Complex Answer Retrieval Track

Summary
The current retrieval systems provide good solutions towards phrase-level retrieval for simple fact and entity-centric needs. The focus of the track is on developing systems that are capable of answering complex information needs by collating relevant information from an entire corpus, much like Wikipedia pages which synthesize knowledge that is globally distributed.

Task
The task is to retrieve, select, and arrange a sequence of k passages P from the provided passage corpus given an outline Q, with ideally:

1. Highest relevance of all passages.
2. Balanced coverage of all query facets as defined through headings in the outline.
3. Maximizing topical coherence, minimizing topic switches, i.e., first all passages about one topic, then all passages of the next topic while avoiding to interleave multiple topics. The number of passages k is given with the topic.

Data
The following data is provided-
Paragraph Corpus: A corpus of 20 million paragraphs that are harvested from paragraphs on Wikipedia pages from a snapshot of 2016 (with hyperlinks preserved). The collection was de-duplicated before the data release. These paragraphs are to be used for the passages ranking task.

AllButBenchmark/knowledge base: Nearly all Wikipedia pages (2016) are provided as standardized information on entities for training. Full meta information for each page including disambiguation pages and redirects, as well as the full page contents, is provided. To avoid train/test signal leakage, the dump is also offered as five-folds that are consistent with the large train corpus.

Queries/Outlines: Query outlines structured into title, and a hierarchy of headings are provided.

Evaluation of Runs
Training/evaluation annotations are provided in the trec_eval compatible Qrels format. Both automatic (binary) and a manual (graded) evaluation procedures are used.
automatic benchmark: Wikipedia pages are split into paragraphs and outlines to generate a large-scale automatic benchmark for training and automatic evaluation and asking the system to put paragraphs back to their place of origin. For each section, in training stubs Q the passages

originating from the section are indicated. To derive a ground truth for entity ranking, the entities linked from the section are indicated.

Approach to the task

Lucene's BM25 implementation can be used to create a candidate set. The set can be re-ranked by using neural networks or Markov Chain models.