Proposition Salience Guidelines

V0.1

Katherine Conhaim & Amir Zeldes

Preamble

Discourse, whether written or spoken, is not a flat sequence of equally important propositions: some utterances are all but placeholders (“hold on” or “let me tell you”), others are informative but tangential (“incidentally, she had no other neighbors”), and others still may form the main message in a story (“at this turning point, war broke out with Iceland”).

While there can be different ways of measuring the relative salience of propositions in discourse, a good salience metric will be **graded** (ratio-scaled and not just binary salient/non-salient), **consistent** and **comprehensive**, meaning it should involve a minimum amount of subjectivity, while still accounting for the limited salience some evaluators will perceive for more tangential cases. In this project, we rely on **multiple summarization** as the underlying, relatively stable source for salience criteria, by equating the degree of salience with **the proportion of summaries mentioning a proposition**. This allows for some subjectivity (different summaries prioritize different content), but also considerable stability (the most salient propositions will tend to be mentioned in all summaries).

The guidelines below specify what we consider to be propositions, how we determine whether something has been mentioned, and the kinds of annotations that we add to propositions to indicate the nature of their alignment to summary contents. Data for the project is derived from the Georgetown University Multilayer corpus (GUM), an English data set covering a range of spoken and written genres in English.

Markables for Annotation - Propositions & Leftovers

In order to have a decidable and consistent underlying unit of annotation, this project uses the notion of Elementary Discourse Units (EDUs) as defined in Rhetorical Structure Theory (RST) as the target of annotation, specifically following GUM’s RST guidelines. EDUs are typically sentences, or in complex sentences, the smaller clause units that coalesce to form sentences, but also the units that are left over once such sentences and clauses have been segmented. In particular:

- No EDU is larger than a single sentence: *[I saw Kim.] [She was standing outside.]*
- No EDU contains material from more than one speaker: *[You mean she –] [wanted to come too, yes!]*
- EDUs typically corresponds to a single predication, including all of its arguments, which includes infinitival complements and auxiliaries: *[Kim wanted to go home.] [I had been meaning to go too.]*
- In more complex, multi-predicational sentences, each clause is an EDU. This includes adverbial clauses and relative clauses, but not subject or object clauses, which are considered part of the predication of their main verb's EDU: *[I took an umbrella] [because it was raining.]* But: *[They prevented France from joining the competition.] [That France wanted to join alarmed us.]*
- An exception to this rule concerns attribution clauses, which are segmented: *[I know] [that they want to participate.] [They didn't suspect] [that the game was over.]*
- Coordinate verbs are given separate EDUs, unless they share an object: *[I came] [and saw Kim.]* But: *[I greeted and hugged Kim]*
- Propositions may be discontinuous if they are interrupted by other clauses: *[The men]$_1$ [who came by]$_2$ [were dressed in purple.]$_1$*
- Headings and other fragments left over once clauses have been segmented are also made into EDUs: *[... as we have seen in this section.] [2. Methods] [We applied a range of instruments to the problem.]* Although these are not strictly speaking propositions, we must allow for their alignment, since fragments are sometimes the sole trigger for content in a summary. Below when we refer to propositions the term should be understood to possibly refer to such fragments as well.

For more detailed guidelines on EDU segmentation, see the GUM corpus RST annotation guidelines: https://wiki.gucorpling.org/gum/rst

Aligning Propositions to Summaries

For each summary of a document (typically 5 distinct summaries), we align the EDUs that are mentioned in that summary. The process is repeated independently for the same EDUs but with the specific summary in mind.

What is a match?

One guiding principle is to attempt to take the point of view of the summarizer and attempt to infer what portion of a document **triggered** the composition of each part of the summary: triggering propositions are usually considered aligned, though in weaker/incomplete equivalence cases, we speak of an **approximate** alignment (see below).

Multiple matches

If multiple propositions match the same part of the summary, all of them should be highlighted, and all non-initial instances should be linked back to the initial instance's ID (visible on hover). See more under Duplicates below.

Alignment subcategories

Subcategory overview

Normal match - the default alignment type, when propositional content matches summary content well.
Approximate - the match between a proposition in the document and the summary is close enough that we believe it is a trigger for words in the summary, but there are actually substantial differences between the propositions.
Component - a superordinate class aggregating over multiple propositions in the document is mentioned in the summary; the subordinate multiple propositions are tagged **component**.
Duplicate - a proposition may be considered a trigger for part of the summary, but there is another proposition that could equally.

Approximate alignment

If an EDU contains a substantial amount of information which is clearly paralleled in the summary, we generally do not select the approximate alignment type, even if some amount of additional information appears in the proposition which is not paralleled in the summary. In particular, modifiers of entities mentioned in the proposition but not in the summary are not cause for approximate alignment:

> **Summary:** President Macron ordered demonstrators to be cleared away…
>
> [1] French President Emanuel Macron ordered …
>
> Alignment: normal match (the added modifiers "French" and the first name "Emanuel" appearing in the proposition are not grounds for approximate alignment.

By contrast in the following example, the statement in EDU [2] appears to be the trigger for "worsening financial conditions".

> **Summary:** Amid worsening financial conditions in Portugal, the European Union is considering…

[1] Portugal used to be easy living. [2] Now no one can afford to own a car anymore.

Alignment: [2]: approximate, since the summary does not mention anything about affording a car, but this seems to be the trigger for the financial conditions statement. In cases like this, we align, select “approximate”, and use the Notes box to indicate the reason in the format: “triggers worsening financial conditions”.

Note that if this is also the only mention of Portugal, then [1] is also an approximate alignment (triggers Portugal), but not a normal match, since formerly being “easy living” is the main predicate, which is not mentioned.

Duplicates

In cases where a summary portion corresponds to more than one segment of the text but those multiple segments convey roughly the same meaning, we call the subsequent mentions in the text of the repeated content “duplicates”, and note the ID of the initial mention of the duplicated information. To add the repetition index, click the note icon next to the non-initial span and select the ID displayed when hovering over the initial span in the ‘duplicate’ box.

By convention, duplicates are annotated at the non-initial mention, and refer back to the earliest EDU ID of the duplicated span. For example:

> **Summary:** Muhammad Ali was an invincible boxer.
>
> [1] Muhammad Ali, [2] born Cassius Marcellus Clay Jr., [3] was an American professional boxer and activist. [4] He was often referred to as invincible: [5] his fans said that he could not be defeated.

In this example, EDUs [4] and [5] could both be considered to trigger the inclusion of “invincible” in the summary, but the underlined EDU [5] is considered a duplicate of the earlier EDU [4]. As such, [5] should be annotated as a duplicate of [4]. Note that in order to prevent inter-annotator disagreements, the later unit is always annotated as the duplicate, even if an annotator judges the later one to be a ‘better’ match subjectively - as soon as two units are considered to have a duplication relation, the earliest mention is the anchor for duplication. This also applies to more than 2 mentions: mentions 2, 3, 4 etc. will always point back to the ID of the first EDU of the first mention.

Also note that duplication is not the same as coreference: it is possible for two EDUs to mention the same entities or events, but if only one of them is considered a trigger for the summary, other coreferring spans are not necessarily required to be selected. For example:

Summary: the author complains about a tax giveaway to the rich.

[1] On Wednesday, Jeff Bezos was awarded a tax exempt contract by that company. [2] This is just a giveaway to the rich sponsored by the American people.

In this example, [2] triggers the "tax giveaway" portion of the summary. The events in [1] are portrayed as coreferring to the giveaway, which is "the same thing in the world" as the "contract"; however the "contract" EDU [1] is not necessary as a trigger given the existence of [2], and [2] is not a duplicate of [1]. By contrast, the two duplicate units in the Muhammad Ali example are both sufficient for triggering the "invincible" part of the summary.

Component matches

If the summary mentions a superset of several propositions, each of these is considered **component mentioned**. The corresponding propositions should be highlighted, then using the note icon, the category **component** should be selected.

Note that the goal of component annotation is to ensure that superordinate propositions in the summary are not left unaligned, but at the same time, we want to avoid aligning the entire document, or very large parts of it, to general predicates in the summary, such as "this article describes…", which could apply to nearly all propositions. We therefore apply a **maximum component cutoff**:

No component item may correspond to more than **five** EDUs:

- If there are more than five candidates, but they are not equally prominent or representative, a subset of up to the top 5 most compelling alignments may be selected for each component item

Example:

Summary: Governor Ross details his reasons for the recent tax cuts…

[1] I think [2] higher education should be affordable for everyone. [3] The state needs money [4] in order to make these plans feasible. [5] If we don't invest in higher education now, [6] you will see a brain drain [7] just like it happened in neighboring states 10 years ago. [8] We don't want a brain drain like that here.

Alignments: 2, 3, 5, 6, 8 (all 'component'); 8 -> duplicate: 6

Component alignments are by nature not exact, but are only marked as ‘approximate’ if they are: 1. A trigger, but not a good match for what the summary states (not exactly a member in a mentioned set); 2. Mention substantially more information than just the component of the mentioned set. In such cases, we select both ‘component’ and ‘approximate’.

Specific constructions

Questions and answers

If some of the aligned information comes from the answer to a question, prefer highlighting the answer and not the question, unless:

- Some part of the information is only found in the question
- The summary mentions the act of posing a question in some way (including: “Wikinews interviews so-and-so about…”)

Nominal mentions

Headings - a heading can be considered mentioned if it triggers information in the summary, for example “Section 1” is considered mentioned if the summary says “This section…” (since otherwise the summarizer would not know this is specifically a section). Similarly “Method 1” is aligned if the summary mentions “a method to…”, or will be component aligned if the summary mentions “methods” in aggregate.

Names and dates - if the summary mentions a date or time which is inferable from a dateline, then the dateline has been mentioned. If the date or time is also inferable elsewhere in the document, then it is part of a duplicate mention. Similarly if someone’s full name appears in an image caption, heading or letter signature, and the name is also mentioned in the summary, then the caption, heading or other such unit is considered ‘mentioned’ since it triggers the name.

If the caption or heading consists solely of a name or date which is mentioned in the summary, then it is a normal match, and not ‘approximate’, since that unit is literally mentioned in the summary. All such units (mentions of the same name and nothing else in the unit) would be duplicates of each other.

If a name appears in full only in a unit whose proposition is not mentioned in the summary, but the summary mentions the name and there is no other source triggering it, then the unit mentioning the name is considered an ‘approximate’ mention (essential for triggering the information, but not a good alignment).

If there are multiple possible sources for a name and none of them meet the normal criteria for their proposition being mentioned, we prefer:

1. Fragment mentions - prefer a plain mention EDU consisting of the name to one with an additional predicate, not mentioned in the summary. In this case, any subsequent mention of the name as an entire fragment EDU *is* a duplicate of the first fragment mention, and the alignment is *not* approximate (since the name is literally mentioned).
2. Earliest proposition mention - if there is no such fragment, and the name only appears in a proposition that does not appear in the summary, align only the earliest EDU mentioning the name, and mark it approximate. In this case, subsequent EDUs with the name are *not* aligned as duplicates.

Note that fragments such as names can also be components (for example if a summary mentions a family, but the text only has each member of the family in a separate EDU). Again the same preferences apply (e.g. for each family member, prefer a fragment, and handle duplicates as above).

Propositions triggering modifiers - if a nominal modifier such as a participle used adjectivally refers to a predication, then a full clause headed by a corresponding predicate is considered mentioned. For example, if a summary mentions a “home made clock” and the document contains a sentence saying that someone “made the clock at home”, then that proposition has been mentioned, and this is considered a normal match. Note that as always, some deviations are allowed (for example, the clock is understood to be made at home, even though “at home” does not actually appear in the aligned sentence headed by “made”).