THE UNIVERSITY OF MELBOURNE

Comp90042 Workshop Week 12

May 2022

Table of Contents

1. Summarisation

2. Ethics

Table of Contents

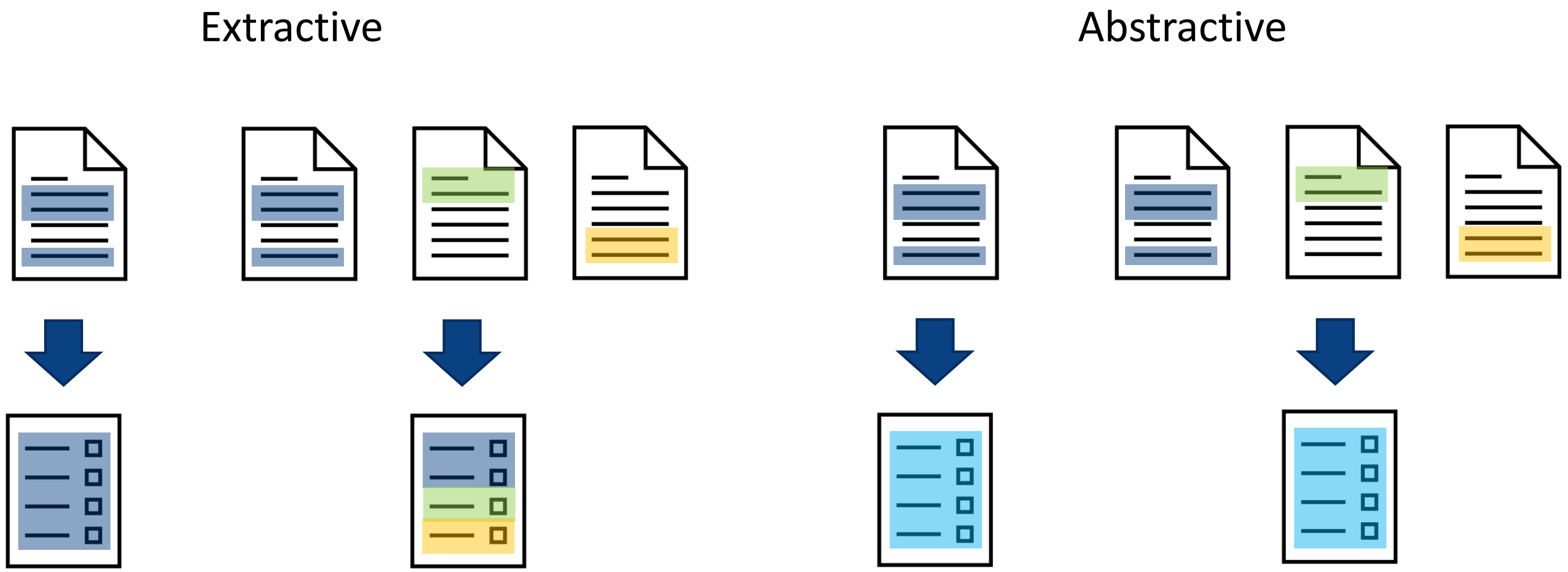
1. Summarisation

2. Ethics

Summarisation

1. What is **Summarisation**?

Distilling information from an input text to capture the key points or messages



Extractive

Abstractive

Summarisation

2. Datasets in **Summarisation**

English: CNNDM, XSUM, NYT, Gigaword, Multi-News

Chinese: LCSTS, CLSTS, TTNews

Indonesian: Liputan6

Multilingual: MLSUM, WikiLingua

Good resource:
github.com/xcfcode/Summarization-Papers

Summarisation

3. Content Selection in Extractive **Summarisation**

- TF-IDF
- Log likelihood ratio
- Sentence centrality
- RST Parsing
- Neural approaches …

Summarisation

4. Log-Likelihood Ratio

Measures a statistical discrepancy of a word between two sources

We can use log likelihood ratio to **measure how salient a word is to a document by comparing its frequency statistics to that of a background corpus**.

This is useful for summarisation because it provides an unsupervised method to extract salient words/passages from a document.

Summarisation

4. Log-Likelihood Ratio

Intuition: a word is salient if its probability in the **input corpus** is very different to a **background corpus**

$$-2log(\frac{\binom{N_d}{F_d} p^{F_d}(1-p)^{N_d-F_d} \times \binom{N_b}{F_b} p^{F_b}(1-p)^{N_b-F_b}}{\binom{N_d}{F_d} p_d^{F_d}(1-p_d)^{N_d-F_d} \times \binom{N_b}{F_b} p_b^{F_b}(1-p_b)^{N_b-F_b}}$$

Loglikelihood ratio

What is the log likelihood ratio of word w in document d if: F_d(w) = 1, F_b(w) = 20, N_d = 30, and N_b = 4000, where b denotes the background corpus, F the frequency and N the total number of word tokens.

$$-2log(\frac{\binom{N_d}{F_d}p^{F_d}(1-p)^{N_d-F_d}\times\binom{N_b}{F_b}p^{F_b}(1-p)^{N_b-F_b}}{\binom{N_d}{F_d}p_d^{F_d}(1-p_d)^{N_d-F_d}\times\binom{N_b}{F_b}p_b^{F_b}(1-p_b)^{N_b-F_b}}$$

$$p = \frac{1+20}{30+4000} = \frac{21}{4030} \qquad p_d = \frac{1}{30} \qquad p_b = \frac{20}{4000} = \frac{1}{200}$$

$$\mathrm{LLR(w)} = -2\log\left(\frac{\binom{30}{1}p^1(1-p)^{29}\times\binom{4000}{20}p^{20}(1-p)^{3980}}{\binom{30}{1}p_d^1(1-p_d)^{29}\times\binom{4000}{20}p_b^{20}(1-p_b)^{3980}}\right)$$

$$\mathrm{LLR(w)} = -2\log\left(\frac{0.134320\times 0.0875}{0.374132\times 0.089058}\right)$$

$$\mathrm{LLR(w)} = -2\log(0.3527) = -2 * (-1.04203)$$

$$\mathrm{LLR(w)} = 2.08407$$

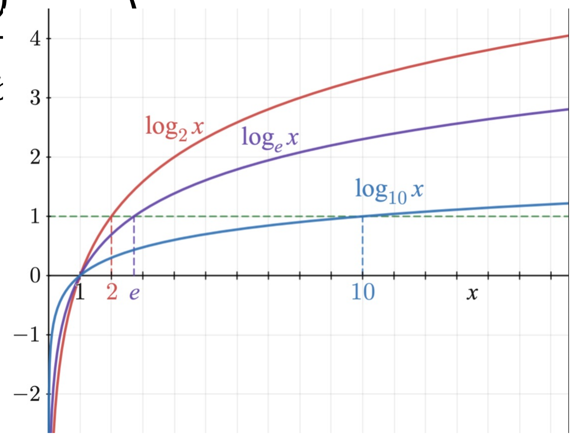
Loglikelihood ratio

What is the log likelihood ratio of word w in document d if: F_d(w) = 1, F_b(w) = 20, N_d = 30, and N_b = 4000, where b denotes the background corpus, F the frequency and N the total number of word tokens.

$$-2log(\frac{\binom{N_d}{F_d}p^{F_d}(1-p)^{N_d-F_d}\times\binom{N_b}{F_b}p^{F_b}(1-p)^{N_b-F_b}}{\binom{N_d}{F_d}p_d^{F_d}(1-p_d)^{N_d-F_d}\times\binom{N_b}{F_b}p_b^{F_b}(1-p_b)^{N_b-F_b}}$$

$$p=\frac{1+20}{30+4000}=\frac{21}{4030} \qquad p_d=\frac{1}{30} \qquad p_b=\frac{20}{4000}=\frac{1}{200}$$

$$\mathrm{LLR(w)}=-2\log\Bigg(\frac{\binom{30}{1}p^1(1-p)^{29}\times\binom{4000}{20}p^{20}(1-p)^{3980}}{\binom{30}{1}p_d^1(1-p_d)^{29}\times\binom{4000}{20}p_b^{20}(1-p_b}$$

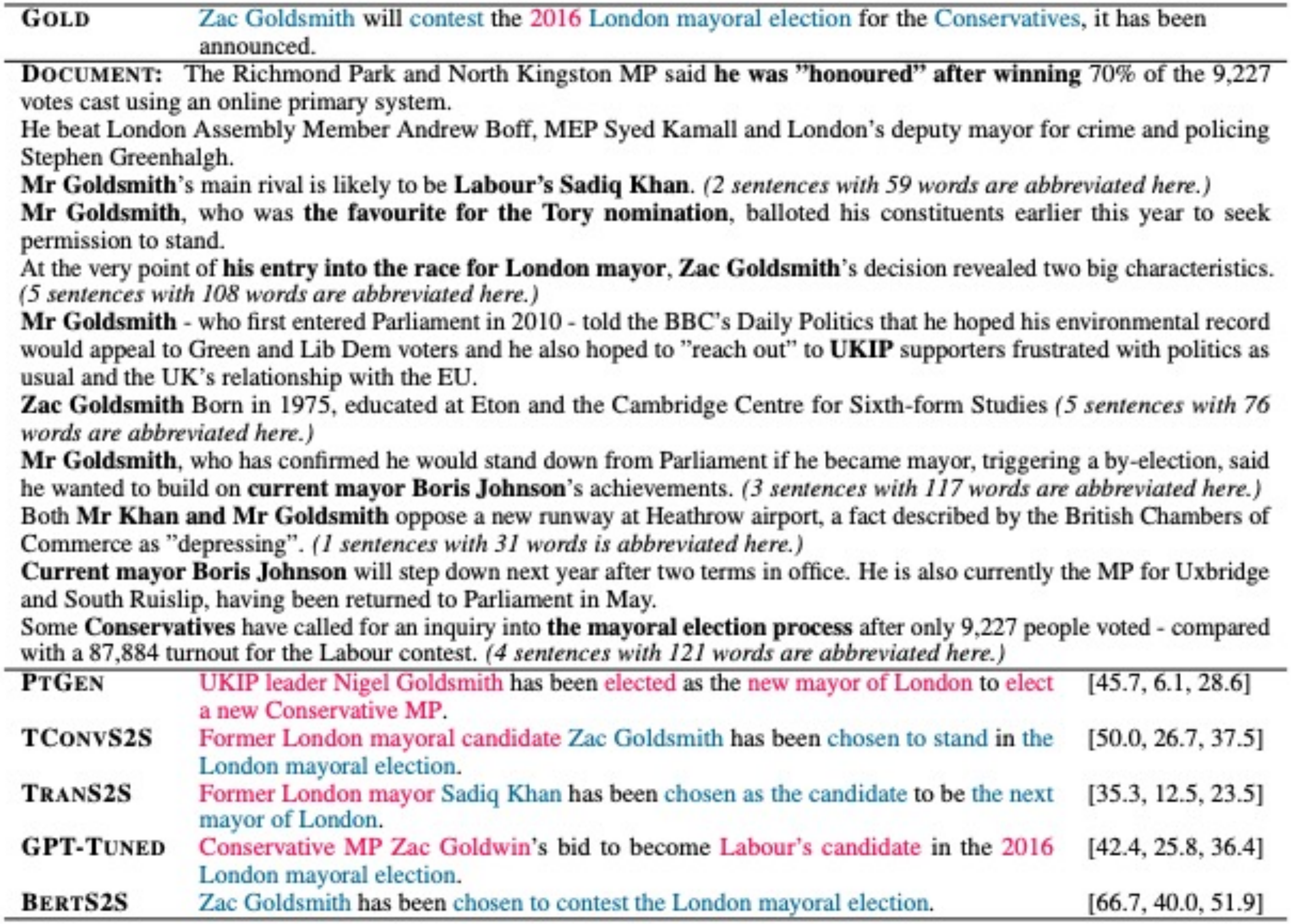


$$\mathrm{LLR(w)}=\ -2\log\left(\frac{0.134320\times0.0875}{0.374132\times0.089058}\right)$$

$$\mathrm{LLR(w)}=-2\log(0.3527)=-2*(-1.04203)$$

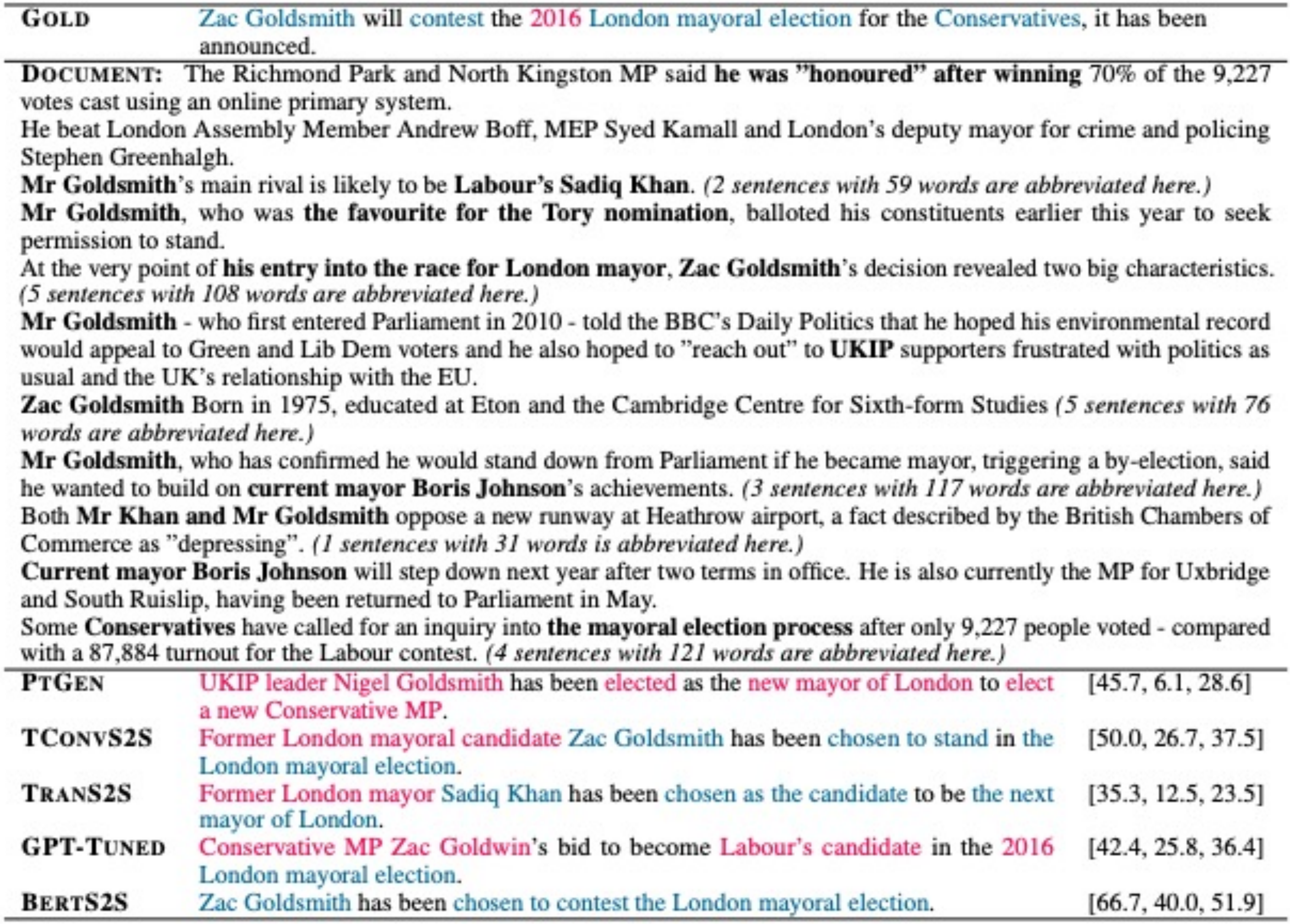
$$\mathrm{LLR(w)}=2.08407$$

What’s wrong here?



GOLD	Zac Goldsmith will contest the 2016 London mayoral election for the Conservatives, it has been announced.	
DOCUMENT:	The Richmond Park and North Kingston MP said **he was "honoured" after winning** 70% of the 9,227 votes cast using an online primary system. He beat London Assembly Member Andrew Boff, MEP Syed Kamall and London’s deputy mayor for crime and policing Stephen Greenhalgh. **Mr Goldsmith**’s main rival is likely to be **Labour’s Sadiq Khan**. *(2 sentences with 59 words are abbreviated here.)* **Mr Goldsmith**, who was **the favourite for the Tory nomination**, balloted his constituents earlier this year to seek permission to stand. At the very point of **his entry into the race for London mayor**, **Zac Goldsmith**’s decision revealed two big characteristics. *(5 sentences with 108 words are abbreviated here.)* **Mr Goldsmith** - who first entered Parliament in 2010 - told the BBC’s Daily Politics that he hoped his environmental record would appeal to Green and Lib Dem voters and he also hoped to "reach out" to **UKIP** supporters frustrated with politics as usual and the UK’s relationship with the EU. **Zac Goldsmith** Born in 1975, educated at Eton and the Cambridge Centre for Sixth-form Studies *(5 sentences with 76 words are abbreviated here.)* **Mr Goldsmith**, who has confirmed he would stand down from Parliament if he became mayor, triggering a by-election, said he wanted to build on **current mayor Boris Johnson**’s achievements. *(3 sentences with 117 words are abbreviated here.)* Both **Mr Khan and Mr Goldsmith** oppose a new runway at Heathrow airport, a fact described by the British Chambers of Commerce as "depressing". *(1 sentences with 31 words is abbreviated here.)* **Current mayor Boris Johnson** will step down next year after two terms in office. He is also currently the MP for Uxbridge and South Ruislip, having been returned to Parliament in May. Some **Conservatives** have called for an inquiry into **the mayoral election process** after only 9,227 people voted - compared with a 87,884 turnout for the Labour contest. *(4 sentences with 121 words are abbreviated here.)*	
PTGEN	UKIP leader Nigel Goldsmith has been elected as the new mayor of London to elect a new Conservative MP.	[45.7, 6.1, 28.6]
TCONVS2S	Former London mayoral candidate Zac Goldsmith has been chosen to stand in the London mayoral election.	[50.0, 26.7, 37.5]
TRANS2S	Former London mayor Sadiq Khan has been chosen as the candidate to be the next mayor of London.	[35.3, 12.5, 23.5]
GPT-TUNED	Conservative MP Zac Goldwin’s bid to become Labour’s candidate in the 2016 London mayoral election.	[42.4, 25.8, 36.4]
BERTS2S	Zac Goldsmith has been chosen to contest the London mayoral election.	[66.7, 40.0, 51.9]

What’s wrong here?



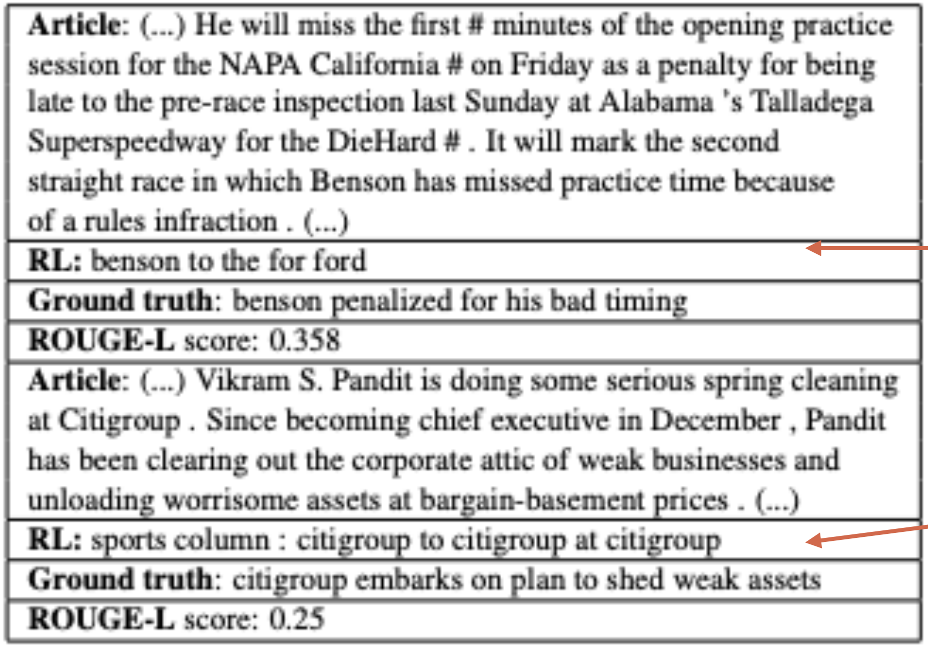
GOLD	Zac Goldsmith will contest the 2016 London mayoral election for the Conservatives, it has been announced.	
DOCUMENT:	The Richmond Park and North Kingston MP said **he was ”honoured” after winning** 70% of the 9,227 votes cast using an online primary system. He beat London Assembly Member Andrew Boff, MEP Syed Kamall and London’s deputy mayor for crime and policing Stephen Greenhalgh. **Mr Goldsmith**’s main rival is likely to be **Labour’s Sadiq Khan**. *(2 sentences with 59 words are abbreviated here.)* **Mr Goldsmith**, who was **the favourite for the Tory nomination**, balloted his constituents earlier this year to seek permission to stand. At the very point of **his entry into the race for London mayor**, **Zac Goldsmith**’s decision revealed two big characteristics. *(5 sentences with 108 words are abbreviated here.)* **Mr Goldsmith** - who first entered Parliament in 2010 - told the BBC’s Daily Politics that he hoped his environmental record would appeal to Green and Lib Dem voters and he also hoped to ”reach out” to **UKIP** supporters frustrated with politics as usual and the UK’s relationship with the EU. **Zac Goldsmith** Born in 1975, educated at Eton and the Cambridge Centre for Sixth-form Studies *(5 sentences with 76 words are abbreviated here.)* **Mr Goldsmith**, who has confirmed he would stand down from Parliament if he became mayor, triggering a by-election, said he wanted to build on **current mayor Boris Johnson**’s achievements. *(3 sentences with 117 words are abbreviated here.)* Both **Mr Khan and Mr Goldsmith** oppose a new runway at Heathrow airport, a fact described by the British Chambers of Commerce as ”depressing”. *(1 sentences with 31 words is abbreviated here.)* **Current mayor Boris Johnson** will step down next year after two terms in office. He is also currently the MP for Uxbridge and South Ruislip, having been returned to Parliament in May. Some **Conservatives** have called for an inquiry into **the mayoral election process** after only 9,227 people voted - compared with a 87,884 turnout for the Labour contest. *(4 sentences with 121 words are abbreviated here.)*	
PTGEN	UKIP leader Nigel Goldsmith has been elected as the new mayor of London to elect a new Conservative MP.	[45.7, 6.1, 28.6]
TCONVS2S	Former London mayoral candidate Zac Goldsmith has been chosen to stand in the London mayoral election.	[50.0, 26.7, 37.5]
TRANS2S	Former London mayor Sadiq Khan has been chosen as the candidate to be the next mayor of London.	[35.3, 12.5, 23.5]
GPT-TUNED	Conservative MP Zac Goldwin’s bid to become Labour’s candidate in the 2016 London mayoral election.	[42.4, 25.8, 36.4]
BERTS2S	Zac Goldsmith has been chosen to contest the London mayoral election.	[66.7, 40.0, 51.9]

Hallucinations

Copying mechanism, Pointer-Generator networks

ROUGE (??)

$$ROUGE_N = \frac{\sum_{S \in Ref} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in Ref} \sum_{gram_n \in S} Count(gram_n)}$$



Article: (...) He will miss the first # minutes of the opening practice session for the NAPA California # on Friday as a penalty for being late to the pre-race inspection last Sunday at Alabama 's Talladega Superspeedway for the DieHard # . It will mark the second straight race in which Benson has missed practice time because of a rules infraction . (...)
RL: benson to the for ford
Ground truth: benson penalized for his bad timing
ROUGE-L score: 0.358
Article: (...) Vikram S. Pandit is doing some serious spring cleaning at Citigroup . Since becoming chief executive in December , Pandit has been clearing out the corporate attic of weak businesses and unloading worrisome assets at bargain-basement prices . (...)
RL: sports column : citigroup to citigroup at citigroup
Ground truth: citigroup embarks on plan to shed weak assets
ROUGE-L score: 0.25

Some problems with ROUGE:

1. Stopwords can contribute a lot

2. The same word repeated

Table of Contents

1. Summarisation

2. Ethics

Ethics

You have an idea of building a comment generation system for news articles. Training data can be created by mining news articles and their comments on the web. What are the ethical implications of such application?

Ethical implications:

Bias – Where is the data coming from? How can I avoid racism, misogyny, etc.?

Dual use – What is the primary use of my application? How could it be also used?

Privacy – Do I have the rights from users to manipulate their data?

Environmental concerns – What kind of model am I using?

THE UNIVERSITY OF
MELBOURNE

Thank you