Tasks

Tasks within ML and NLP

- Question Answering
- Coreference Resolution
- Keyword Extraction
- Relationship Extraction
- Federated Learning
- Large Scale Multi-Label Learning

Question Answering

Approaches

Fine-Tuning Sentence-BERT for Question Answering

CapitalOne produced a tutorial (mirror) about using sentence-transformers for Question Answering.

- They use SBERT because it is optimised for fast compute on individual sentence and has good general performance on a number of NLP tasks. They are most interested in the STS capability of SBERT
- They fine-tuned the model on 7k utterances and saw a huge improvement in performance:
  - 52% match rate on 200 test Q-A pairs with no fine tuning
  - 79% match rate on the 200 sample dataset after fine tuning
- They use triplet loss (minimise distance between alike sentences, maximise distance between different sentences)
- This is a nice simple approach where you want to present the answer, as-written, back to the user but its not appropriate for factoid type use cases where we want to highlight the exact word or phrase in the answer text.

Haystack

Haystack is an open source NLP framework for use cases involving large collections of documents. It could be used for searching and ranking type use cases and question answering type use cases.

Haystack works flexibly with existing document stores including DB systems and ElasticSearch.

Coreference Resolution

Co-reference Resolution (CR) is the task of deciding whether two entity mentions refer to the same instance or not.

For example in:

> Joe Biden appeared at the event at 8pm. The president was wearing a Louis Vuitton Tuxedo.

The objective is to identify that Joe Biden and The president are the same entity.

Coreference Detection is related to Relationship Extraction (RE) - in fact you could even say that CR is a special case of RE in the sense that we are interested in the special relationship between entity mentions when they both refer to the same entity.

In-Document Coreference Resolution

This is the "normal" CR case in which you're trying to resolve mentions of entities within the same document e.g. a single news article.

Approaches

- **2022-10-23** A recent blog post from explosion / spaCy shows how they have implemented end-to-end CR in their excellent NLP pipeline but as of writing they do not provide a trained model and they require you to have a copy of the Ontonotes dataset.

Cross-Document Coreference Resolution

Cross-Document Coreference Resolution (CDCR) is when you try to link named entity references across multiple input documents. A use case might be identifying that a number of news articles do actually refer to the same person (e.g. "Joe Biden", "The President").

CDCR is challenging because there are so many possible entities and thus $O(n^2)$ comparisons to make between candidates.

Approaches

- In 2021 we proposed CD^2CR - a CDCR approach across documents and domains that allows us to match mentions of people, places, technologies etc across scientific papers and news articles that discuss them.

Keyword Extraction

Graph-Based Keyword Extraction

Graph-based approaches like TextRank allows the extraction of keywords + phrases based on their centrality to the semantics of the other words in the document.

- https://github.com/SkBlaz/rakun2 - RaKUn 2.0, a very fast keyphrase extraction algorithm suitable for large-scale keyphrase retrieval. It has an associated preprint.

Relationship Extraction

Relationship Extraction (RE) is a task that is related to Coreference Resolution but with a focus on identifying relationships between entities.

In the following example:

> James, the CTO at Filament AI, lives in the South of England.

We want to identify the following relationships:

```
(James, isCTOof, Filament AI)
(James, livesIn, England)
```

Approaches

- FewRel by the NLP group at Tsinghua University propose a few-shot model which takes 2 entity mentions, in context, and N possible relationship types - it predicts which of the relationships is most likely to apply. It can also predict when none of the relationships are applicable.

Federated Learning

- Flower is a federated learning framework with compatibility with Torch, Tensorflow and others

Large Scale Multi-Label Learning

The Keras website has a tutorial on how to do multi-label learning with a large number of labels:

- https://keras.io/examples/nlp/multi_label_classification/ (mirror)