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...

Contribution
- 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

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
- 2 cnn.com documents
- Sentimental news corpus that the world No. 1 Novak Champions will be able to play in Friday’s opening rubber of the semifinals against Argentina in Belgrade. [...] Sentimental news corpus that the world No. 1 Novak Champions will be able to play in Friday’s opening rubber of the semifinals against Argentina in Belgrade. [...] Sentimental news corpus that the world No. 1 Novak Champions will be able to play in Friday’s opening rubber of the semifinals against Argentina in Belgrade. [...]

Encoder Attention

Decoder Attention

Source Factors

The Multi-Aspect News Corpus

Find it here:

\[ A = \{ \text{tvshowbiz, travel, health, sciencetech, sports, news} \} \]

\[ w_P \sim U(1, 5) \text{ sentences per paragraph} \]

\[ w_D \leq 1000 \text{ words per document} \]

\[ n_{DA} \sim U(1, 4) \text{ aspects per document} \]

\[ n_{DS} \text{ document-summary pairs per document} \]

\[ n = 284,701 \text{ train} / 1,000 \text{ valid} / 1,000 \text{ test documents} \]

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: \( \geq 2,000 \) words
- Average: \( \leq 1,000 \) words

Exp 3: Natural multi-topic Docs
- Data: Reuters news
- MTurk: assign topic to document-summaries pair
- 2 summaries per document
- lead-2: first two sentences as one summary each