The activityNet-Entity dataset consists of multiple documents. Each document is consisting of one or more than one video segments, and each video segment has a segment id and process class. Every video segment has video frames and textual part.

Each video frame has attribute name called frame-index and crowds. Each video frame showing one or more than one bounding box. Each bounding box has attributes name bounding-box id and process-bounding box. Process bounding box has values ybr (y coordinate of the bottom right corner of bounding box), ytl (y coordinate of the top left corner of bounding box), xbr (x coordinate of the bottom right corner of bounding box), and xtl (x coordinate of the top left corner of bounding box).

The textual-part of video-segment is described by captions, and each caption is identified by caption-id. The textual part has tokens, and each token can be either textual-entity or non-textual-entity mention. Textual-entity mentions are linked with DBpedia and WikiData knowledgebases.