1 | problem types

1.1 | seq2seq

Given an input sequence, generate an output sequence of tokens.

1.2 | question answering (QA)

given a question, find or generate an answer

1.2.1 | free form question answering

generate answer text given a question (and optionally, context)

1.2.2 | open domain question answering (OpenQA, ODQA)

the model is only given a question (no context)

1. closed book question answering (CBQA)
   the model answers the question directly
2. open book question answering
   the model answers using information from a knowledge base, knowledge graph, corpus, or other info source
   (a) distantly supervised open-domain question-answering (DS-QA)
       find answers in collections of unlabeled text (open book question answering on a corpus)

1.3 | reading comprehension (RC)

close reading to understand a short (paragraph) of natural text, usually to answer questions

1.4 | information retrieval (IR)

given a question, the task of finding relevant paragraphs from a corpus (ex. by the retriever to pass to the reader)

2 | internal representation

2.1 | knowledge graph (KG)

a set of entities and relationships between them, ex. 'barak obama' 'in' 'us presidents', and 'barak obama' 'in' 'fathers'. details are highly implementation dependent

2.2 | corpus

raw, natural text. for example, the wikipedia corpus is the raw wikitext of wikipedia (maybe without links, etc. but human text.)

3 | model architectures

3.1 | retriever-reader

The model consists of a retriever model and a reader model.

3.1.1 | retriever

Smaller model that scans the corpus for relevant paragraphs to pass to the reader

3.1.2 | reader

Larger model that does reading comprehension on the input context and the question. It may predict the answer span, or generate a free-form response, etc.

3.2 | MLP (multi-layer perceptron)

Bog standard feed forward neural network, wikipedia

4 | techniques

4.1 | answer re-ranking

Store a bunch of possible answers in memory, and re-rank them (as new information (more paragraphs) are processed, or by considering other passages that came up with the same answer). Then, the answer will have to agree with multiple paragraphs.

4.1.1 | strength-based re-ranking

find sources that corroborate an answer by seeing how many passages say any given answer is likely, and how much the passage supports that answer

4.1.2 | coverage-based re-ranking

for each answer, concat contexts that gave that answer and see if the expanded context answers the question better ’could entail the question?’ via match-lstm (or other means)

4.1.3 | sources

1. Wang et al. evidence aggregation for answer reranking in odqa

4.2 | term frequency, inverse document frequency (TF-IDF)

Used for scoring how related a document is to a word. good for IR and keyword extraction

Product of term frequency (how often a term appears in a document) and inverse document frequency (how often that term appears in other documents).

4.2.1 | **term frequency**

Number of times a word appears in a document, normalized either by the number of occurrences of the most common word or by length of the document.

4.2.2 | **inverse document frequency**

log((Number of documents) / (number of documents containing the word))

Between 0 and 1, with 0 meaning a common word and ~1 meaning a very rare word

4.2.3 | **sources**

1. `https://monkeylearn.com/blog/what-is-tf-idf/`