Don't Give Me the Details, Just the Summary!
Topic-Aware Convolutional Neural Networks for Extreme Summarization

Shashi Narayan   **Shay B. Cohen**   **Mirella Lapata**

Institute for Language, Cognition and Computation
School of Informatics

THE UNIVERSITY OF EDINBURGH

Edinburgh NLP
University of Edinburgh
Natural Language Processing

Neural Summarization is a Hot Topic!

Grand total of 104 papers in ACL conferences since 2016!

Story highlights

North Carolina reports first case of mosquito-borne virus called chikungunya

Chikungunya is primarily found in Africa, East Asia and the Caribbean islands

Virus is not deadly, but it can be painful, with symptoms lasting for weeks

A debilitating, mosquito-borne virus called chikungunya has made its way to North Carolina, health officials say. It's the state's first reported case of the virus.

The patient was likely infected in the Caribbean, according to the Forsyth County Department of Public Health. Chikungunya is primarily found in Africa, East Asia and the Caribbean islands, but the Centers for Disease Control and Prevention has been watching the virus,+ for fear that it could take hold in the United States -- much like West Nile did more than a decade ago.

The virus, which can cause joint pain and arthritis-like symptoms, has been on the U.S. public health radar for some time. About 25 to 28 infected travelers bring it to the United States each year, said Roger Nasci, chief of the CDC's Arboviral Disease Branch in the Division of Vector-Borne Diseases.

3

CNN and DailyMail Datasets

+ Large-scale datasets (92K and 220K documents)

+ Both articles and summaries are written by human (journalists)

+ News highlight generation is a natural summarization task

CNN and DailyMail Datasets

+ Large-scale datasets (92K and 220K documents)

+ Both articles and summaries are written by human (journalists)

+ News highlight generation is a natural summarization task

− Encourage models to be extractive

− Extractive methods outperform abstractive ones on these datasets (Narayan et al. 2018, Zhang et al. 2018)

− Difficult to do better than LEAD baseline

Human-written Summaries are Extractive!

Article: Queen Victoria spent her holidays in Osborne House on the Isle of Wight. …
She would travel to Portsmouth by train and then by ferry to Ryde. From Ryde there was a railway line that passed not far from Osborne House but the nearest station was at Wootton, more than two miles from the property. So, in 1875, a station was built at Whippingham, the closest point on the line to Osborne House – just to serve the Royal residence. … The building is now a five-bedroom family home, currently on the market for £625,000, while the track has become a cycle path. …

Human-written Summary (Story highlights)

- Queen Victoria's holiday residence was Osborne House on the Isle of Wight
- But her journeys there involved train and ferry ride and then another train ride to a station more than two miles from the property
- In 1875, a station was built at Whippingham just to serve Royal residence
- Building is now a five-bedroom home, currently on the market for £625,000

Human-written Summaries are Extractive!

Article: Queen Victoria spent her holidays in Osborne House on the Isle of Wight. … She would travel to Portsmouth by train and then by ferry to Ryde. From Ryde there was a railway line that passed not far from Osborne House but the nearest station was at Wootton, more than two miles from the property. So, in 1875, a station was built at Whippingham, the closest point on the line to Osborne House – just to serve the Royal residence. … The building is now a five-bedroom family home, currently on the market for £625,000, while the track has become a cycle path. …

Human-written Summary (Story highlights)

- Queen Victoria's holiday residence was Osborne House on the Isle of Wight
- But her journeys there involved train and ferry ride and then another train ride to a station more than two miles from the property
- In 1875, a station was built at Whippingham just to serve Royal residence
- Building is now a five-bedroom home, currently on the market for £625,000

7

Abstractive Summaries are Often Extractive!

Article: Queen Victoria spent her holidays in Osborne House on the Isle of Wight. … She would travel to Portsmouth by train and then by ferry to Ryde. From Ryde there was a railway line that passed not far from Osborne House but the nearest station was at Wootton, more than two miles from the property. So, in 1875, a station was built at Whippingham, the closest point on the line to Osborne House – just to serve the Royal residence. … The building is now a five-bedroom family home, currently on the market for £625,000, while the track has become a cycle path. …

Generated Summary (Pointer-Generator, See et al. 2017)

- Queen Victoria spent her holidays in Osborne House on the Isle of Wight.
- She would travel to Portsmouth by train and then by ferry to ryde.
- Building is now a five-bedroom family home, currently on the market for £625,000.

Our Contributions

❏ **A New Dataset**

    ❏ **Suitable for Abstractive Summarization**

❏ **A New Topic-Aware Convolutional Model**

    ❏ **Suitable for Contextual Understanding and Abstraction**

A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.

Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon.

The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane.

The plane's only two occupants were unharmed, it added.

The Diario de Noticias newspaper quoted an eyewitness who said the plane had been flying at a low altitude over the beach, although he did not realise anything was wrong until other beachgoers began running.

One young witness told Reuters news agency: "I was near the water when I saw the plane. I called my parents, the plane fell on the sand and ran over two people, fatally hurting them and another was injured, I think, but I'm not sure, people were running away.

"The plane is still there, but the ambulances arrived quickly. I think maybe the fuel ran out because I find it weird that it landed on the beach."

BBC

Extreme Summarization

The **XSum Dataset**

BBC

ONE-LINE SUMMARY: *A man and a child have been killed* after a *light* aircraft made an emergency landing on a beach in Portugal.

DOCUMENT: Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon.
The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane.
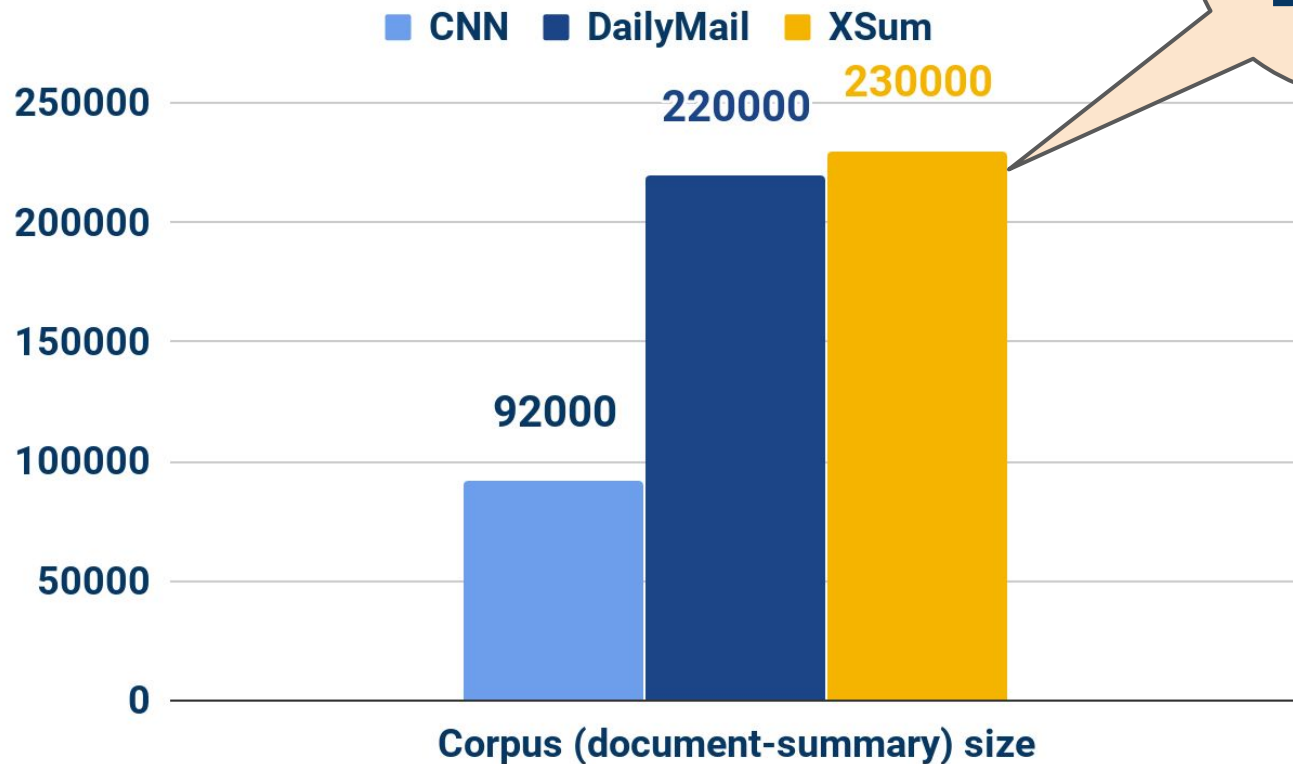[*6 sentences with 139 words are abbreviated from here.*]
Other reports said the victims had been sunbathing when the plane made its emergency landing.
[*Another 4 sentences with 67 words are abbreviated from here.*]
Video footage from the scene carried by local broadcasters showed a small recreational plane parked on the sand, apparently intact and surrounded by beachgoers and emergency workers.
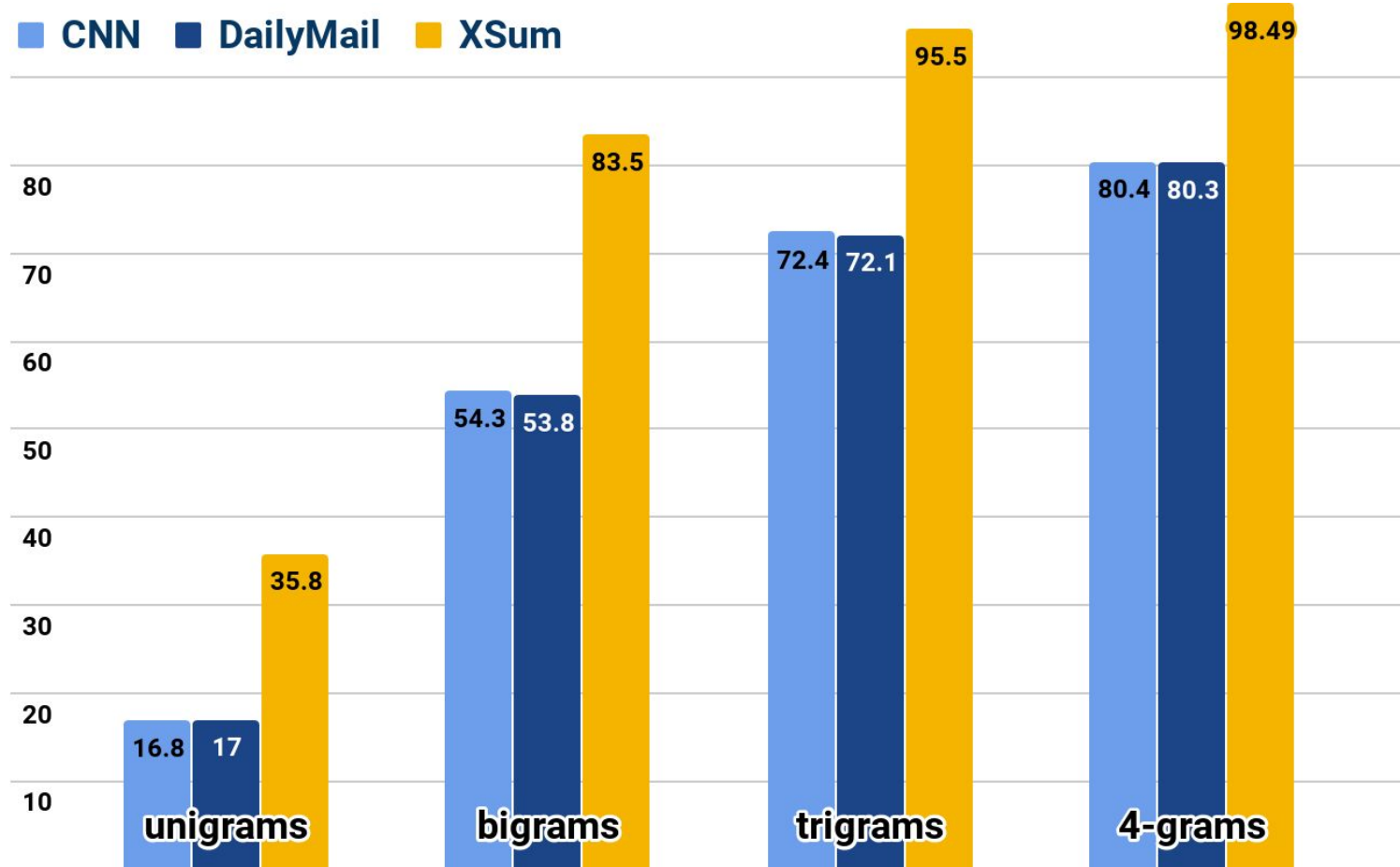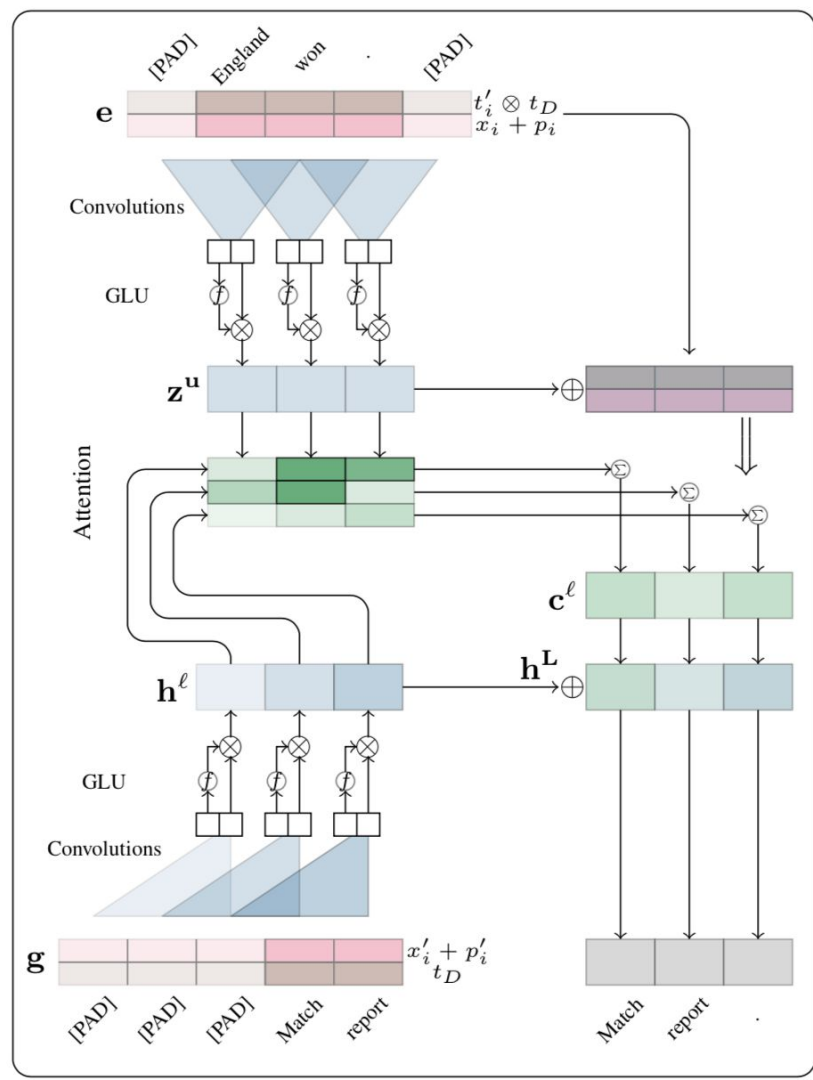[*Last 2 sentences with 19 words are abbreviated.*]

XSum Dataset Size

Our Dataset

Legend: CNN, DailyMail, XSum

- CNN: 92000
- DailyMail: 220000
- XSum: 230000

Corpus (document-summary) size

Percentage of Novel N-grams in Gold Summaries
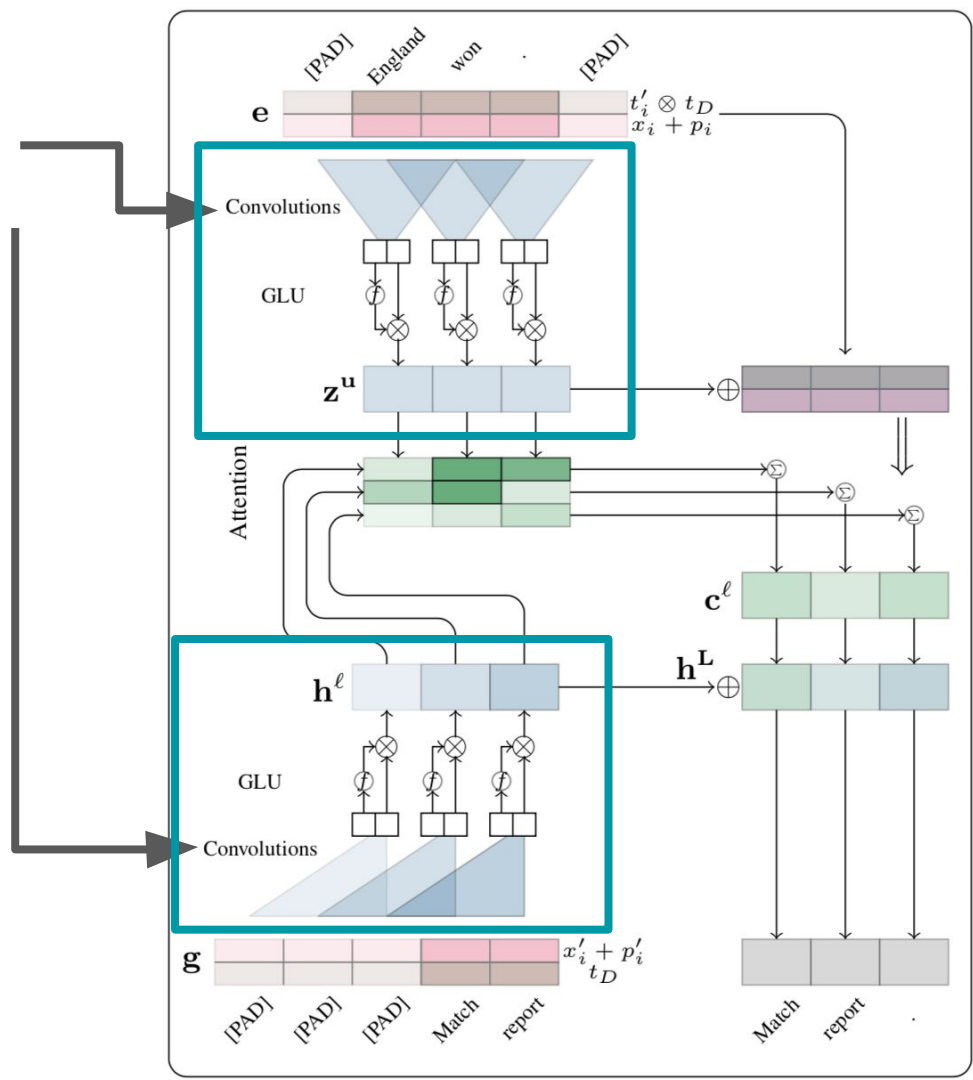

13

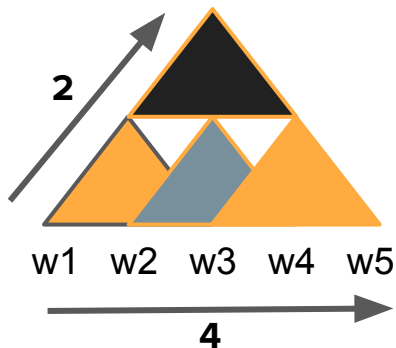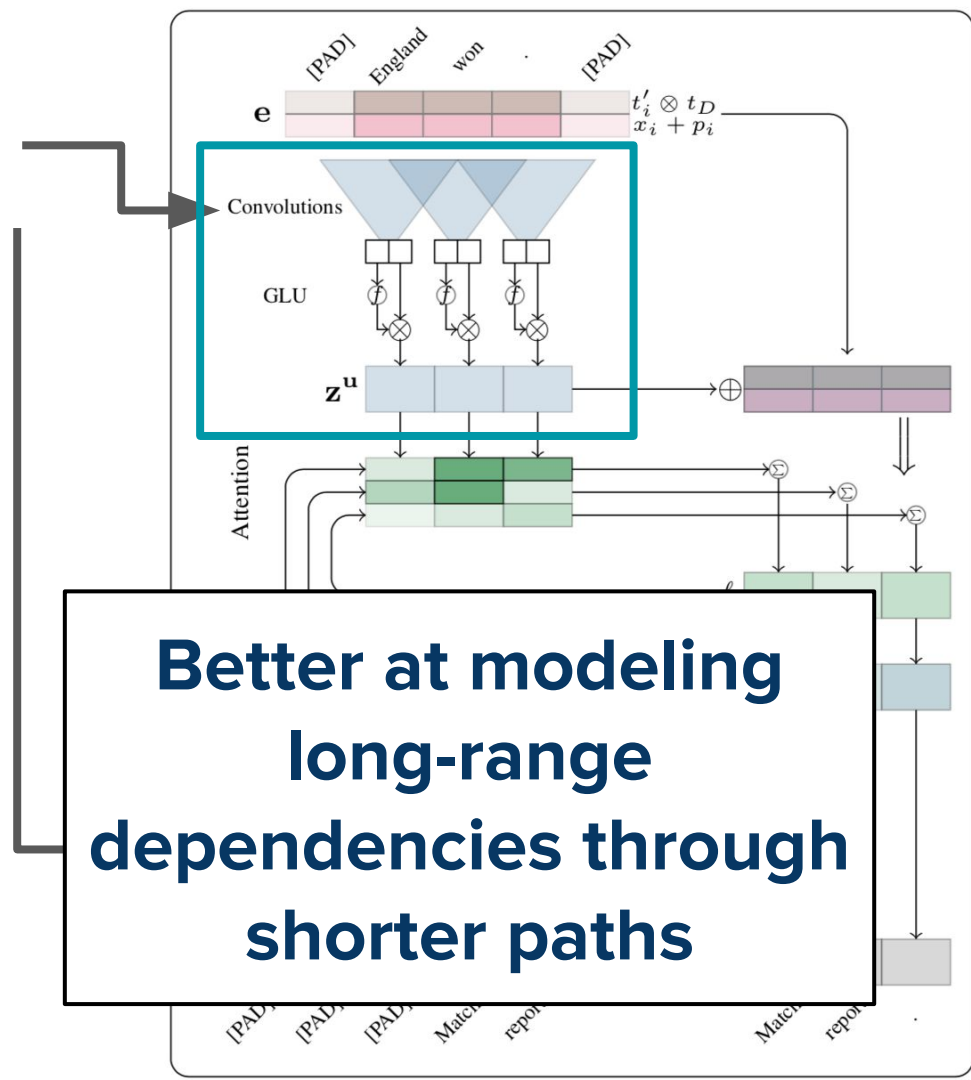# Topic-Aware Convolutional Seq-to-Seq Model for Abstractive Summarization

Multi-layer Convolution Hierarchical Representation

➢ Models document with **stacked convolutional layers**, rather than as a **chain structure**

➢ **Efficient fully convolutional** structure

[Gehring et al, 2017]

Multi-layer Convolution Hierarchical Representation

➢ Models document with **stacked convolutional layers**, rather than as a **chain structure**

➢ **Efficient fully convolutional** structure

w1   w2   w3   w4   w5

2

4

[Gehring et al, 2017]

Better at modeling long-range dependencies through shorter paths

Multi-layer Convolution Hierarchical Representation

➢ Models document with **stacked convolutional layers**, rather than as a **chain structure**

➢ **Efficient fully convolutional** structure

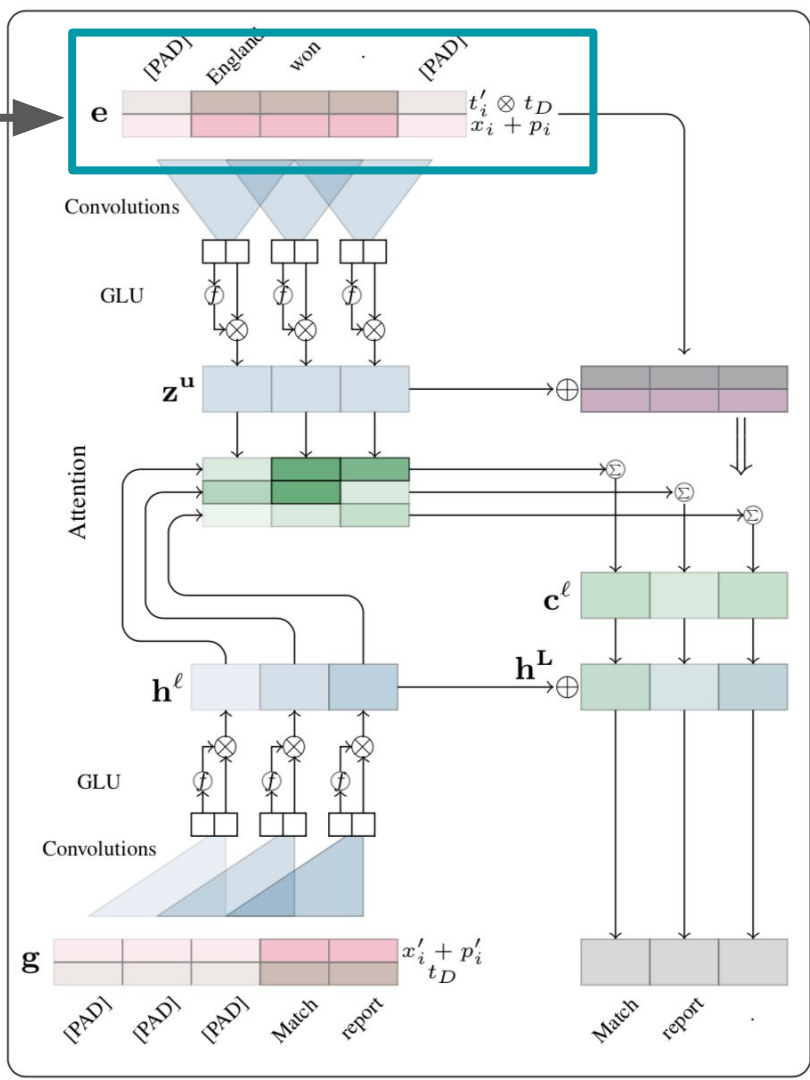➢ **Multi-hop Attention** between Encoder and Decoder

[Gehring et al, 2017]
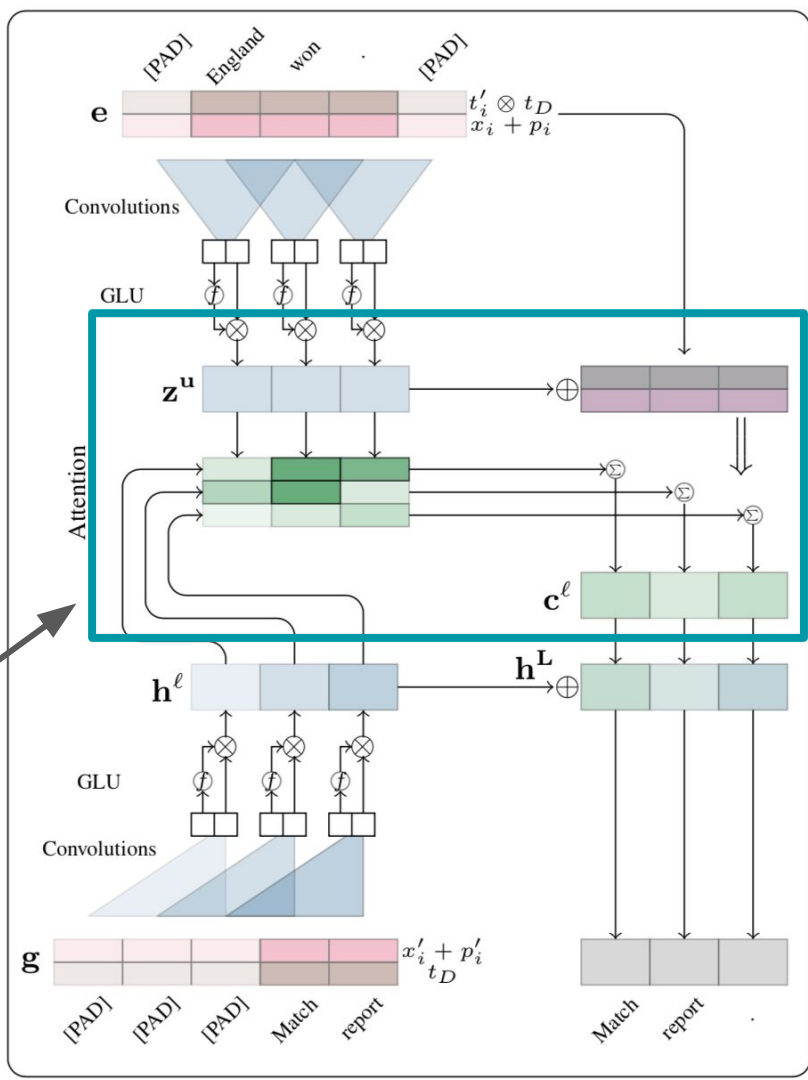
Topic-Sensitive Embeddings

➤ **Encoder Embeddings**
- Word embedding
- Position embedding
- **Word topic distribution**
- **Document topic distribution**

Topic-Sensitive Embeddings

➤ **Encoder Embeddings**
  ○ Word embedding
  ○ Position embedding
  ○ **Word topic distribution**
  ○ **Document topic distribution**

Latent Dirichlet Allocation (Blei et al. 2003) for word and document topic distributions
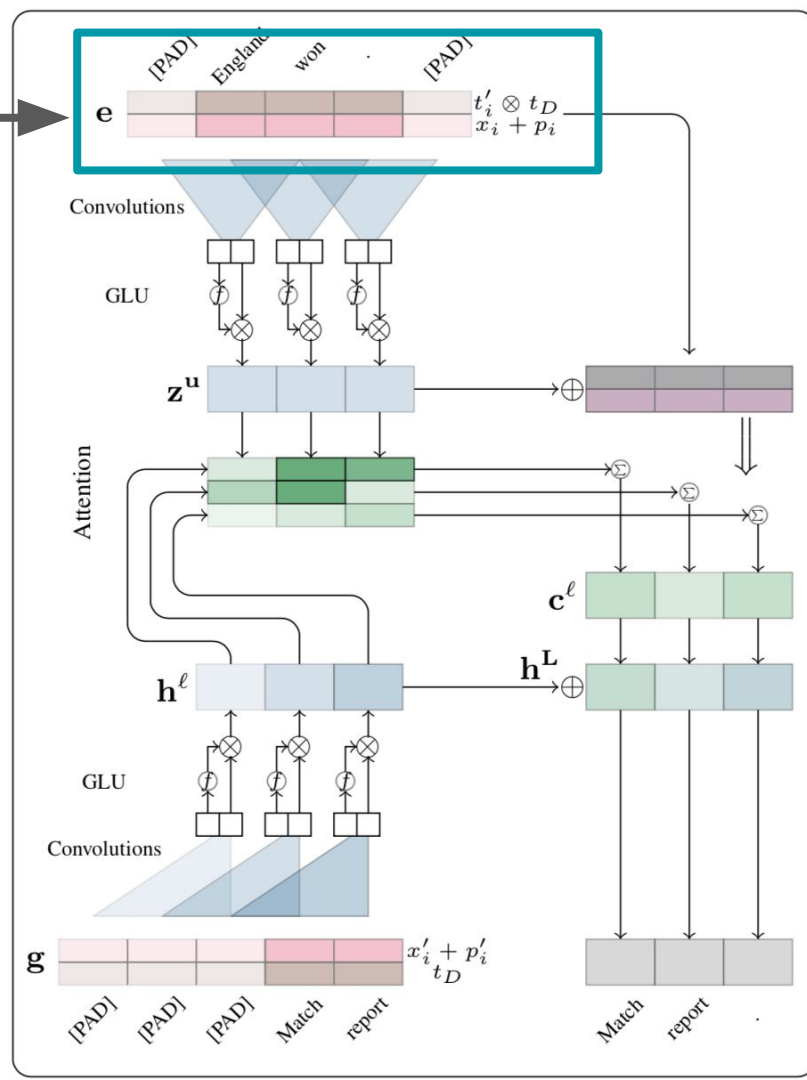
Topic-Sensitive Embeddings

➤ **Encoder Embeddings**
  - Word embedding
  - Position embedding
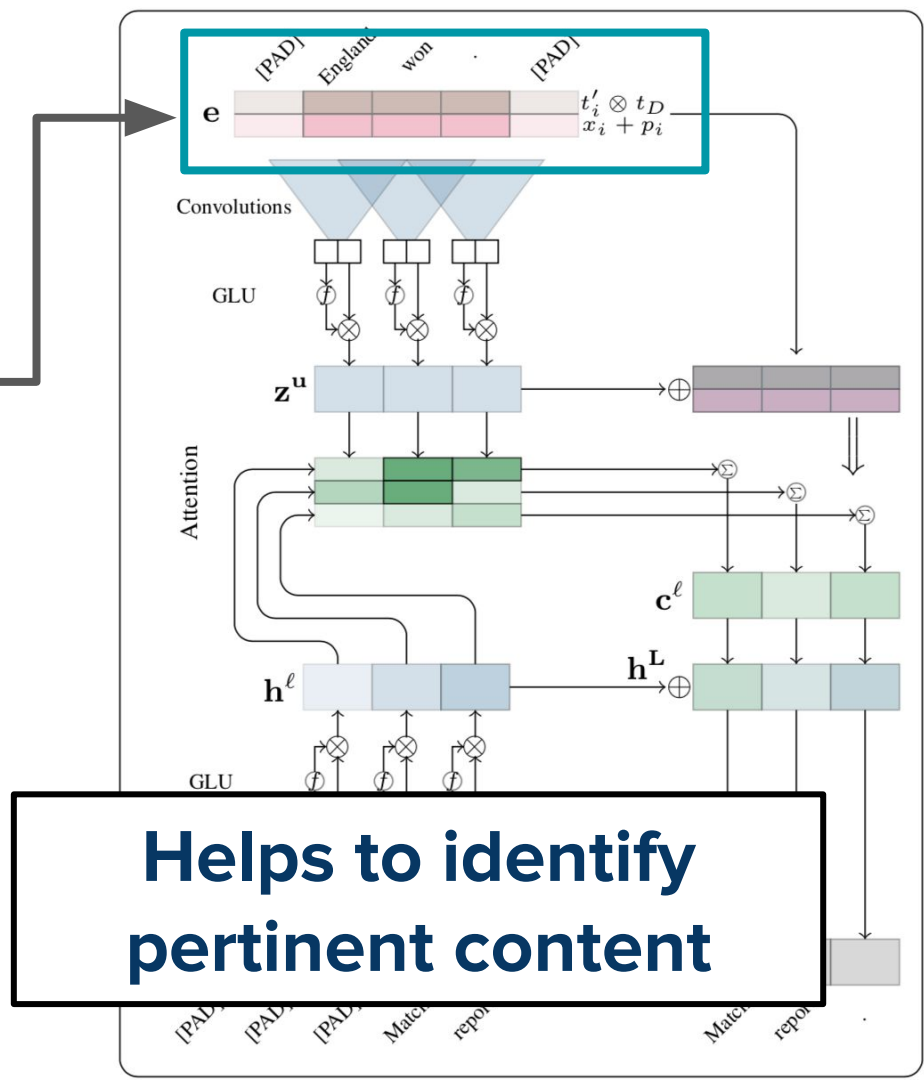 - **Word topic distribution**
  - **Document topic distribution**

Helps to identify pertinent content

Topic-Sensitive Embeddings
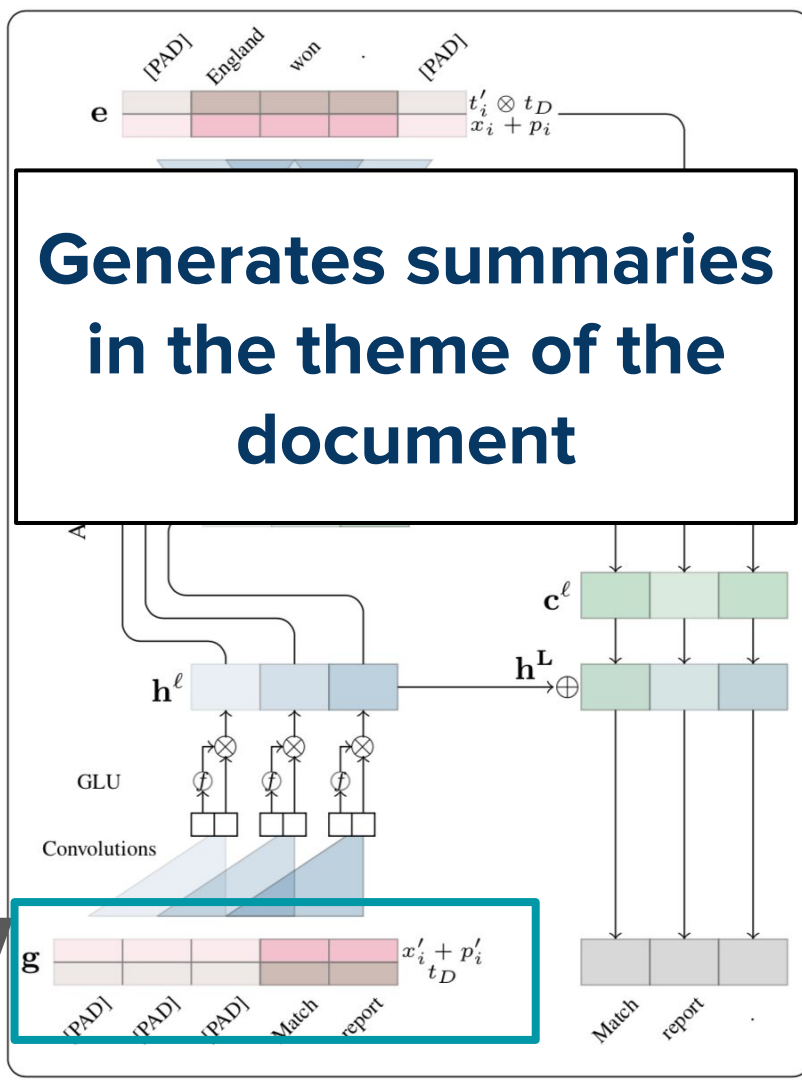
➤ **Decoder Embeddings**
  ○ Word embedding
  ○ Position embedding
  ○ **Document topic distribution**

Topic-Sensitive Embeddings

➢ **Decoder Embeddings**
  ○ Word embedding
  ○ Position embedding
  ○ **Document topic distribution**

Generates summaries in the theme of the document
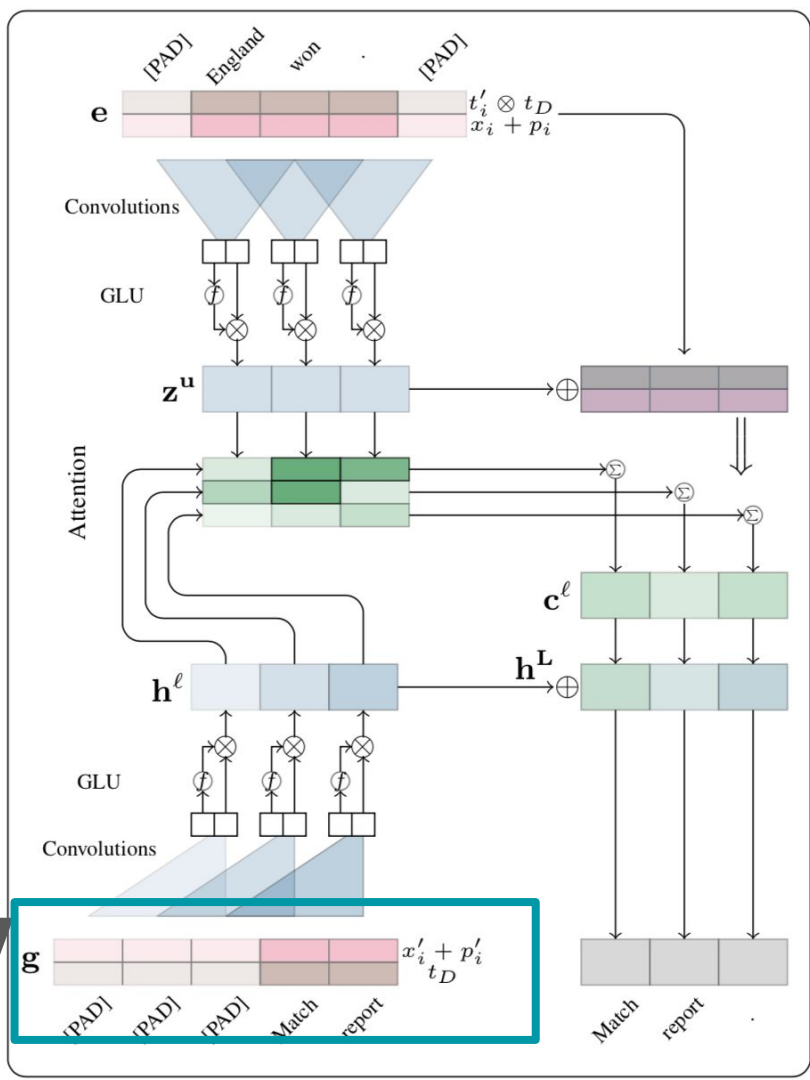
Topic-Sensitive Embeddings

Identify pertinent content and generate summary in the same theme
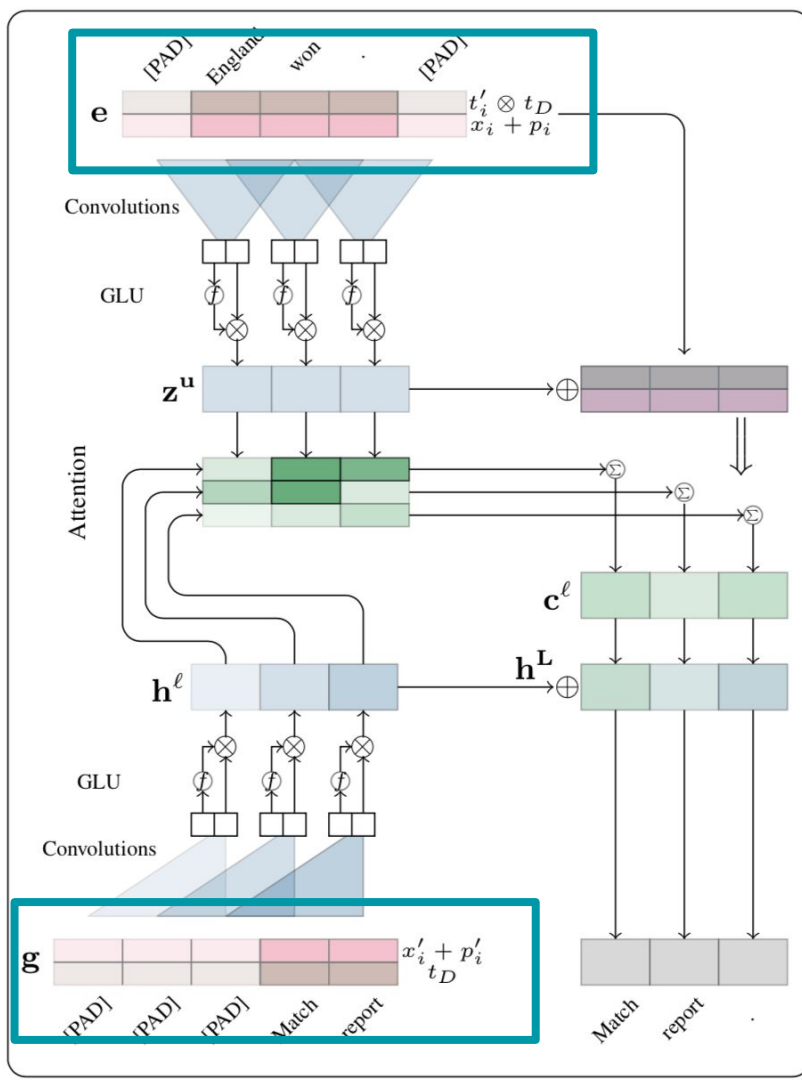
Topic-Aware Convolutional Seq-to-Seq Model

$$p(y_{i+1}|y_i, \ldots, y_i, D, t_D, t')$$

Multi-layer Convolution Hierarchical Representation

Topic-Sensitive Embeddings

Multi-hop Attention between Encoder and Decoder

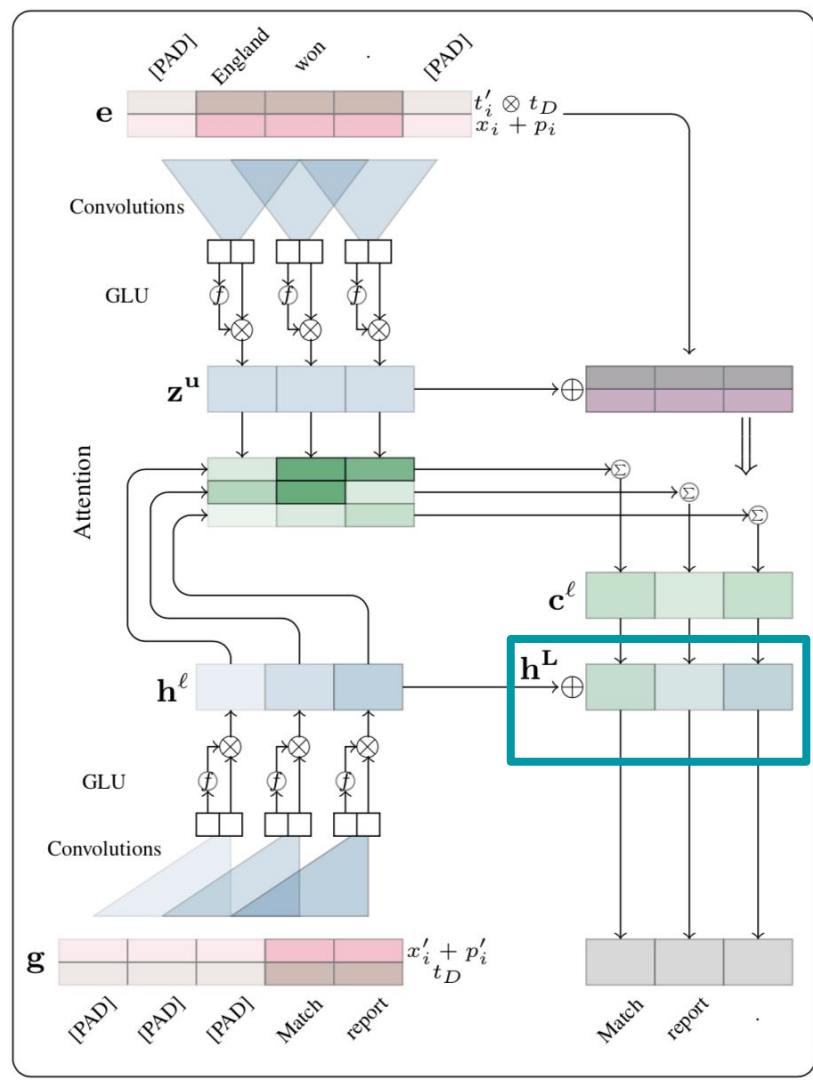# Topic Information Captures Document Theme in Generated Summaries

GOLD

Zac Goldsmith will contest the 2016 London mayoral election for the conservatives, it has been announced.

Topic Information Captures Document Theme in Generated Summaries

GOLD

Zac Goldsmith will contest the 2016 London mayoral election for the conservatives, it has been announced.

Without topic information

Pointer Generator

UKIP leader Nigel Goldsmith has been elected as the new mayor of London to elect a new conservative MP.

Convolutional

London mayoral candidate Zac Goldsmith has been elected as the new mayor of London.

Topic Information Captures Document Theme in Generated Summaries

GOLD

Zac Goldsmith will contest the 2016 London mayoral election for the conservatives, it has been announced.

Without topic information

Pointer Generator

UKIP leader Nigel Goldsmith has been elected as the new mayor of London to elect a new conservative MP.

Convolutional

London mayoral candidate Zac Goldsmith has been elected as the new mayor of London.

Our Model

Topic Convolutional

Former London mayoral candidate Zac Goldsmith has been chosen to stand in the London mayoral election.

27

Abstractiveness: Novel N-Grams

% of novel n-grams in generated summaries

Legend: PtGen (See et al) · Convolutional · Topic-Convolutional · Gold

unigrams: PtGen 27.4, Convolutional 31.3, Topic-Convolutional 30.7, Gold 35.8

bigrams: PtGen 73.3, Convolutional 79.5, Topic-Convolutional 79.2, Gold 83.5
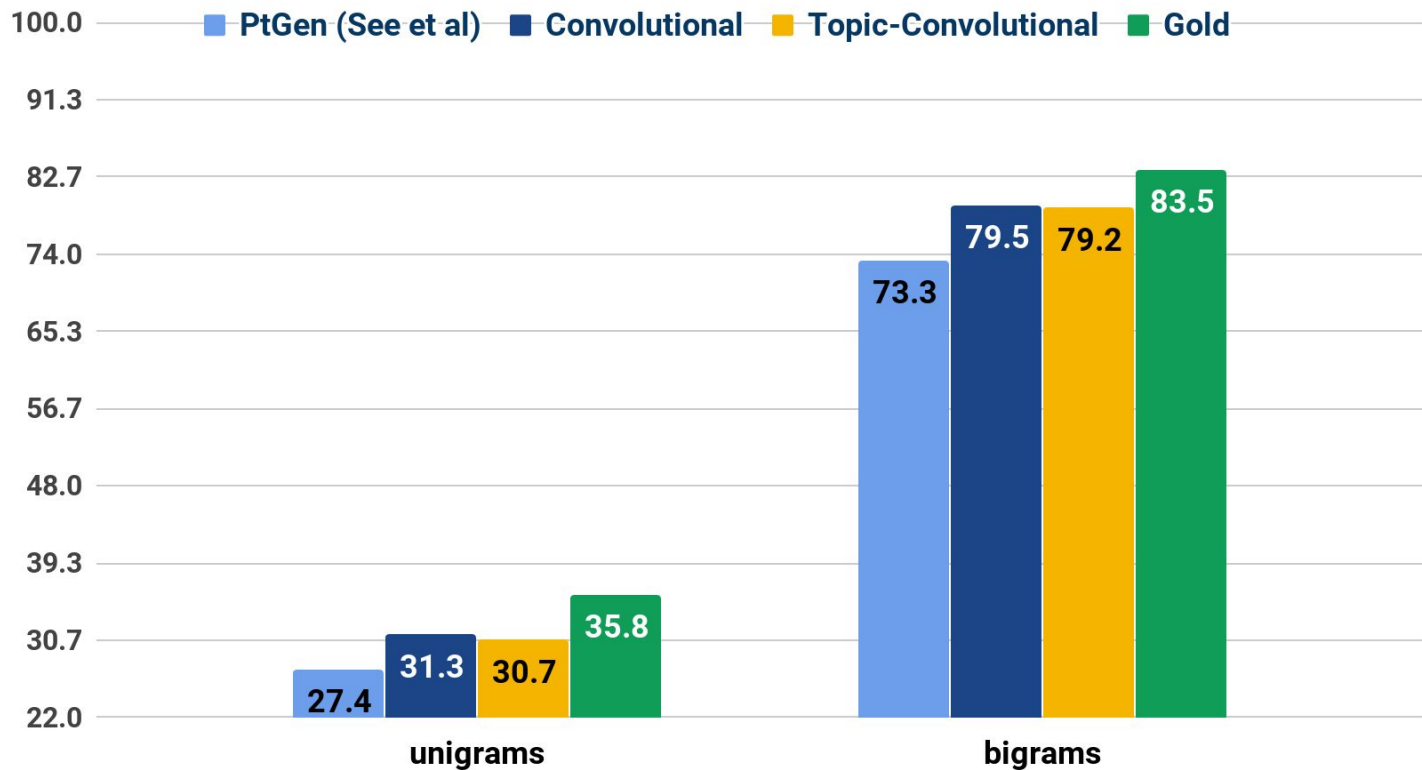
Abstractiveness: Novel N-Grams

% of novel n-grams in generated summaries
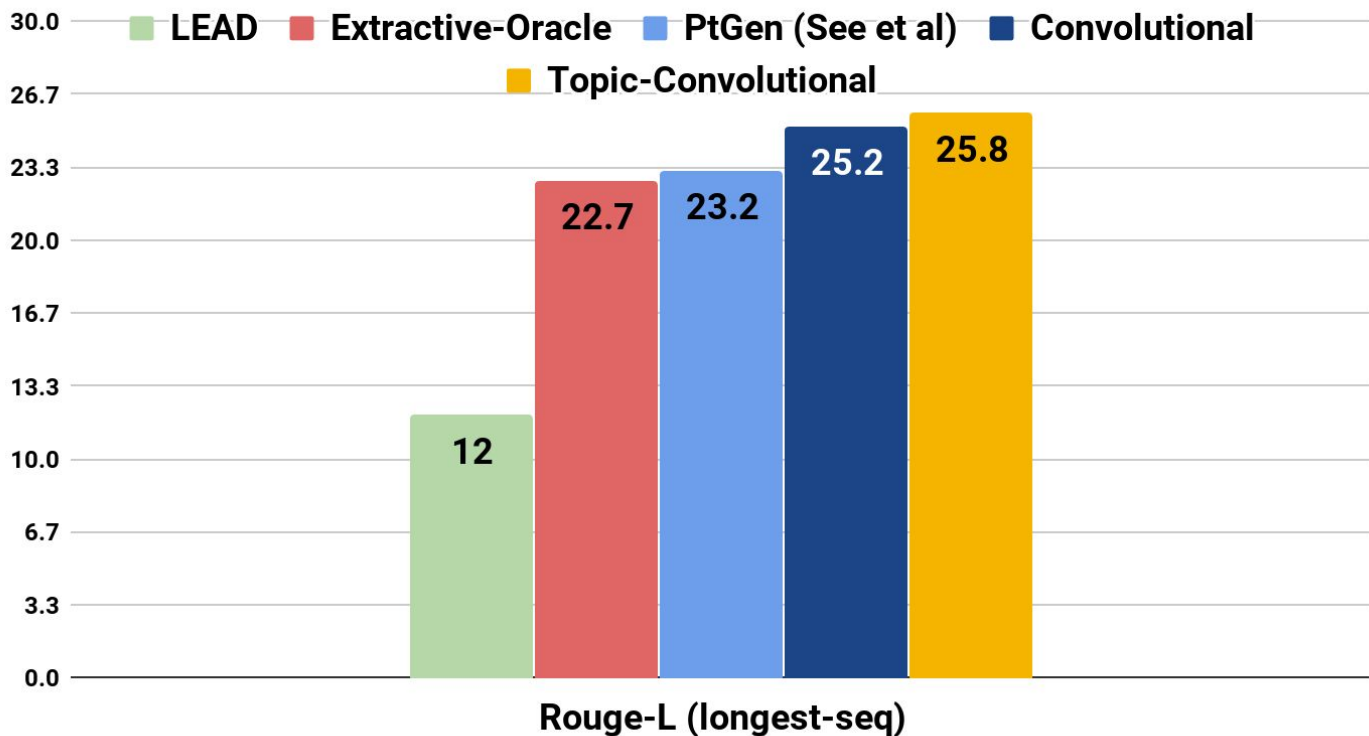

But are those summaries informative?

Informativeness: Automatic Evaluation with ROUGE

ROUGE (n-gram overlap) F-Scores

Informativeness:
QA-based Human Evaluation

ROUGE is not a reliable metric for Informativeness!

ONE-LINE SUMMARY: *A man and a child have been killed* after a *light* *aircraft made an emergency landing* on *a beach in Portugal*.

DOCUMENT: Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon.
The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane.
[*6 sentences with 139 words are abbreviated from here.*]
Other reports said the victims had been sunbathing when the plane made its emergency landing.
[*Another 4 sentences with 67 words are abbreviated from here.*]
Video footage from the scene carried by local broadcasters showed a small recreational plane parked on the sand, apparently intact and surrounded by beachgoers and emergency workers.
[*Last 2 sentences with 19 words are abbreviated.*]

Informativeness:
QA-based Human Evaluation

ONE-LINE SUMMARY: *A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.*
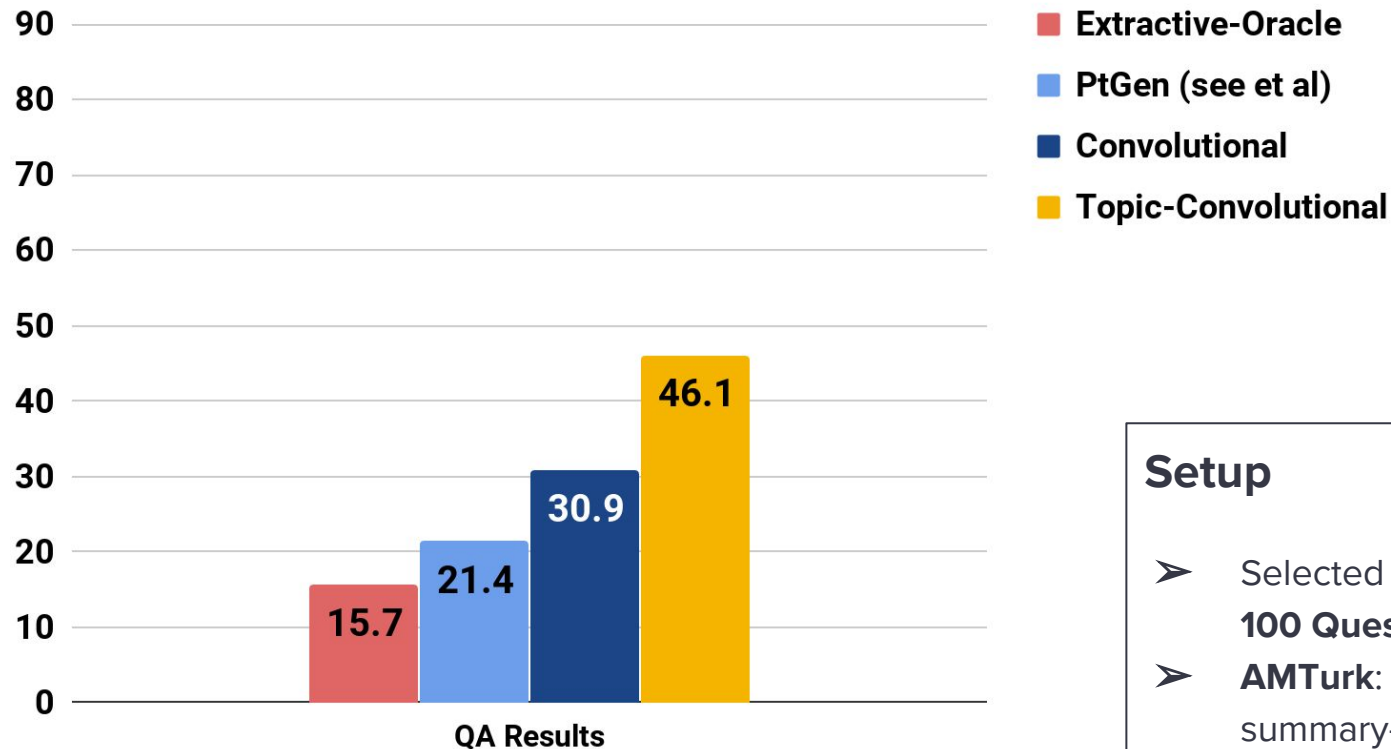
Document is not Shown.

Question Set

1. **Who died in the accident?**
   a. **A man and a child**
2. **Where did the accident happen?**
   a. **A beach in Portugal**
3. **What caused an accident?**
   a. **Landing of a light aircraft**

Informativeness: QA-based Human Evaluation

% of the Correctly Answered Questions

Legend:
- Extractive-Oracle
- PtGen (see et al)
- Convolutional
- Topic-Convolutional

Chart values:
- Extractive-Oracle: 15.7
- PtGen (see et al): 21.4
- Convolutional: 30.9
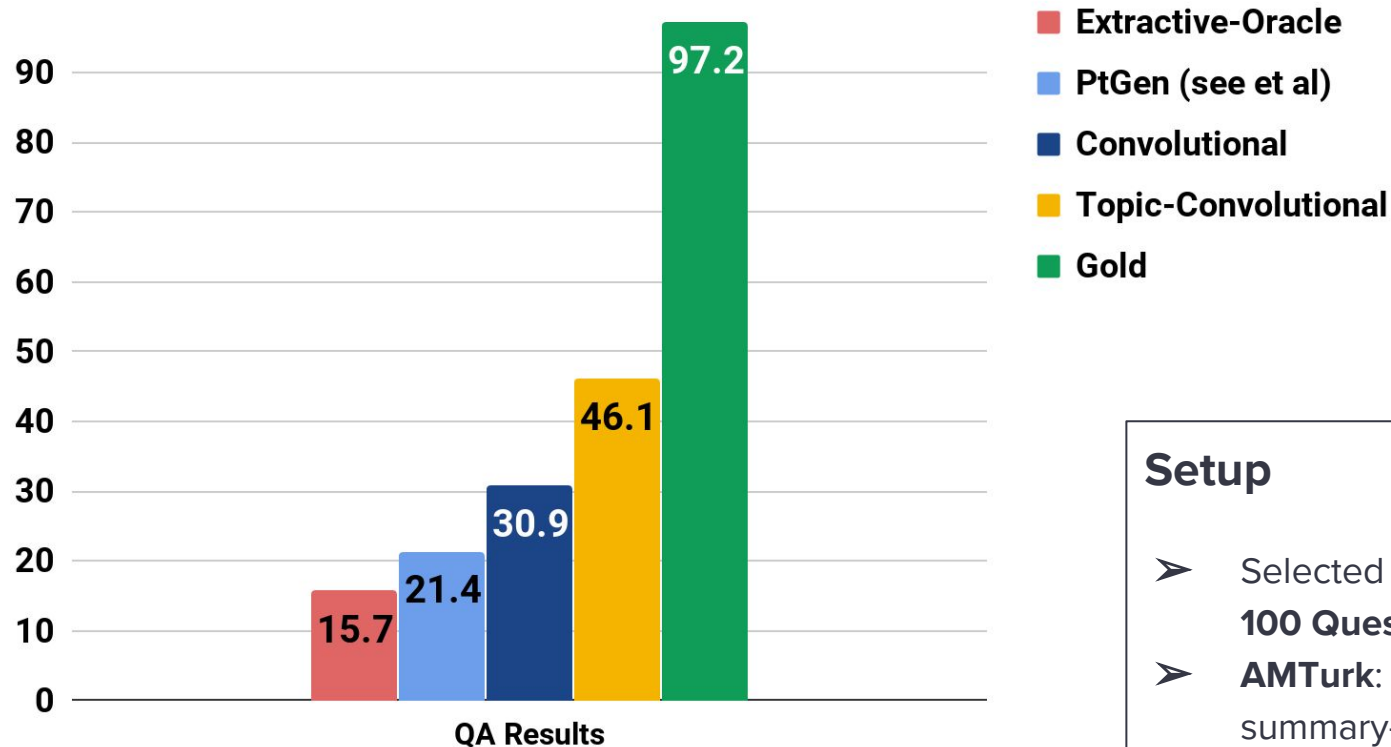- Topic-Convolutional: 46.1

X-axis: QA Results

Setup

➢ Selected **50 Documents with 100 Questions**

➢ **AMTurk**: **5 annotations** per summary-question pair.

33

Informativeness: QA-based Human Evaluation

% of the Correctly Answered Questions

Legend:
- **Extractive-Oracle**
- **PtGen (see et al)**
- **Convolutional**
- **Topic-Convolutional**
- **Gold**

Chart values: 15.7, 21.4, 30.9, 46.1, 97.2 — QA Results

Setup

➢ Selected **50 Documents with 100 Questions**

➢ **AMTurk**: **5 annotations** per summary-question pair.

Conclusions

➢ **Introduced "extreme summarization" together with a large-scale dataset to push the boundaries of abstractive methods**

➢ **Proposed a model with high-level document knowledge to recognize pertinent content and generate informative summaries**

➢ **Proposed a QA-based human evaluation to access informativeness**

Thank you!

Our code and dataset are available here:
https://github.com/shashiongithub/XSum

LDA: Topics Learned

➢ (charge, court, murder, police, arrest, guilty, sentence, boy, bail, space, crown, trial)

➢ (church, abuse, bishop, child, catholic, gay, pope, school, christian, priest, cardinal)

➢ (council, people, government, local, housing, home, house, property, city, plan, authority)

➢ (clinton, party, trump, climate, poll, vote, plaid, election, debate, change, candidate, campaign)

➢ (country, growth, report, business, export, fall, bank, security, economy, rise, global, inflation)

➢ (hospital, patient, trust, nhs, people, care, health, service, staff, report, review, system, child)

Ablation Rouge Results on the XSum: Test Set

Models	R1	R2	RL
Random	15.16	1.78	11.27
LEAD	16.30	1.60	11.95
EXT-ORACLE	29.79	8.81	22.66
SEQ2SEQ	28.42	8.77	22.48
PTGEN	29.70	9.21	23.24
PTGEN-COVG	28.10	8.02	21.72
CONVS2S	31.27	11.07	25.23
T-CONVS2S ($\text{enc}_{t'}$)	31.71	11.38	25.56
T-CONVS2S ($\text{enc}_{t'}, \text{dec}_{t_D}$)	31.71	11.34	25.61
T-CONVS2S ($\text{enc}_{(t', t_D)}$)	31.61	11.30	25.51
T-CONVS2S ($\text{enc}_{(t', t_D)}, \text{dec}_{t_D}$)	**31.89**	**11.54**	**25.75**