NATURAL LANGUAGE PROCESSING

LECTURE 13: Applications

INDEX

How to exploit model?

- Sequence Classification
  - Sentiment Analysis
- Token Classification
  - NER
  - QA
- Similarity Measure
  - Retrieval

NLP Roadmap

Task

- Sentiment Classification
- Machine Reading
- Machine Translation
- Language Model
- ...

Formulation

- Text Classification
- Token Classification
- Retrieval
- Text Generation

Model

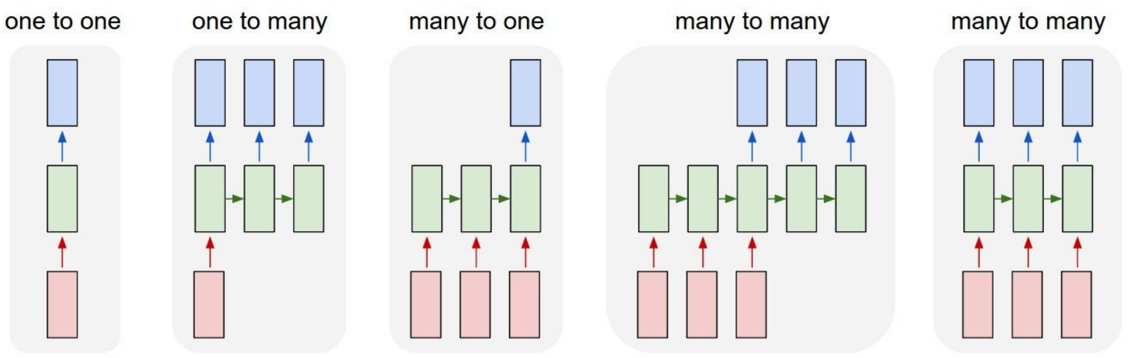
- RNNs
- Encoder-Decoder
- Attention
- Transformer

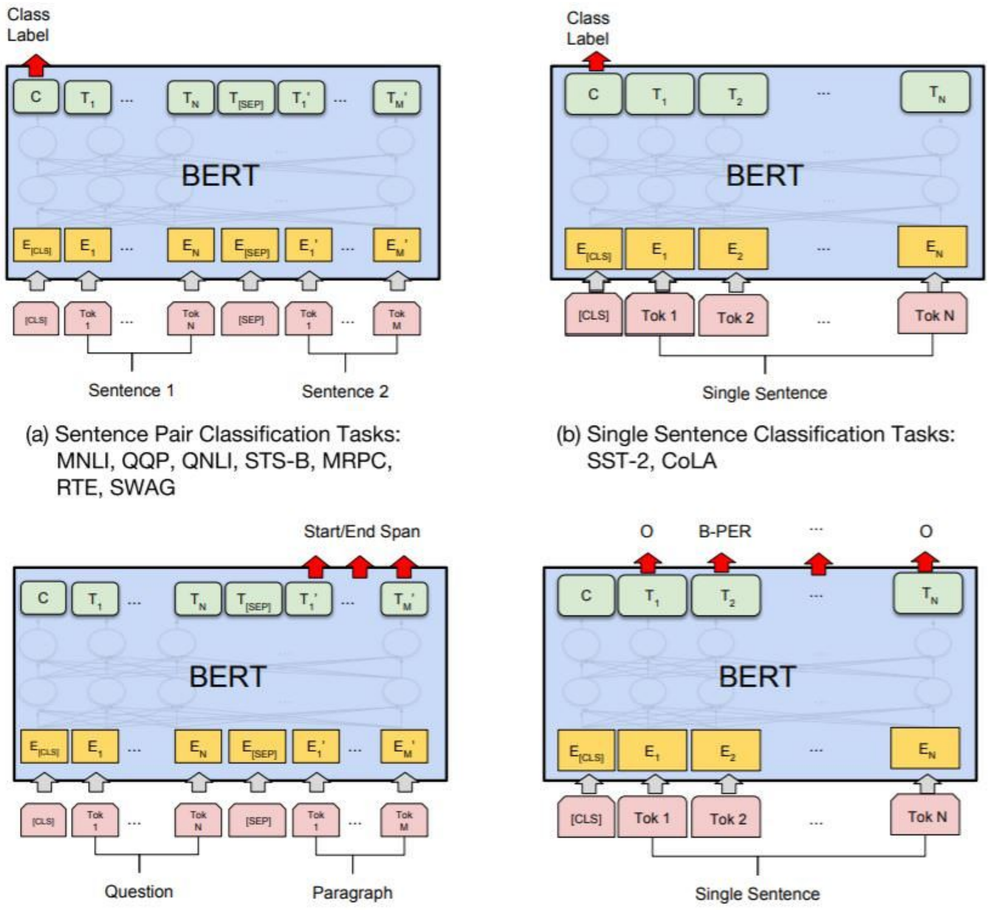
Learning

- Vanilla
- Pretraining & Finetuning
- In-context Learning

Recap

Various Model Architecture





(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

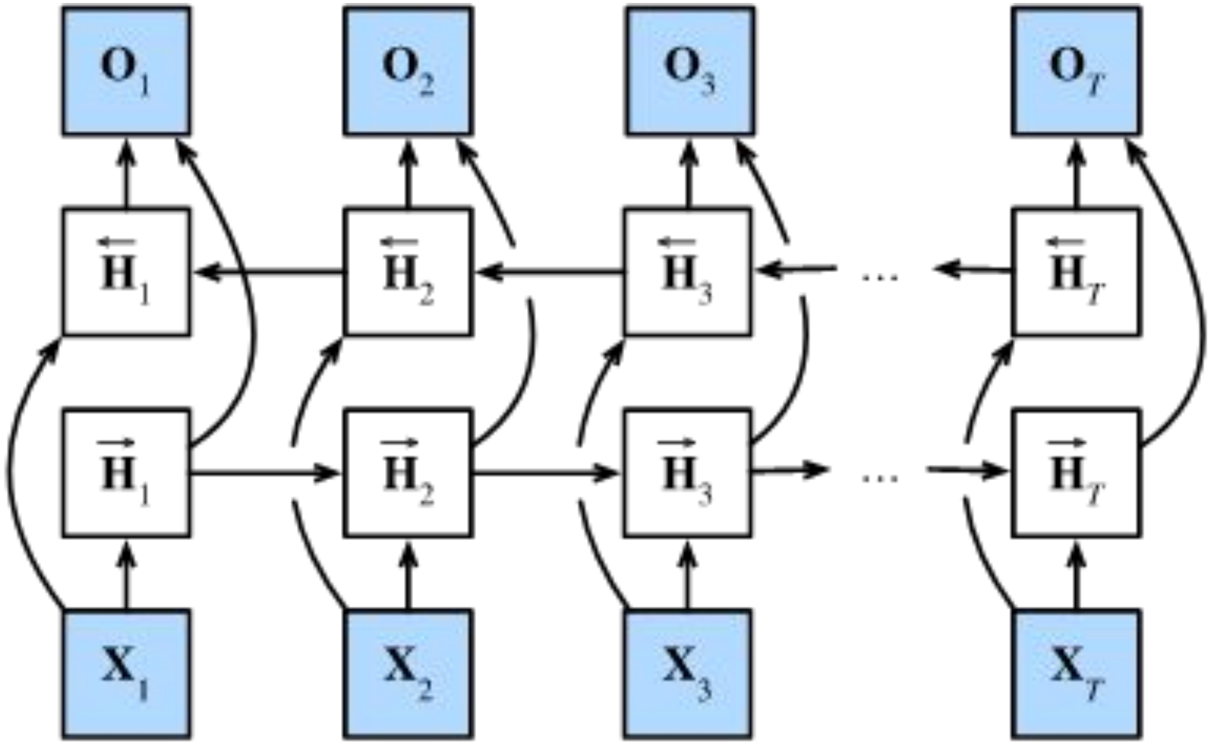
(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Recap

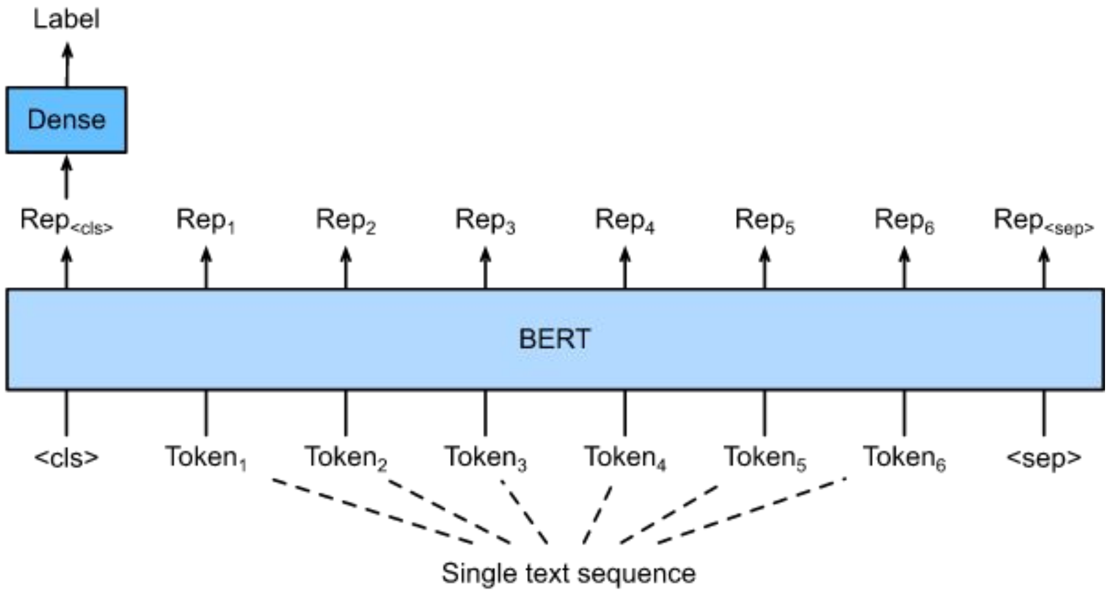
Bi-directional RNNs

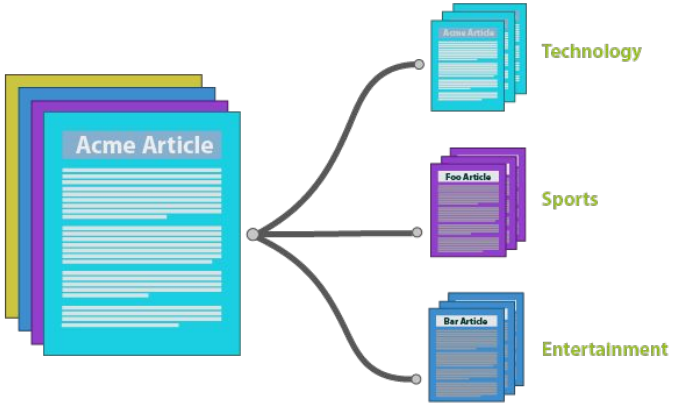


Text Classification

Text Classification is also known as sequence classification.

- In text classification, classify the entire text into categories
- extract “prototype” representation from entire token representation.
- E.g., spam classifier, sentiment analysis, article classifier



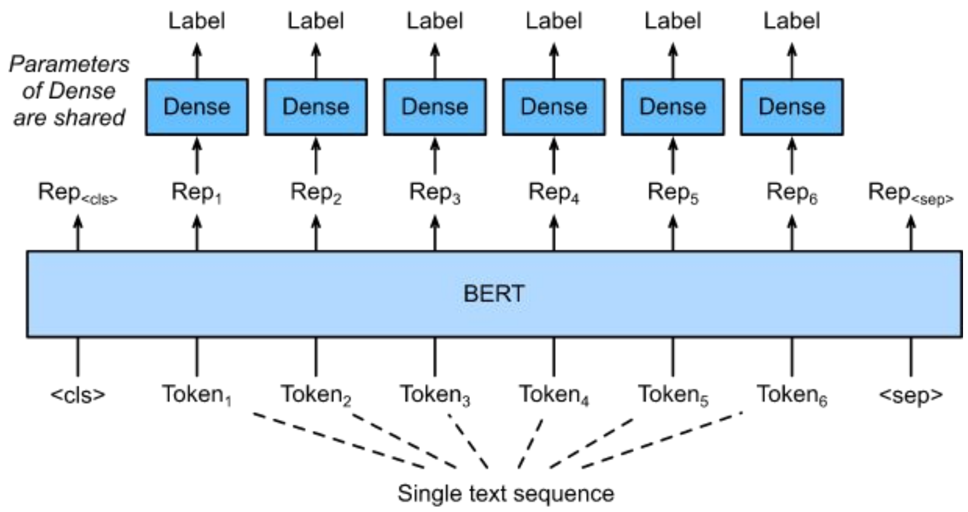


Token Classification

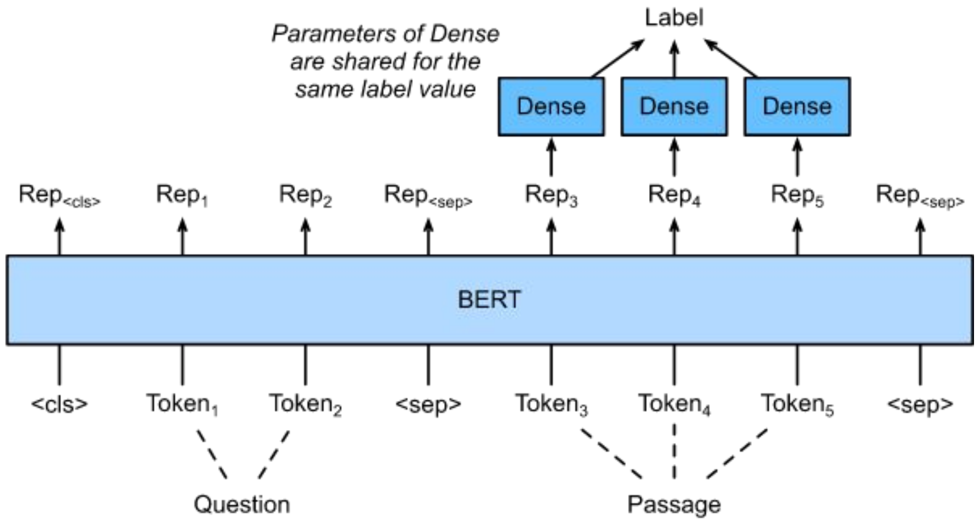
Token Classification is also known as sequence tagging.

- In token classification, classify each token of the text.

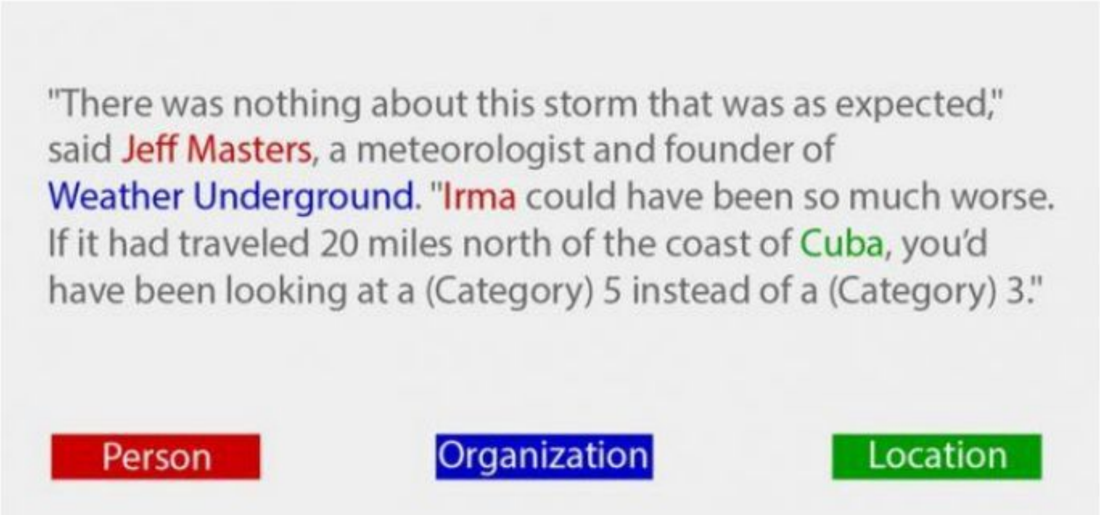
Text Tagging



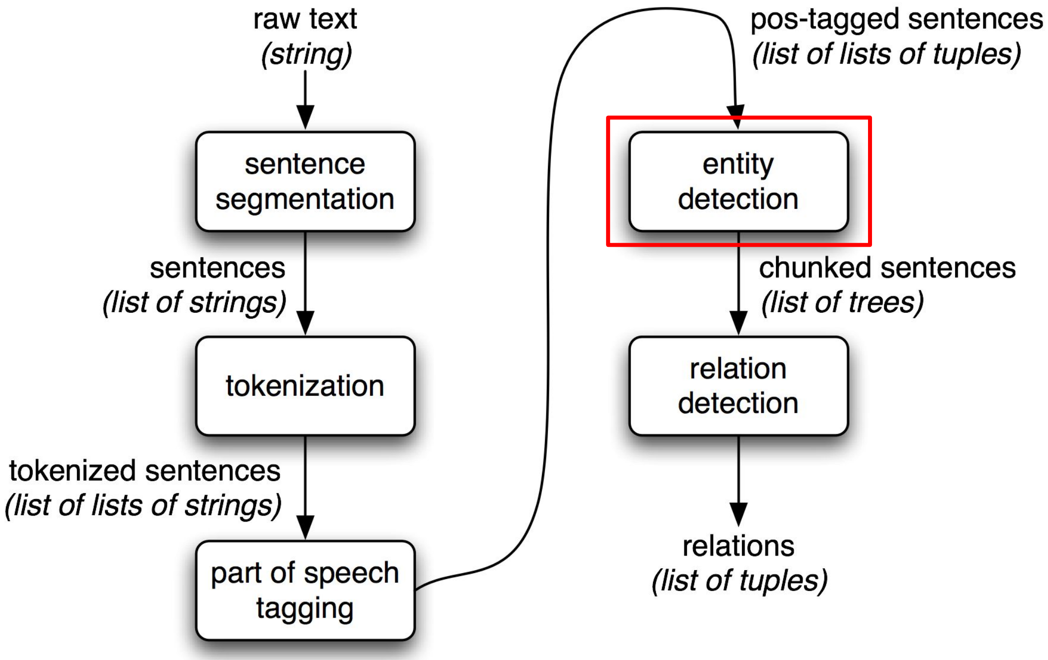
QA



Named Entity Recognition



▲ NER example



▲ Information extraction pipeline

https://www.nltk.org/book/ch07.html

Named Entity Recognition

In information extraction, a **named entity** is a real-world object, such as a person, location, organization, product, etc., that can be denoted with a proper name. It can be abstract or have a physical existence. Examples of named entities include Barack Obama, New York City, Volkswagen Golf, or anything else that can be named. Named entities can simply be viewed as entity instances (e.g., New York City is an instance of a city).

Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.

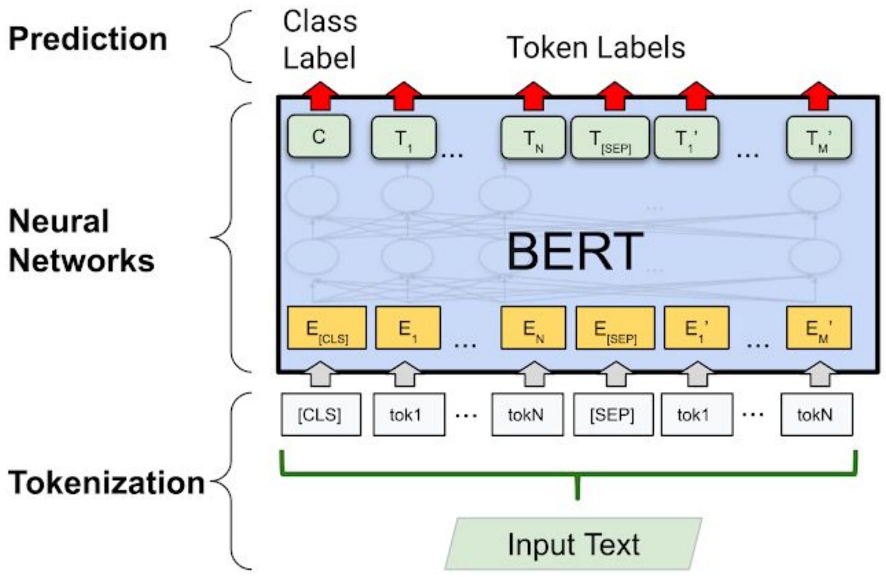
Example)

- Original Sentence : “EU rejects german call to boycott british lamb.”
- Ground Truth Entity : EU-ORG, german-MISC, british-MISC

Named Entity Recognition

NER as BIO tagging (Token-level prediction)

B - Begin / I - Interior / O-out



Ex1) EU rejects german call to boycott british lamb.

→ Process into [“eu”, “reject”,”#s”, “german”, 'to', 'boycott', 'british', 'lamb', '.']

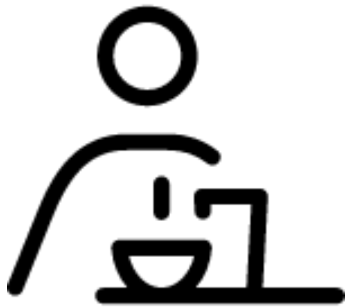
→ label : [“B-ORG”, “O”, “O”, “B-MISC”, “O”, “O”, “B-MISC”, “O”, “O”]

Ex2) Barack Obama was the president of the United States

10 minutes break

&

Leave questions in chat



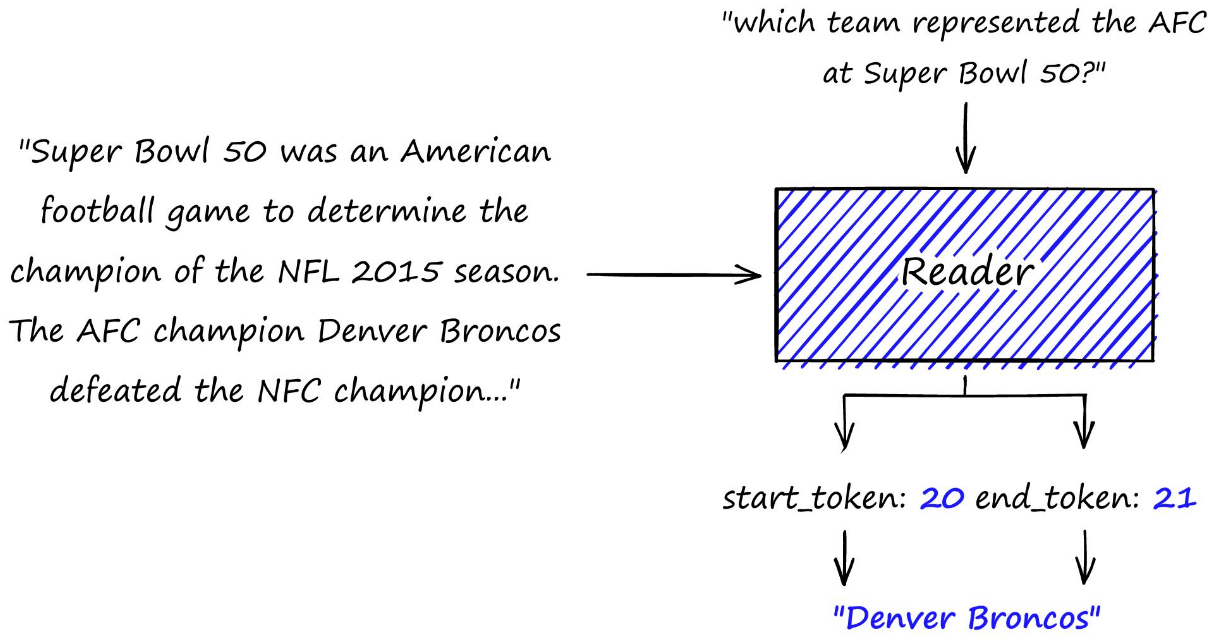
Machine Reading Comprehension (MRC)

Question Answering (Extractive)

Hypothesis :

Ground truth answer always in the paragraph

- Input is Context and question
- Expected Output is a span in the context
- Classifying start, end and others



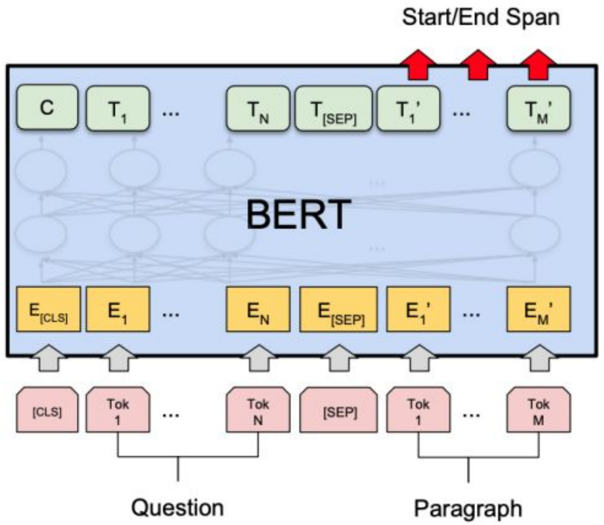
QA

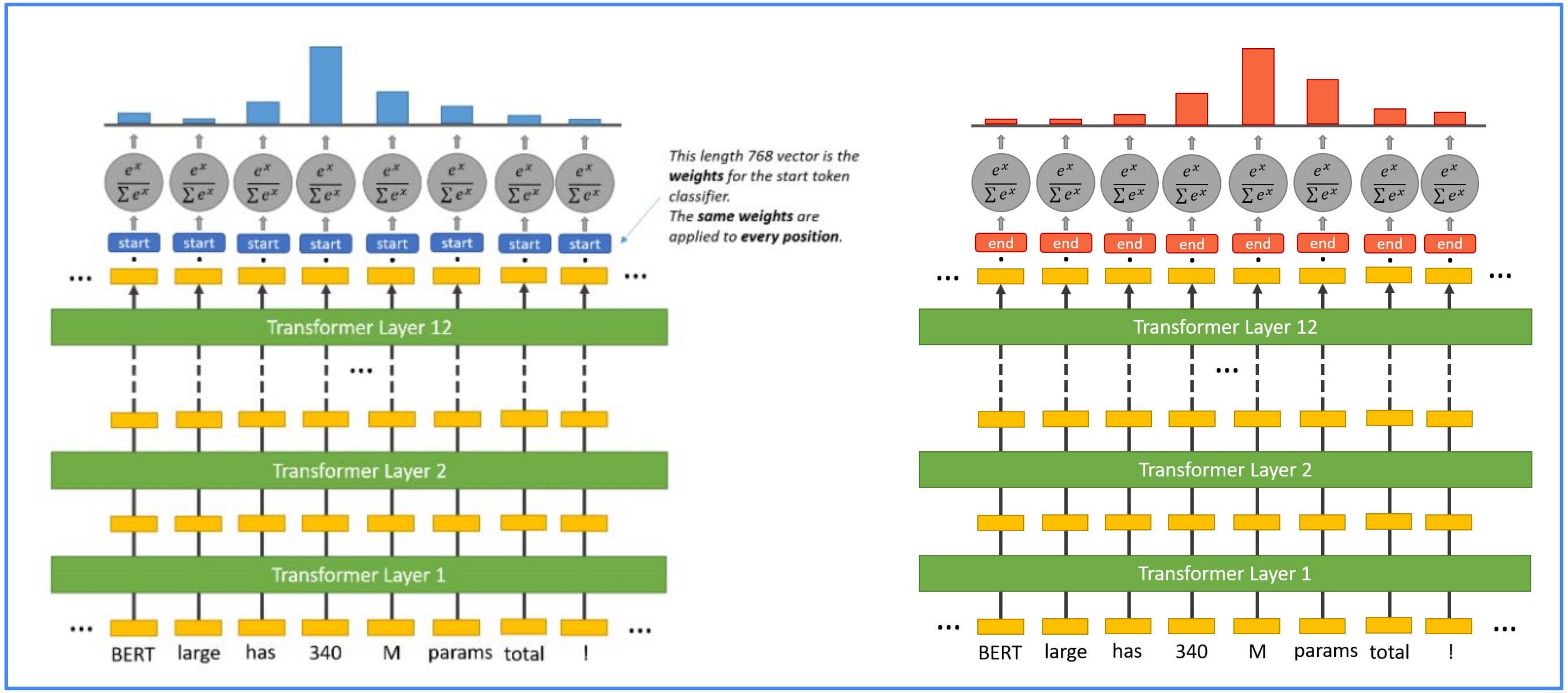
Question Answering (Extractive) with BERT

$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

$$p_{\text{start}}(i) = \text{softmax}_i(\mathbf{w}_{\text{start}}^\mathsf{T}\mathbf{H})$$

$$p_{\text{end}}(i) = \text{softmax}_i(\mathbf{w}_{\text{end}}^\mathsf{T}\mathbf{H})$$





QA example

Datasets: SQuAD, CoQA


```
{
  "version": "1.0",
  "data": [
    {
      "source": "wikipedia",
      "id": "3zotghdk5ibi9cex97fepx7jetpso7",
      "filename": "Vatican_Library.txt",
      "story": "The Vatican Apostolic Library (), more commonly calle
established in 1475, although it is much older, it is one of the olde
codices from throughout history, as well as 1.1 million printed books
philosophy, science and theology. The Vatican Library is open to anyc
published between 1801 and 1990 can be requested in person or by mail
manuscripts, to be made available online. \n\nThe Vatican Secret Arch
\n\nScholars have traditionally divided the history of the library ir
the initial days of the library, dated from the earliest days of the
      "questions": [
        {
          "input_text": "When was the Vat formally opened?",
          "turn_id": 1
        },
        {
          "input_text": "what is the library for?",
          "turn_id": 2
        },
    "answers": [
      {
        "span_start": 151,
        "span_end": 179,
        "span_text": "Formally established in 1475",
        "input_text": "It was formally established in 1475",
        "turn_id": 1
      },
      {
        "span_start": 454,
        "span_end": 494,
        "span_text": "he Vatican Library is a research library",
        "input_text": "research",
        "turn_id": 2
      },
```


▲ CoQA dataset example

QA example

QA model with pre-trained BERT model

- Question: “Who is the acas director?”
- Answer: “Agnes karin ##gu.”
- Bert uses **wordpiece tokenization.**
  - In BERT, rare words get broken down into subwords/pieces.
  - Wordpiece tokenization uses ## to delimit tokens that have been split.
  - “Karin” is a common word → maintain
  - “Karingu” is a rare word → “Karin” and “##gu”.

Long term dependency in QA

Long term dependency in QA

- A model needs to be sufficiently aware of distant tokens
- When dealing with long text and paragraphs, LSTM is not good enough

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

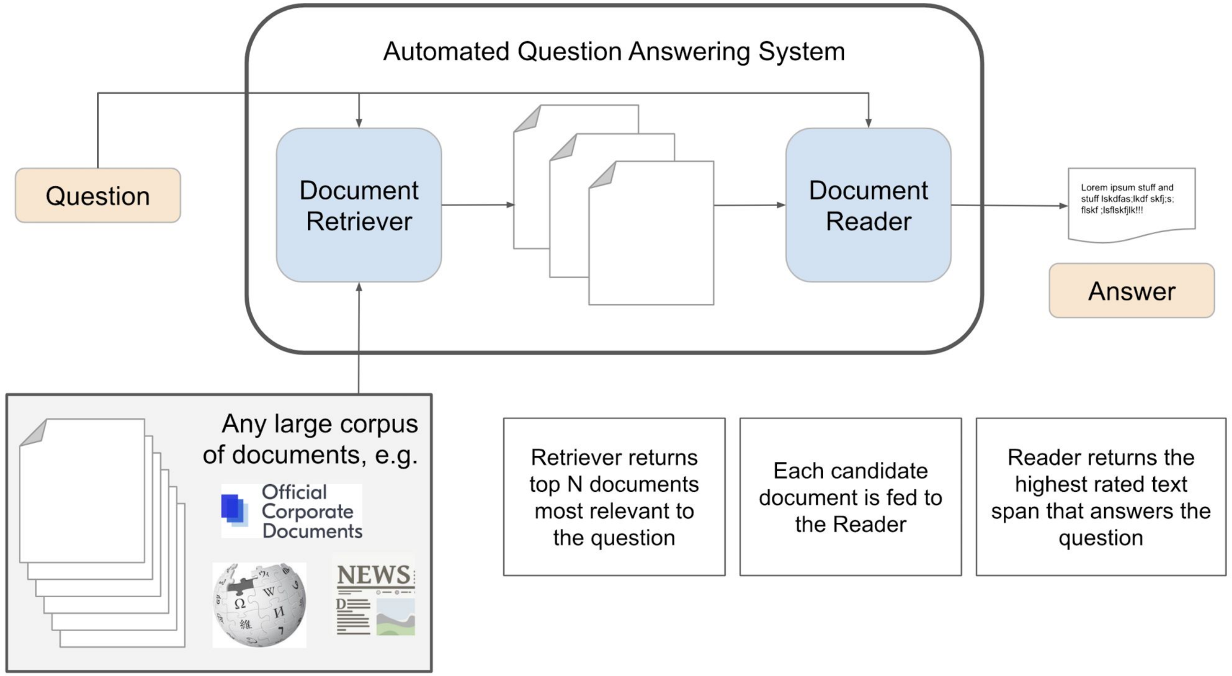
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

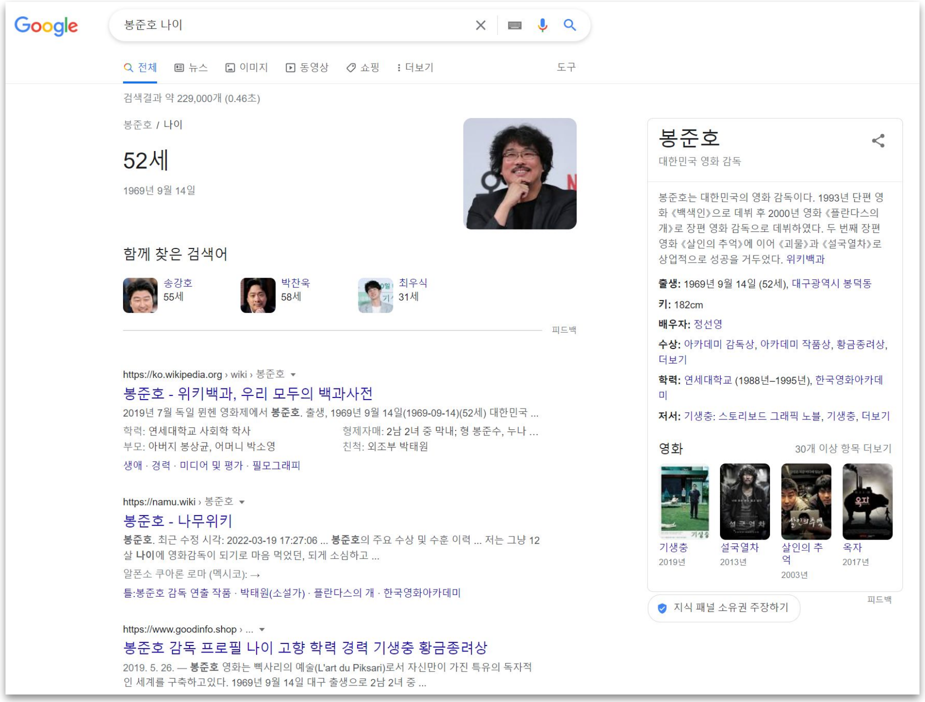
Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Retrieval

Open Domain QA / Entity Retrieval

1. Retriever searches for the most relevant documents in response to a query
2. Reader gives the selected documents a closer look by passing them through a pre-trained QA language model.
3. The model then returns the text passages that it deems most likely to answer the query.

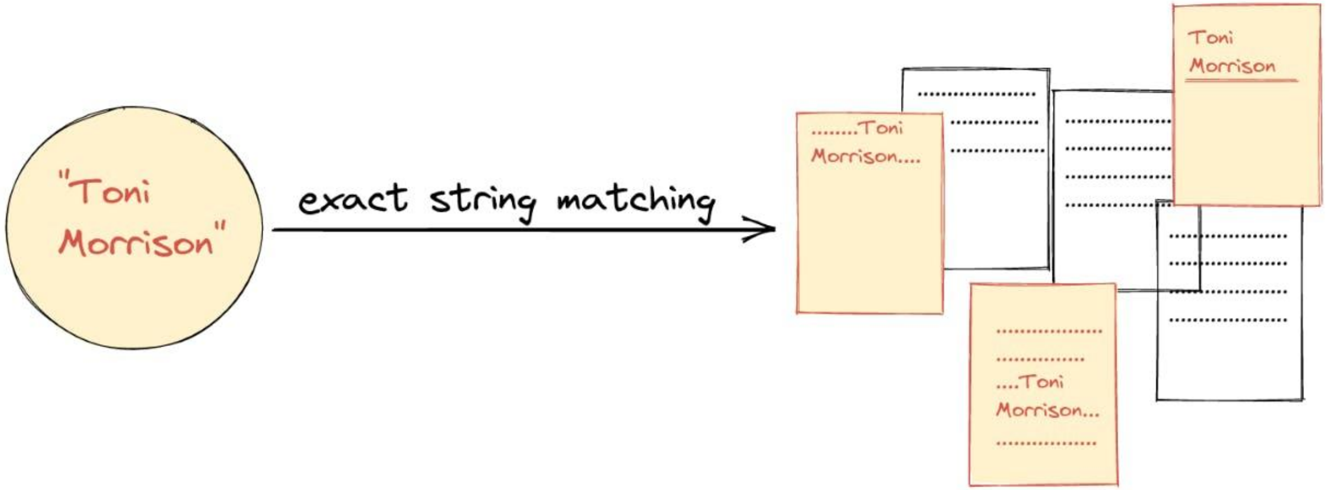




Retrieval

Retrieval

- Minimize candidates of possible documents from millions of passages.
- Question and Passage similarity measure

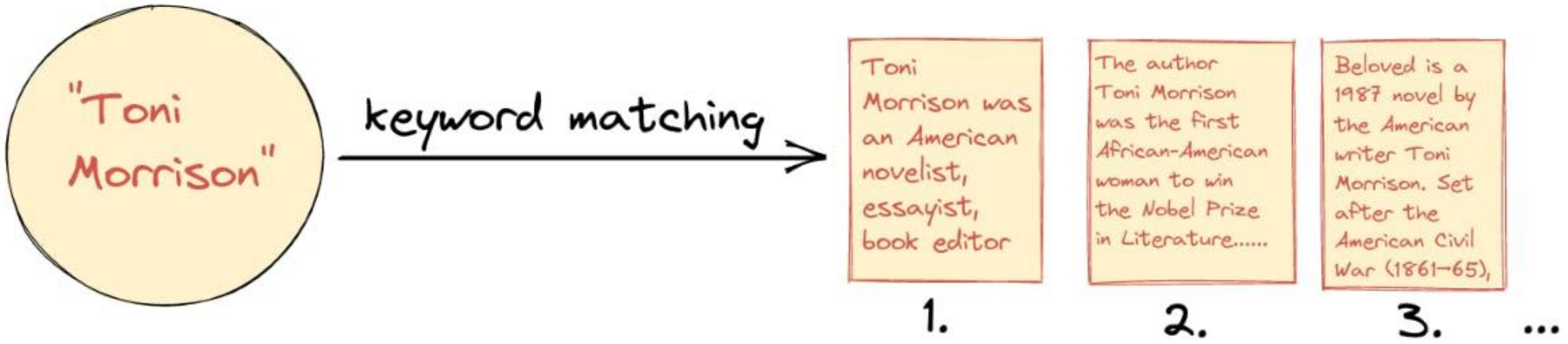


Retrieval

Question and Passage similarity measure

Sparse Retriever

- Bag Of Words (BOW)
- TF-IDF

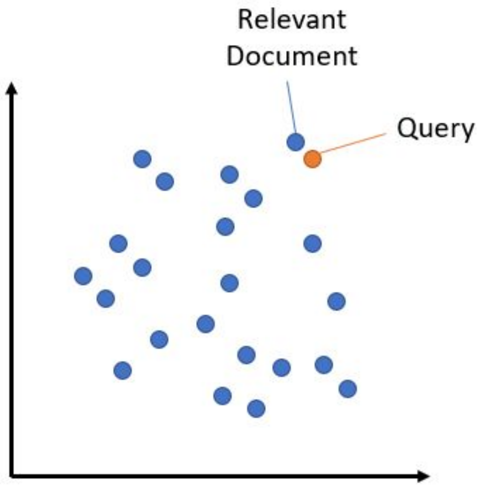


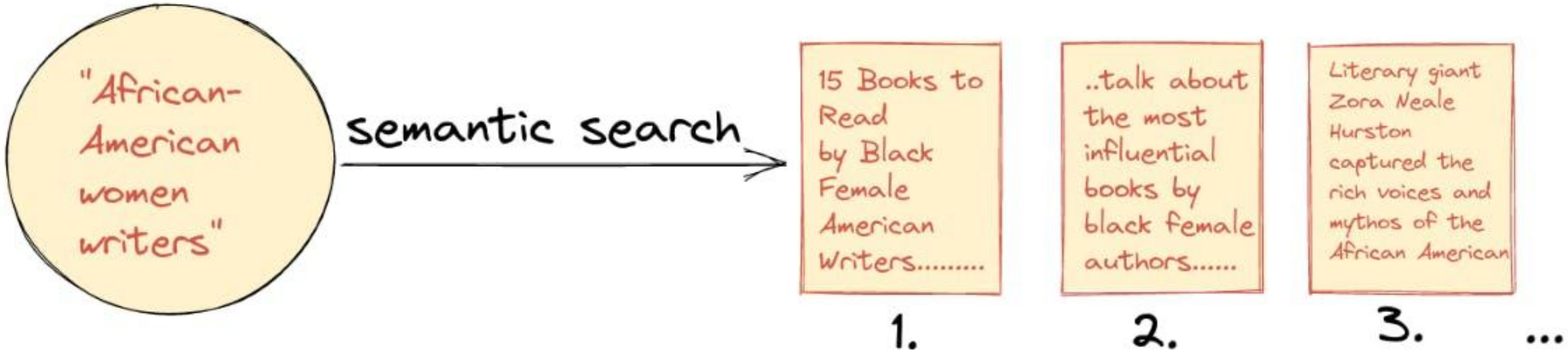
Retrieval

Question and Passage similarity measure

Dense Retriever

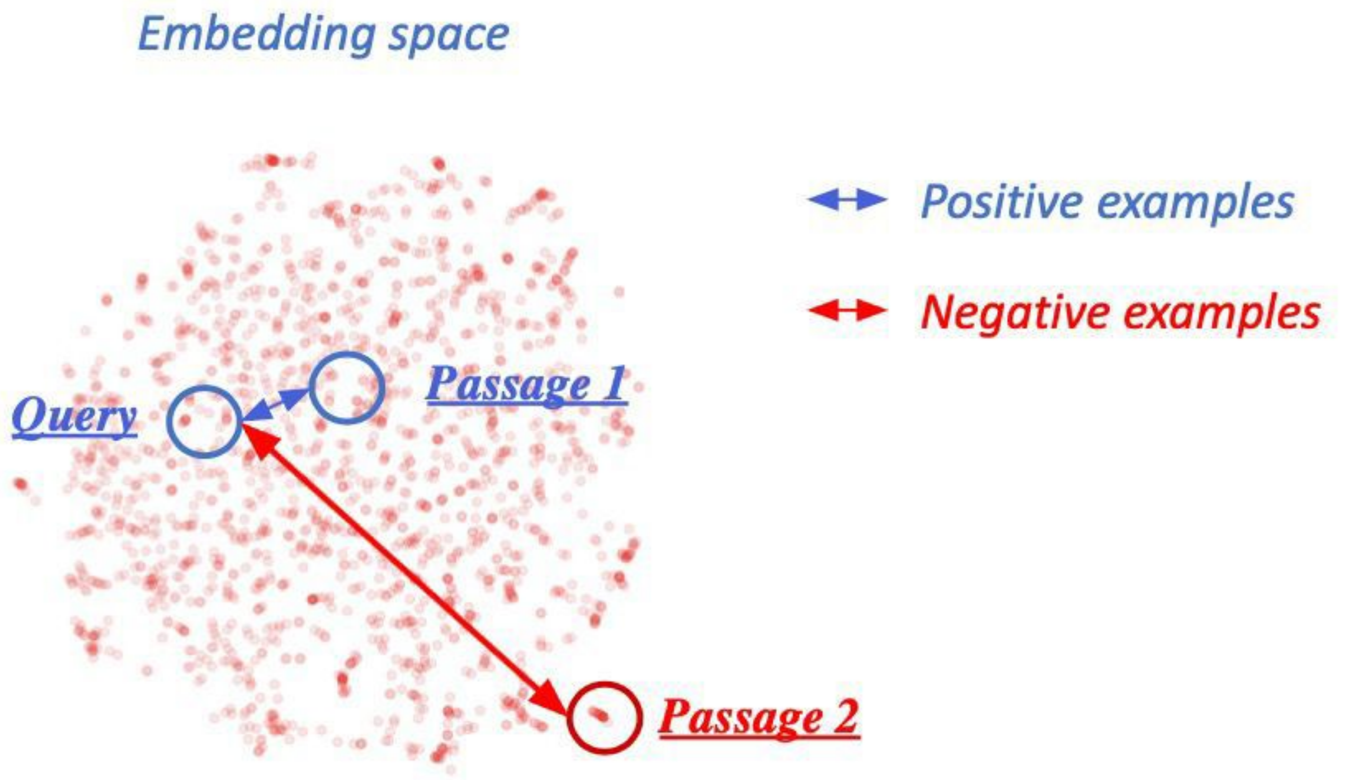
- Query: question
- Passage: document





Retrieval

Dense Retriever



References

- Overall
  - https://d2l.ai/chapter_natural-language-processing-applications/finetuning-bert.html
- QA
  - http://web.stanford.edu/class/cs224n/slides/cs224n-2021-lecture11-qa-v2.pdf (CS224n)
  - https://medium.com/analytics-vidhya/question-answering-system-with-bert-ebe1130f8def
  - https://blog.paperspace.com/how-to-train-question-answering-machine-learning-models/
- Retrieval
  - https://www.deepset.ai/blog/understanding-semantic-search
- Word-piece tokenization
  - https://ai.googleblog.com/2021/12/a-fast-wordpiece-tokenization-system.html