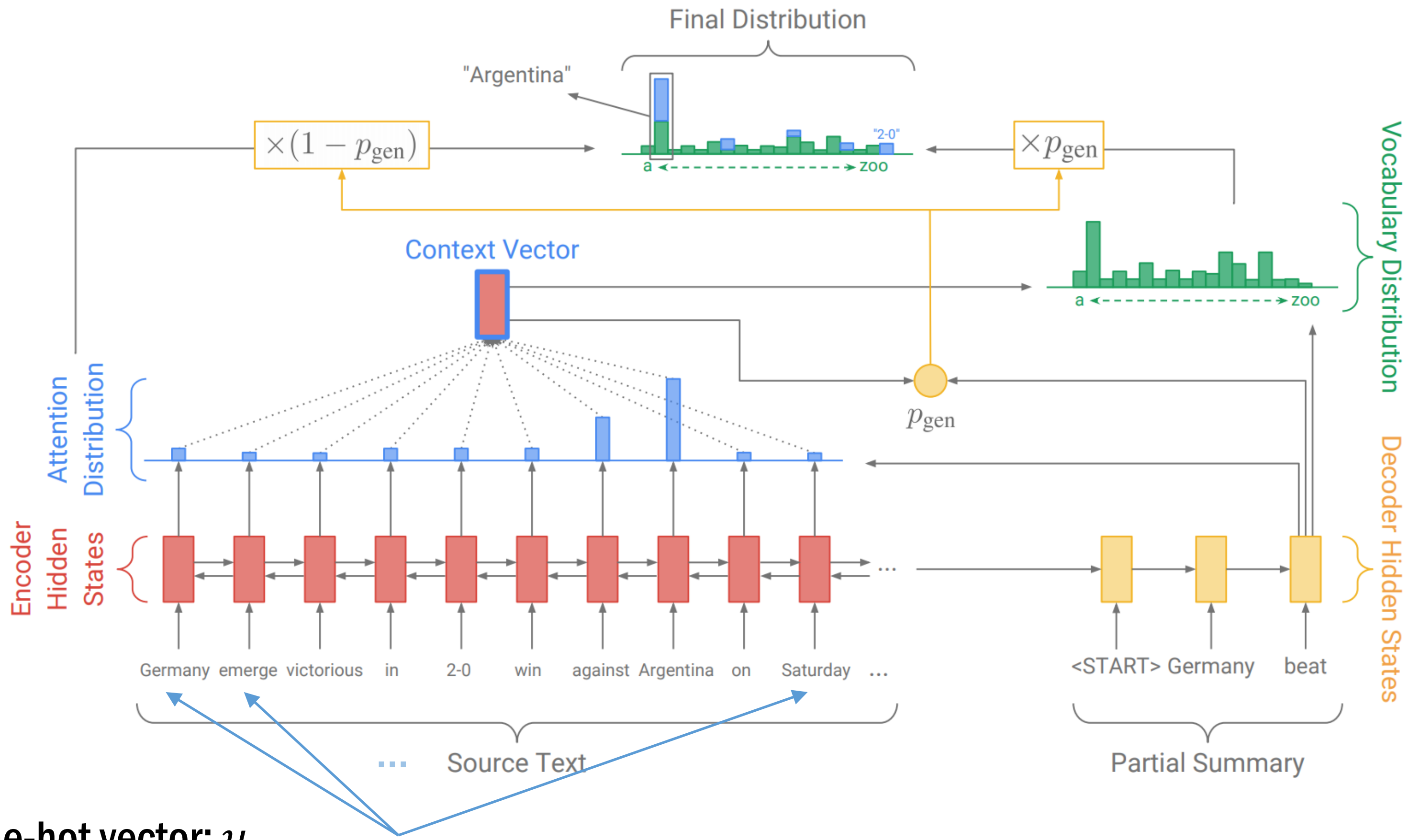
  - a' and a'' can convey two topics
  - Add (a', u_{t_1}, s_1) and (a'', u_{t_2}, s_2) into the **final** dataset



1. Randomly pick (a_1, u_{t_1}, s_1) and (a_2, u_{t_2}, s_2) from the intermediate dataset such that $t_1 \neq t_2$.
2. Make a new article a' by sequentially picking up lines from a_1 and a_2. Each addition of a new line is done by randomly selecting one of a_1 or a_2 and extracting out a new line from the beginning of it. This ensures that the lines from a_1 occur in the same order in a' as originally in a_1, and the same thing is true for a_2 too. This ensures that the sequential flow of content is retained in the merger.
3. Add (a', u_{t_1}, s_1) to the final dataset.
4. Repeat step 2 to get a new article a'' and add (a'', u_{t_2}, s_2) to the final dataset.
5. Discard (a_1, u_{t_1}, s_1) and (a_2, u_{t_2}, s_2) from the intermediate dataset.
6. Repeat steps $1 - 5$ until the entire intermediate dataset is exhausted or all remaining instances in it have the same topic.

Method

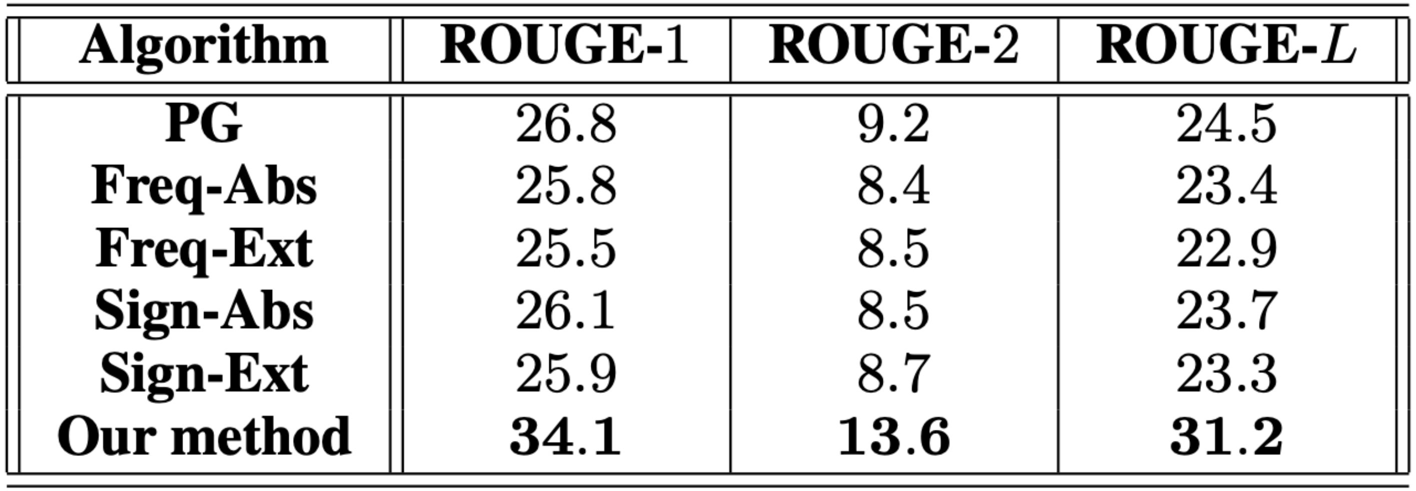
- **Topic aware** pointer-generator network



Topic one-hot vector: u_t

Experiment

- Performance on the created dataset

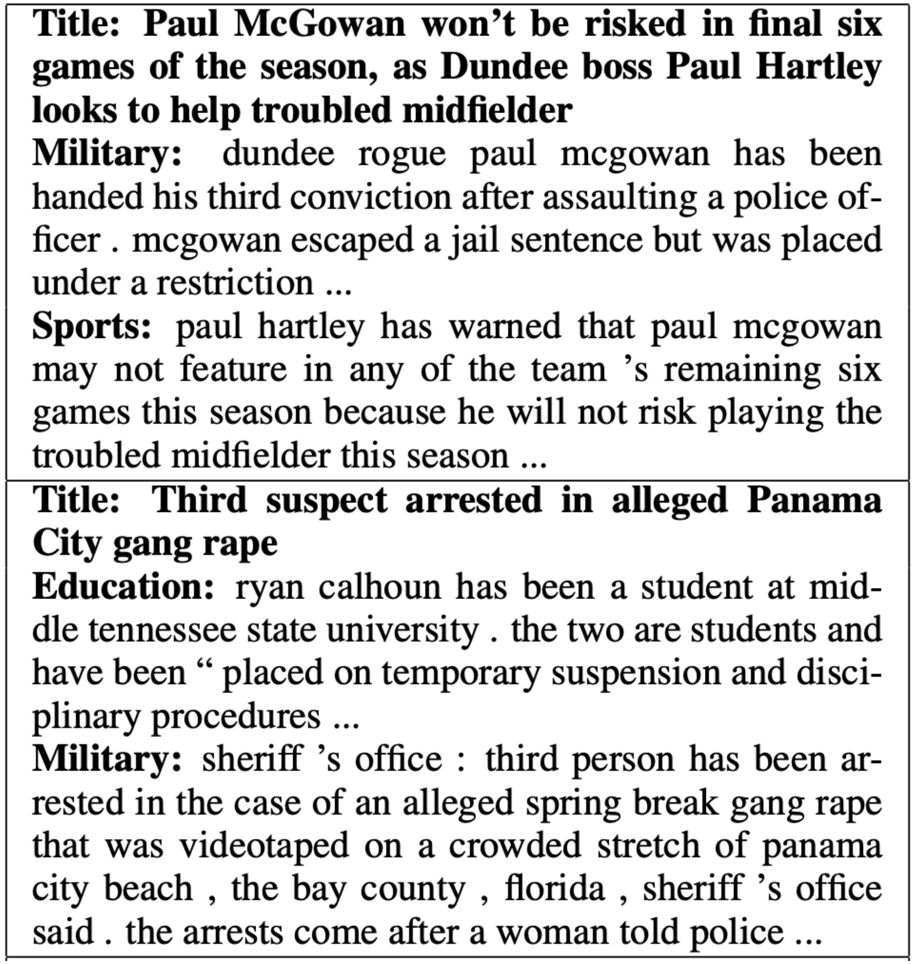


Algorithm	ROUGE-1	ROUGE-2	ROUGE-L
PG	26.8	9.2	24.5
Freq-Abs	25.8	8.4	23.4
Freq-Ext	25.5	8.5	22.9
Sign-Abs	26.1	8.5	23.7
Sign-Ext	25.9	8.7	23.3
Our method	**34.1**	**13.6**	**31.2**

Table 3: ROUGE F1 scores obtained by various methods on the final test set

Experiment

- Performance on multi-topic articles
  - Case study:



Title: Paul McGowan won’t be risked in final six games of the season, as Dundee boss Paul Hartley looks to help troubled midfielder

Military: dundee rogue paul mcgowan has been handed his third conviction after assaulting a police officer . mcgowan escaped a jail sentence but was placed under a restriction ...

Sports: paul hartley has warned that paul mcgowan may not feature in any of the team ’s remaining six games this season because he will not risk playing the troubled midfielder this season ...

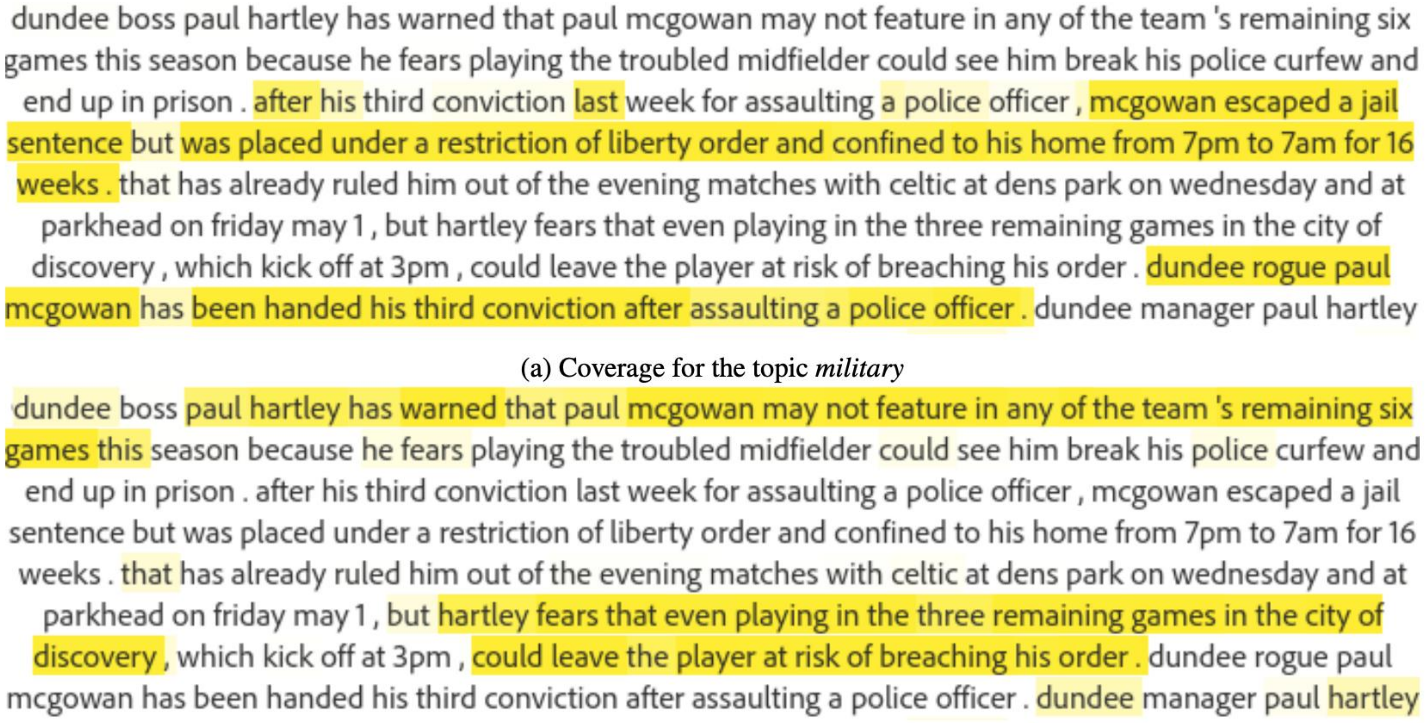
Title: Third suspect arrested in alleged Panama City gang rape

Education: ryan calhoun has been a student at middle tennessee state university . the two are students and have been “ placed on temporary suspension and disciplinary procedures ...

Military: sheriff ’s office : third person has been arrested in the case of an alleged spring break gang rape that was videotaped on a crowded stretch of panama city beach , the bay county , florida , sheriff ’s office said . the arrests come after a woman told police ...

Experiment

- Performance on multi-topic articles
  - Attention coverage visualization:



dundee boss paul hartley has warned that paul mcgowan may not feature in any of the team 's remaining six games this season because he fears playing the troubled midfielder could see him break his police curfew and end up in prison . after his third conviction last week for assaulting a police officer , mcgowan escaped a jail sentence but was placed under a restriction of liberty order and confined to his home from 7pm to 7am for 16 weeks . that has already ruled him out of the evening matches with celtic at dens park on wednesday and at parkhead on friday may 1 , but hartley fears that even playing in the three remaining games in the city of discovery , which kick off at 3pm , could leave the player at risk of breaching his order . dundee rogue paul mcgowan has been handed his third conviction after assaulting a police officer . dundee manager paul hartley

(a) Coverage for the topic *military*

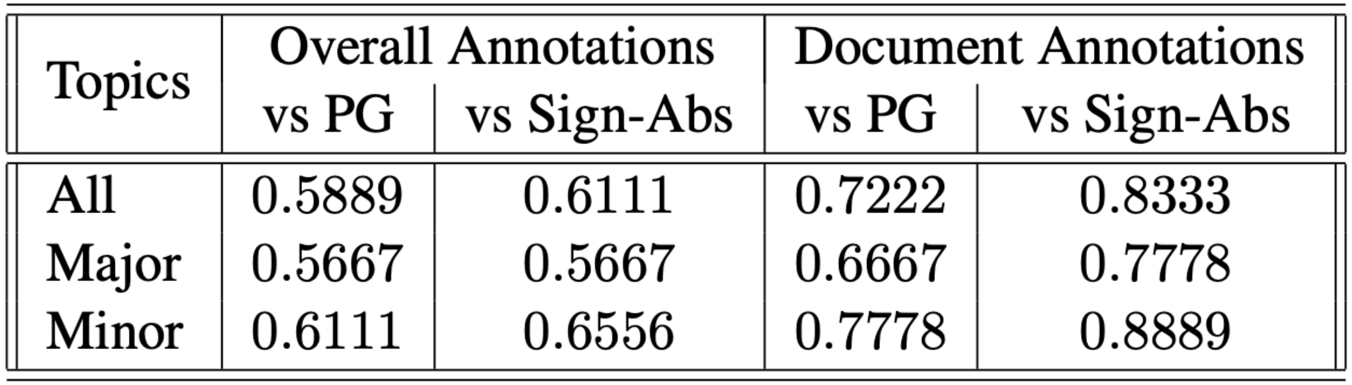
dundee boss paul hartley has warned that paul mcgowan may not feature in any of the team 's remaining six games this season because he fears playing the troubled midfielder could see him break his police curfew and end up in prison . after his third conviction last week for assaulting a police officer , mcgowan escaped a jail sentence but was placed under a restriction of liberty order and confined to his home from 7pm to 7am for 16 weeks . that has already ruled him out of the evening matches with celtic at dens park on wednesday and at parkhead on friday may 1 , but hartley fears that even playing in the three remaining games in the city of discovery , which kick off at 3pm , could leave the player at risk of breaching his order . dundee rogue paul mcgowan has been handed his third conviction after assaulting a police officer . dundee manager paul hartley

(b) Coverage for the topic *sports*

Figure 1: Variation in the attention coverage while summarizing an article for different topics

Experiment

- Human evaluation of performance



Topics	Overall Annotations		Document Annotations	
	vs PG	vs Sign-Abs	vs PG	vs Sign-Abs
All	0.5889	0.6111	0.7222	0.8333
Major	0.5667	0.5667	0.6667	0.7778
Minor	0.6111	0.6556	0.7778	0.8889

Table 6: Evaluation of summaries of the proposed approach against Pointer Generator Framework and Topic Signature based Summarizer by human annotators

Conclusion

- Key contributions:
  - **New task**: Summarize a **single** document into **multiple** summaries with **different topics** of interest
  - **A novel approach** to artificially create a topic-centric training corpus

- One inspiration
  - *“Sometimes when the target dataset is unavailable, we can consider create an artificial dataset using existing data”*