Inducing Document Structure for Aspect-based Summarization

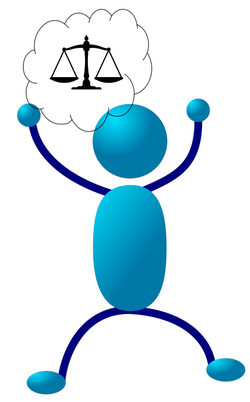
Lea Frermann, Melbourne University, lea.frermann@unimelb.edu.au
Alexandre Klementiev, Amazon Research, klementa@amazon.com

In a nutshell...

steffi graf reluctantly paid 1.3 million marks to charity as part of a **settlement** with german prosecutors who dropped their **tax evasion investigation** of the **tennis player**. […] she wanted to put the **media circus** about her **tax affairs** behind her and concentrate on **tennis** . graf , 27 , who recently **lost** her ranking as **world number one** women 's tennis player to 16-year-old martina hingis after 94 weeks o**n** top , was quoted by the magazine as saying […] it was worth it to **avoid further litigation** . […] the **seven-times wimbledon champion** , who has not played since the semifinals of a **tournament**

prosecutors dropped their investigation last month after probing graf 's finances for nearly two years when she agreed to their offer to pay a sum to charity last month as part of a settlement with german prosecutors who dropped their tax evasion investigation of the tennis player , a news magazine

seven-times wimbledon champion could make a return to the court at the end of april in the german open . former family tax adviser joachim eckardt received two and a half years for complicity .

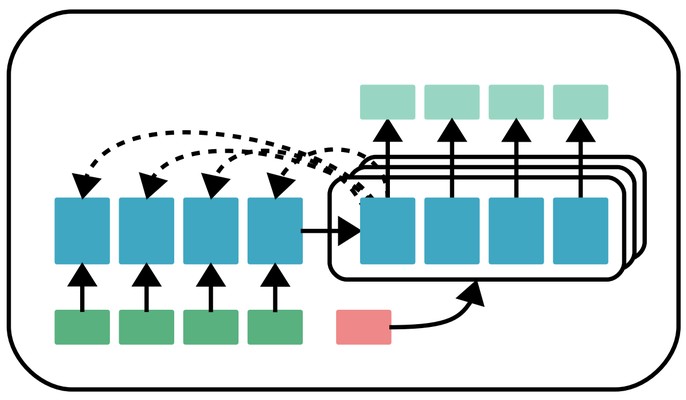


Contributions

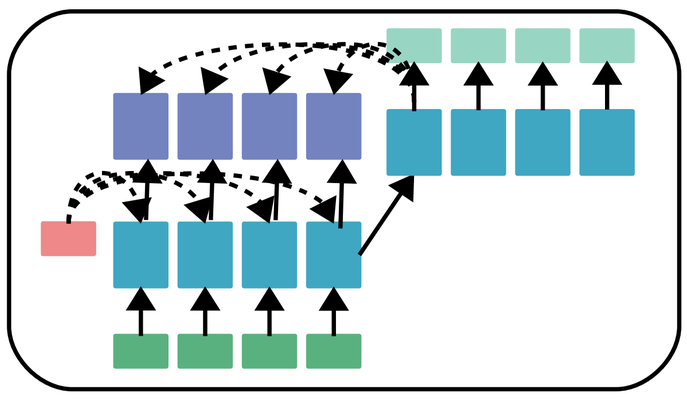
- joint learning of latent aspects and summarization
- aspect segmentation of the document induced without explicit supervision
- synthetic training paradigm
- three neural models for aspect-aware summarization

Models

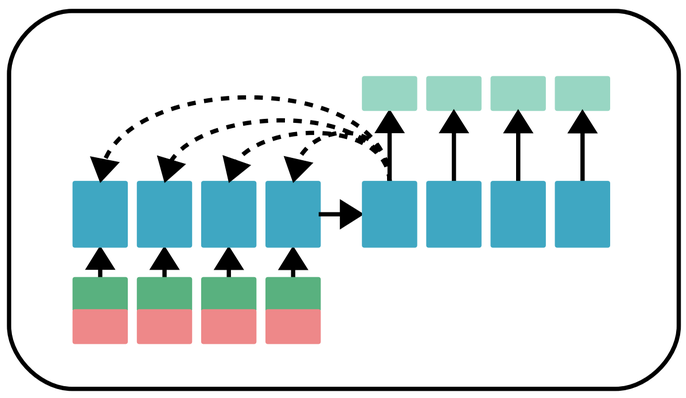
Decoder Attention



Encoder Attention



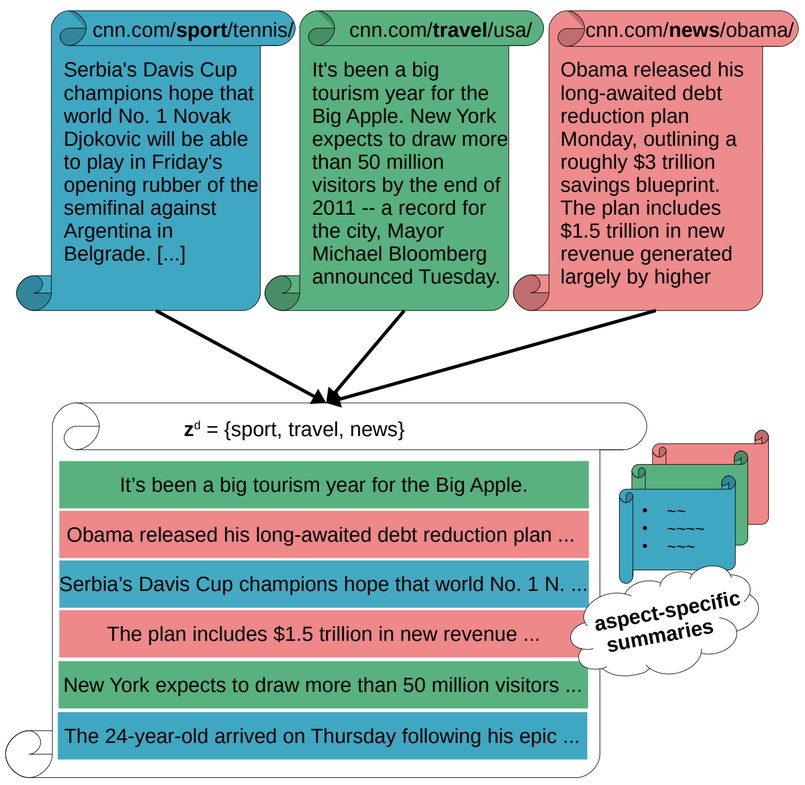
Source Factors



Synthetic Training Data I

- CNN / DM documents
- document URLs ⇒ original aspect
- interleave to multi-aspect documents
- pair with original (∼ aspect-specific) summaries

Synthetic Training Data II



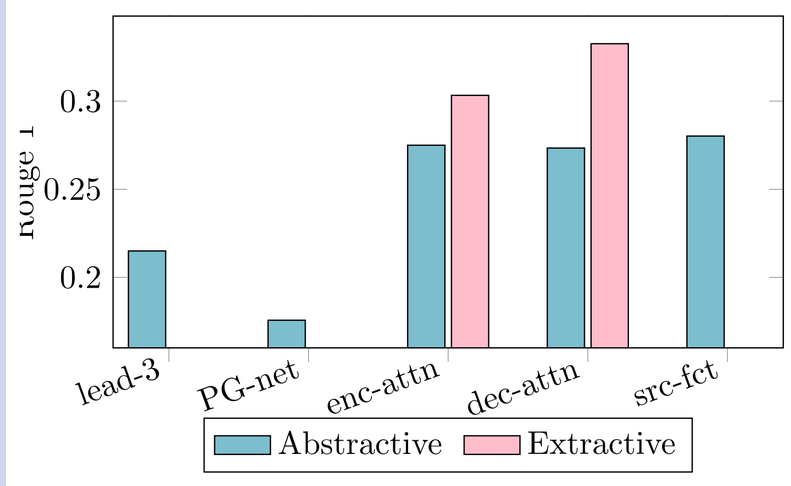
The Multi-Aspect News Corpus

Find it here:

$\mathcal{A}$ = `tvshowbiz, travel, health, sciencetech, sports, news`

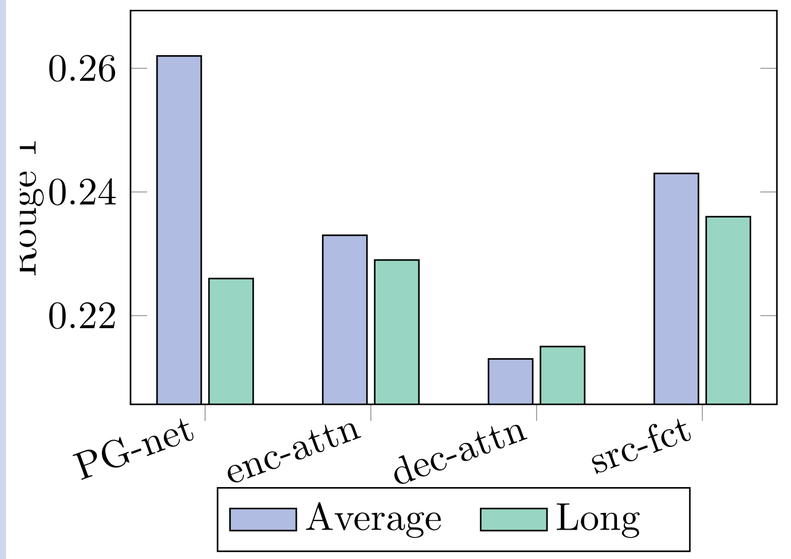
- $w_d^p \sim U(1,5)$ sentences per paragraph
- $w_d \leq 1000$ words per document
- $n_d \sim U(1,4)$ aspects per document
- n_d document-summary pairs per document n
- 284,701 train / 1,000 valid / 1,000 test documents

Exp 1: Synthetic Multi-topic Docs



- Data: synthetic documents
- PG-net (See et al., 2017)
- lead-3: first 3 sentences

Exp 2: Natural long Docs



- Data: original CNN / DM documents
- Long: $\geqq$ 2,000 words
  Average: $\leqq$ 1,000 words

Exp 3: Natural multi-topic Docs

	lead-2	enc-attn	dec-attn	sf
accuracy	0.540	0.543	0.553	0.553
diversity	0.127	0.177	0.197	0.133

- Data: Reuters news
- MTurk: assign topic to document-summary pair
  2 summaries per document
- lead-2: first two sentences as one summary each