TAC 2014 Biomedical Summarization Task Description

Introduction

The Biomedical Summarization track of TAC encourages the development of systems that can summarize biomedical research papers. The automatic biomedical summarization task is defined as follows:

Given: A topic consisting of a Reference Paper (RP) and 10 Citing Papers (CPs) that all contain citations to the RP. In each CP, the text spans (i.e., citances) have been identified that pertain to a particular citation to the RP.

Task 1a: For each citance, identify the spans of text (cited text spans) in the RP that most accurately reflect the citance. These are of the granularity of a sentence fragment, a full sentence, or several consecutive sentences (no more than 5).

Task 1b: For each cited text span, identify what facet of the paper it belongs to, from a predefined set of facets.

Task 2: Finally, generate a structured summary of the RP and all of the community discussion of the RP represented in the citances. The length of the summary should not exceed 250 words.

Training Data

Twenty topics are provided as training data for the track. Each topic contains a Reference Paper and 10 Citing Papers, annotations of the citations to the Reference Paper that are found in the Citing Papers, and a summary of the Reference Paper that takes into consideration not only the abstract for the Reference Paper itself, but also those parts of the Reference Paper that were mentioned by the Citing Papers and how the Citing Papers discussed the Reference Paper. Each topic contains annotations and summaries from 4 different human annotators.

Each annotation file consists of the following pipe "|" separated fields:

1. Topic ID: An ID for the topic, e.g., "D1401_TRAIN"
2. Citance Number: A numeric ID for the citance for this topic, keyed by a unique combination of Reference Article, Citing Article, and Citation Marker Offset.
3. Reference Article: Name of the reference paper in Documents_Text
4. Citing Article: Name of the citing paper in Documents_Text
5. Citation Marker Offset: pair of start and end offsets defining the span of text in the Citing Article for the Citation Marker of the citance.
6. Citation Marker: Span of text defined by the Citation Marker Offset above.
7. Citation Offset: Pair of start and end offsets defining the span of text in the Citing Article that contains the citation and conveys the authors' discussion of the citation.
8. Citation Text: Span of text defined by the Citation Offset above. Although the exact Citation

Text span may vary between annotators, all Citation Text spans for the same Topic ID and Citance Number must have some overlap (minimally containing the same Citation Marker).

9. Reference Offset: One to three pairs of offsets defining the span(s) of text in the Reference Article that the authors of the Citing Article are discussing in the citation.
10. Reference Text: Span(s) of text defined by the Reference Offset above. Discontiguous spans defined by multiple pairs of offsets are separated by ellipses "..." and do not necessarily appear in the same order as their corresponding offset spans in Field 9.
11. Discourse Facet: one of {Hypothesis_Citation, Method_Citation, Results_Citation, Implication_Citation, Discussion_Citation}.
12. Annotator: An alphabetic ID for the annotator

For each unique combination of Topic ID and Citance Number, the annotator uses his/her judgment to determine Citation Offset (which defines Citation Text), Reference Offset (which defines Reference Text), and Discourse Facet.

Different annotators may disagree about what constitutes the exact boundaries of the citance (i.e. the Citation Offset may differ between two annotators even for the same Topic ID and Citance Number); however, the individual citances returned by the 4 annotators must overlap (minimally at the citance marker), and the citance for a given Topic ID and Citance Number can be taken to be the union of the citances returned by the 4 annotators.

Evaluation Data

30 topics are provided as evaluation data. Each evaluation topic contains 1 Reference Paper and 10 Citing Papers, and 4 annotation files (from 4 different human annotators) that are in the same format as for the training data, except that fields 9-12 are omitted.

System Output

Task 1

A single file consisting of exactly one line for each unique combination of Topic ID and Citance Number in the annotations in the Evaluation Data. Each line consists of the following pipe "|" separated fields:

1. Topic ID: An ID for the topic, e.g., "D1401_EVAL" (From the Evaluation Data annotations.)
2. Citance Number: the numeric ID for the citance for this topic keyed by a unique combination of Reference Article, Citing Article, and Citation Marker Offset. (From the Evaluation Data annotations.)
3. Reference Offset: One to three pairs of offsets defining the portions of the Reference Article that the authors of the Citing Article are discussing in the citation. (Same format as Field 9 in the training data annotation files)
4. Reference Text: Span of text defined by the Reference Offset above. Discontiguous spans defined by multiple pairs of offsets are separated by ellipses "..." This field is for readability only and may be left empty. (Same format as Field 10 of the training data annotation files).
5. Discourse Facet: one of {Hypothesis_Citation, Method_Citation, Results_Citation,

Implication_Citation, Discussion_Citation}. (Same format as Field 11 of the training data annotation files).

6. Run ID: A unique ID for the submission. The run ID should be the team name followed by a number (1-5) for the submission.

Task 2

A run will comprise exactly one file per Topic ID on the Evaluation Data, containing the summary of the Reference Paper for that topic that takes into consideration not only the abstract for the Reference Paper itself, but also those parts of the Reference Paper that were mentioned by the Citing Papers and how the Citing Papers discussed the Reference Paper. Each summary file must be in UTF-8 and have the same name as the Topic ID (e.g., "D1401_EVAL"). Please include a summary file for each Topic ID, even if the file is empty. The files must be in a directory whose name should be the concatenation of the Team ID and a number (1-2) for the run. (For example, if the Team ID is "SYSX" then the directory name for the first run should be "SYSX1".) Please package the directory in a tarfile and gzip the tarfile before submitting it to NIST.

Scoring

Task 1a

Evaluation is based on overlap between the RP span in the system output vs. those in the gold standard output. For a pair of character offsets (i, j), the ***span*** of (i, j) is the set of non-negative integers $K = \{k: i \leq k < j\}$; the span of multiple offset pairs is the union of the span of each offset pair.

For a given Topic ID + Citance Number, we define $WeightedRecall(S|M)$ and $WeightedPrecision(S|M)$ for a System returning RP span S, with respect to a set M of annotations from m humans, containing RP spans $G_1, G_2, \dots, G_m$:

$$WeightedRecall\ (S|M = \{G_1, G_2, \dots G_m\}) \overset{\text{def}}{=} \frac{|S \cap G_1| + |S \cap G_2| + \cdots + |S \cap G_m|}{|G_1| + |G_2| + \cdots + |G_m|}$$

$$WeightedPrecision\ (S|M = \{G_1, G_2, \dots G_m\}) \overset{\text{def}}{=} \frac{|S \cap G_1| + |S \cap G_2| + \cdots + |S \cap G_m|}{m \times |S|}$$

The per-citance score of the System is F1 of $WeightedRecall(S|M)$ and $WeightedPrecision(S|M)$ given the set of 4 human annotations, and the overall score of the System is the mean F1 over all Topic ID + Citance Number combinations in the Evaluation Data.

The metric is similar to ROUGE-1 in that it give more credit for returning character offsets that are in multiple human annotations, but also includes a precision-oriented component since there is no predetermined size limit for the RP span that the System can return. $WeightedRecall = 1$ is achieved by returning all character offsets that are in the union of the RP spans returned by the humans; $WeightedPrecision = 1$ is achieved by returning only character offsets that are in the intersection of the RP spans returned by the humans; and perfect F1 is achievable only when the human annotators return identical RP spans.

Task 1b

For a given Topic ID + Citance Number, we define $WeightedAccuracy$ for a System returning discourse facet f:

$$WeightedAccuracy = \frac{\#\ of\ annotators\ who\ returned\ the\ same\ discource\ facet\ f}{\#\ of\ annotators}$$

The Discourse Facet score of a System is the mean $WeightedAccuracy$ over all Topic ID + Citance Number combinations in the Evaluation Data.

Task 2

Each summary submitted by a System will be scored against the human-authored summaries for the same topic using ROUGE-2. All summaries will be truncated to 250 words before evaluation.

Submissions

Evaluation data will be released on the first day of the evaluation window, and participants must submit their system output to NIST before the end of the last day of the evaluation window. Up to 5 different runs may be submitted for each of Task 1 and Task 2. Submitted runs should be ranked according to their expected overall score.