Beyond Paragraphs: NLP for Long Sequences

Iz Beltagy, Arman Cohan, Hannaneh Hajishirzi, Sewon Min, Matthew Peters

NAACL-HLT 2021

Slides and Code: https://github.com/allenai/naacl2021-longdoc-tutorial

Presenters

Iz Beltagy
AI2

Arman Cohan
AI2

Hanna Hajishirzi
University of Washington, AI2

Sewon Min
University of Washington

Matthew Peters
AI2

UWNLP AI2

Overview

- Tutorial for long document and multi-document NLP, focusing on problems unique to long document setting
- Typical setting: >500-1000 tokens

Overview

- Tutorial for long document and multi-document NLP, focusing on problems unique to long document setting
- Typical setting: >500-1000 tokens
- Dual focus:
  - Review state-of-the-art methods (especially Transformers)
  - Practical implementation details (with hands on coding!)

Overview

- Tutorial for long document and multi-document NLP, focusing on problems unique to long document setting
- Typical setting: >500-1000 tokens
- Dual focus:
    - Review state-of-the-art methods (especially Transformers)
    - Practical implementation details (with hands on coding!)
- Typical tasks:
    - Document classification
    - Multihop QA
    - Many more!

Overview

- Tutorial for long document and multi-document NLP, focusing on problems unique to long document setting
- Typical setting: >500-1000 tokens
- Dual focus:
    - Review state-of-the-art methods (especially Transformers)
    - Practical implementation details (with hands on coding!)
- Typical tasks:
    - Document classification
    - Multihop QA
    - Many more!
- Intended audience:
    - Researchers and practitioners with NLP backgrounds interested in document NLP (without assuming prior experience in this area).

Why long sequence NLP?

- Many practical NLP problems require processing long sequences:
  - Scientific literature (typical document is 1K-10K words or longer)
  - Digital humanities (books can have 100,000 words or more: *Harry Potter and the Deathly Hallows* is about 200K words)
  - Multihop QA with multiple documents (average length of context in HotpotQA is 1.3K, Yang et al, 2018)

Why long sequence NLP?

- Many practical NLP problems require processing long sequences:
    - Scientific literature (typical document is 1K-10K words or longer)
    - Digital humanities (books can have 100,000 words or more: *Harry Potter and the Deathly Hallows* is about 200K words)
    - Multihop QA with multiple documents (average length of context in HotpotQA is 1.3K, Yang et al, 2018)
- Fundamental advances in ability to process very long sequences opens up new approaches and application areas (inside, and outside NLP).

Key challenges for long sequence NLP

- Annotating full documents or multiple documents can be difficult, and many NLP datasets focus on short contexts.

Key challenges for long sequence NLP

- Annotating full documents or multiple documents can be difficult, and many NLP datasets focus on short contexts.
- End tasks often require combining information spread over long distances, either in document, or among many documents. Models need to ignore a lot of irrelevant text.

Key challenges for long sequence NLP

- Annotating full documents or multiple documents can be difficult, and many NLP datasets focus on short contexts.
- End tasks often require combining information spread over long distances, either in document, or among many documents. Models need to ignore a lot of irrelevant text.
- Many popular algorithms are designed to work in short sequence setting, and have limitations in long setting:
  - RNN/LSTM: process input sequentially → slow for long sequences
  - Transformers: self-attention is $O(L^2)$ → cannot process long input with current hardware. Many pre-trained LMs limited to 512 tokens (e.g. BERT).

More tutorial details

- What we will cover:
  - Overview of tasks and datasets
  - Graph based methods
  - Methods for extending transformers to long sequences
  - Practical implementation details with hands on coding for abstractive summarization of long documents
- What we won't cover:
  - Retrieval based methods and open domain QA (see Chen and Yih, ACL 2020 tutorial)
  - Multilingual and document translation methods (Bender rule: all tasks and datasets are English language)

UWNLP AI2

Detailed outline

1. Overview of tasks and datasets (10 minutes)
2. Graph based methods (35 minutes)
3. Long sequence transformers (45 minutes)
4. Pretraining / fine-tuning (25 minutes)
5. Hands on use case: document summarization (45 minutes)
6. Future work and conclusion (10 minutes)

Also live Q&A session (check conference schedule).

Tasks and Datasets

Tasks and Datasets - domain overview

- News articles (~500 - 1000 words)
- Wikipedia (~few thousand words)
- Books / stories (~1K - 1M words)
- Technical domains:
    - PubMed, Medline
    - arXiv (computer science / math / physics)
    - Patents

UWNLP

Tasks and Datasets - task overview

- Single document tasks
    - Classification
    - Question Answering
    - Information extraction - entity extraction, relationship extraction
    - Coreference
    - Summarization
- Multiple document tasks - many single document tasks plus:
    - Multihop QA
- Long sequence language modeling benchmarks

See also: Hugging Face Datasets, List of NLP datasets, TF Datasets

Document classification

Given input document, classify it into one or more predefined classes.

Dataset	Type	Avg doc. len
IMDB (Mass et al, 2011)	Sentiment (movie reviews)	~300
Hyperpartisan (Kiesel et al. 2019)	News	~700
Arxiv (He et al. 2019)	Arxiv subject area	Full docs (~7000)
SciDocs (Cohan et al. 2020)	Scientific documents - MeSH (Medical Subject Headings) and paper topic	Full docs
Patents (Lee et al. 2020)	US Patent claims	Full claims

Single document Question Answering

Dataset	Domain	Avg doc. len
News QA (Trischler. et al 2016)	Uses CNN / Daily Mail dataset (Hermann et. al 2015)	~700
Narrative QA (Kočiský et al, 2017)	Books + movie scripts - two settings, full document and summary only	~60K (full) ~650 (summary)
Search QA (Dunn et al. 2017)	Jeopardy questions + search snippets	~1850 (~50 snippets per question)
Trivia QA (Joshi et al. 2017)	Open web and Wikipedia (groups multiple documents into single instance)	~3000
Natural Questions (Kwiatkowski et al. 2019)	Wikipedia	Full article (median = 3200)
Qasper (Dasigi et al. 2021)	NLP research papers	Full paper text

Note: some benchmarks (MRQA) truncate contexts.

Information extraction

Lutsenko is a former minister of internal affairs. He occupied this post in the cabinets of Yulia Tymoshenko. The ministry of internal affairs is the Ukrainian police authority.

Subject: *Yulia Tymoshenko* **Object:***Ukrainian*

Relation: country of citizenship

From [Nan et al. (2020)](), adapted from DocRED.

Entity Extraction:

- Return typed spans
- Ontonotes, CoNLL 2003

Relationship extraction:

- Return (subject, relationship, object) tuples
- Full document:
  - SciREX, [Jain et al. (2020)]() (salient entity, relationship), full articles.
- Shorter context:
  - DocRED [Yao et al. (2019)](), Wikipedia
  - BC5CDR, [Li et al. (2016)](), PubMed abstracts

Coreference

"*I voted for Nader because he was most aligned with my values," she said.*

Coreference is task of clustering entity mentions across documents.

- Many datasets for coreference: see Sukthanker et al. 2018, "Anaphora and Coreference Resolution: A Review" for a complete list.
- Most popular single document dataset: CoNLL-2012 shared task, (Pradhan et al, 2012), multidomain, multilingual based on OntoNotes 5 (~500 avg. document length)
- Most popular multi-document: ECB+ (Cybulska and Vossen, 2014), news, within and across document entity and event annotations

Example from Stanford NLP

Multihop Question Answering

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

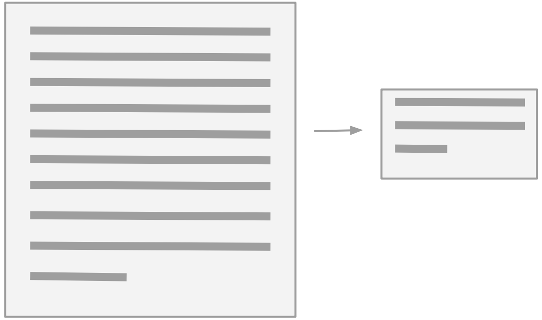
Multihop QA requires combining information from multiple sources to answer a question.

- HotpotQA, Yang et al. (2018): Wikipedia, includes annotations of supporting facts, distractor and open settings (example to left).
- Wikihop / Medhop, Welbl et al. (2018): Wikipedia and Medline, without support annotation.

See also Open Domain QA (Chen and Yih, ACL 2020 tutorial)

UWNLP AI2

Document summarization



Summarize a long input document as significantly shorter document.

Summaries contain both abstractive and extractive elements.

News Domain

Dataset	Input	Output
New York Times (Napoles et al. 2012)	~800	~45
Newsroom (Grusky et al. 2018)	~750	~30
CNN/Daily Mail (Hermann et. al 2015)	~700	~45

Technical domains

Dataset	Input	Output
Arxiv (Cohan et al. 2018)	~4900	~200
Pubmed (Cohan et al. 2018)	~3000	~220
BigPatent (Sharma et al. 2019)	~700	~45

Long sequence language modeling

Dataset	Source	Level	Size
Enwiki8, Text8 (Mahoney 2009)	Wikipedia: Enwiki8 with full markup, Text8 without markup	char	100MB
Wikitext (Merity et al. 2016)	Wikipedia (2 and 103 million word versions)	word	515MB
PG-19 (Rae et al. 2019)	Project Gutenberg books before 1919	word	28K books, 11GB text

- Any corpus (or other modalities, e.g. images, audio) with long contexts can be used. NLP benchmarks listed above.
- See also Long Range Arena, Tay et al. (2021) for a general benchmark across modalities (text, images)