Shallow Discourse Parsing for Open Information Extraction and Text Simplification

Christina Niklaus
Institute of Computer Science
University of St.Gallen
christina.niklaus@unisg.ch

André Freitas
Department of Computer Science
University of Manchester
Idiap Research Institute
andre.freitas@manchester.ac.uk

Siegfried Handschuh
Institute of Computer Science
University of St.Gallen
University of Passau
siegfried.handschuh@unisg.ch

Abstract

We present a discourse-aware text simplification (TS) approach that recursively splits and rephrases complex English sentences into a semantic hierarchy of simplified sentences. Using a set of linguistically principled transformation patterns, sentences are converted into a hierarchical representation in the form of core sentences and accompanying contexts that are linked via rhetorical relations. As opposed to previously proposed sentence splitting approaches, which commonly do not take into account discourse-level aspects, our TS approach preserves the semantic relationship of the decomposed constituents in the output. A comparative analysis with the annotations contained in RST-DT shows that we capture the contextual hierarchy between the split sentences with a precision of 89% and reach an average precision of 69% for the classification of the rhetorical relations that hold between them. Moreover, an integration into state-of-the-art Open Information Extraction (IE) systems reveals that when applying our TS approach as a preprocessing step, the generated relational tuples are enriched with additional meta information, resulting in a novel lightweight semantic representation for the task of Open IE.

1 Introduction

Sentences that present a complex structure can be hard to comprehend by human readers, as well as difficult to analyze by semantic applications (Mitkov and Saggion, 2018). Identifying grammatical complexities in a sentence and transforming them into simpler structures is the goal of syntactic TS. The most relevant method that is used to perform this rewriting step is *sentence splitting*: it divides a sentence into several shorter components with each of them presenting a more regular syntax that is easier to process by both humans (Siddharthan and Mandya, 2014; Ferrés et al., 2016) and machines (Štajner and Popović, 2018; Saha and Mausam, 2018).

We propose a sentence splitting approach that can be used as a preprocessing step to generate an intermediate representation. The objective is to facilitate and improve the performance of downstream tasks whose predictive quality deteriorates with sentence length and complexity (e.g., see Cetto et al. (2018); Saha and Mausam (2018); Heilman and Smith (2010); Štajner and Popović (2018)). Our approach aims to **break down a complex sentence into a set of minimal propositions**, i.e. a sequence of sound, self-contained utterances with a simple and regular structure. Each of them presents a minimal unit of coherent information and, consequently, cannot be further decomposed into meaningful propositions. However, any sound and coherent text is not simply a loose arrangement of self-contained units, but rather a logical structure of utterances that are semantically connected (Siddharthan, 2014). Consequently, when carrying out syntactic TS operations without considering discourse implications, the rewriting may easily result in a disconnected sequence of simplified sentences, making the text harder to interpret. The vast majority of existing structural TS approaches though do not take into account discourse-level aspects. Therefore, they are prone to producing a set of incoherent utterances where important contextual information is lost. Thus, to **preserve the coherence structure** of the input we propose a context-preserving TS approach. It establishes a semantic hierarchy between the split components by (1) setting up a contextual hierarchy and (2) classifying the semantic relationship that holds between them (see Figure 1).

To the best of our knowledge, this is the first time that syntactically complex sentences are *split and rephrased within the semantic context* in which they occur. Our framework differs from previously proposed approaches by using a linguistically grounded transformation stage that applies clausal and phrasal disembedding mechanisms to

Proceedings of 3rd Workshop on Computational Approaches to Discourse, pages 64–76
October 16 – 17, 2022.

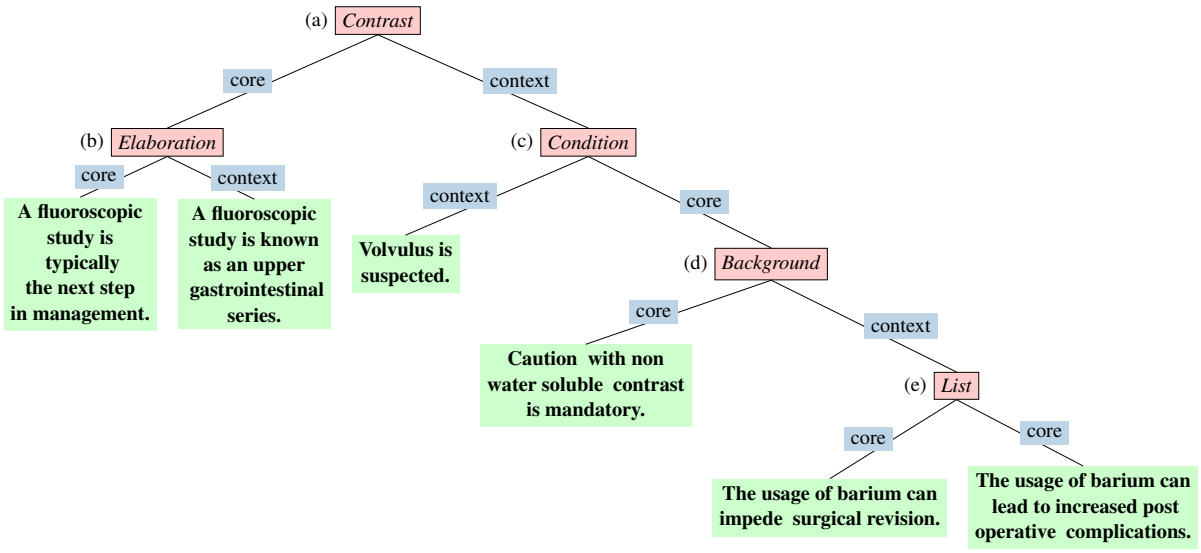

Figure 1: A complex sentence (*"A fluoroscopic study which is known as an upper gastrointestinal series is typically the next step in management, although if volvulus is suspected, caution with non water soluble contrast is mandatory as the usage of barium can impede surgical revision and lead to increased post operative complications."*) is transformed into a semantic hierarchy of simplified sentences in the form of minimal, self-contained propositions that are linked to each other via rhetorical relations. The output presents a regular, fine-grained structure that preserves the context of the input in the form of hierarchically ordered and semantically interconnected sentences.

transform sentences into shorter utterances with a more regular structure. By using a recursive top-down approach, it generates a novel *hierarchical representation* between those units, capturing both their semantic context and relations to other units in the form of rhetorical relations.[1] By taking advantage of the resulting fine-grained representation, the complexity of downstream tasks may be reduced, thus improving their performance. In addition, by incorporating the semantic context of the source sentences, our proposed representation preserves contextual information that is needed to maintain the coherence structure of the input, allowing for a proper interpretation of complex assertions.

In summary, we make the following contributions: (i) We propose a discourse-aware syntactic TS approach which transforms complex sentences into a semantic hierarchy of minimal propositions, resulting in a novel representation that puts a semantic layer on top of the simplified sentences. (ii) The proposed method is linguistically grounded and does not require any training data. (iii) As a proof of concept, we develop a reference implementation. (iv) We perform a comprehensive empirical evaluation, demonstrating that we reach state-of-the-art performance in the classification of both the hierarchical order and the semantic relationship that hold between the split sentences. (v) We show that the semantic hierarchy can be leveraged to extract relational tuples within their semantic context, resulting in a novel lightweight semantic representation for complex text data in the form of normalized and context-preserving tuples.

2 Discourse-Aware Sentence Splitting

We present DISSIM, a discourse-aware TS approach that creates a semantic hierarchy of simplified sentences.[2] It takes a sentence as input and performs a recursive transformation stage that is based upon a small set of 35 hand-crafted rules.

2.1 Transformation Patterns

In the development of the transformation patterns, we followed a principled and systematic procedure, with the goal of eliciting a universal set of transformation rules. They were heuristically determined in a rule-engineering process that was carried out on the basis of an in-depth study of the literature on syntactic sentence simplification, e.g. Siddharthan (2006, 2014, 2002); Siddharthan and Mandya (2014); Evans and Orăsan (2019); Ferrés et al. (2016). Next, we performed a thorough lin-

[1] For this purpose, we make use of a subset of the classical set of RST relations defined in Mann and Thompson (1988) that we adapted from the work of Taboada and Das (2013).

[2] The source code of our framework is available under https://github.com/Lambda-3/DiscourseSimplification (Niklaus et al., 2019a).

guistic analysis of the syntactic phenomena that need to be tackled in the sentence splitting task.[3] The transformation patterns encode syntactic and lexical features that can be derived from a sentence's phrase structure. Each rule specifies (1) how to *split up and rephrase* the input into structurally simplified sentences and (2) how to *set up a contextual hierarchy* between the split components and how to *identify the semantic relationship* that holds between those elements.[4]

2.2 Data Model: Linked Proposition Tree

The transformation algorithm takes a complex sentence as input and recursively transforms it into a semantic hierarchy of minimal propositions. The output is represented as a linked proposition tree (LPT). Its basic structure is depicted in Figure 2. A LPT is a labeled binary tree $LPT = (V, E)$.

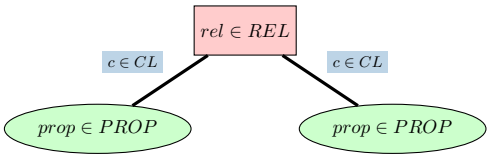

Figure 2: Basic structure of a LPT. It represents the data model of the sem. hierarchy of min. propositions.

Let $V \in \{REL, PROP\}$ be the set of nodes, where $PROP$ is the set of leaf nodes denoting the set of minimal propositions. A $prop \in PROP$ is a triple $(s, v, o) \in CT$, where $CT = \{$*SV, SVA, SVC, SVO, SVOO, SVOA, SVOC*$\}$ represents the set of clause types (Del Corro and Gemulla, 2013). Hence, $s \in S$ denotes a subject, $v \in V$ a verb and $o \in \{O, A, C, OO, OA, OA, \emptyset\}$ a direct or indirect object, adverbial or complement (or a combination thereof). Accordingly, a minimal proposition $prop \in PROP$ is a simple sentence[5] that is reduced to its clause type.[6] Thus, it represents a minimal unit of coherent information where all optional constituents are discarded, resulting in an utterance that expresses a single event consisting of a predicate and its core arguments.

Furthermore, let $REL = \{$*Contrast, List, Disjunction, Cause, Result, Temporal, Background, Condition, Elaboration, Explanation, Spatial, Attribution, Unknown*$\}$ be the set of rhetorical relations, comprising the set of inner nodes. A $rel \in REL$ represents the semantic relationship that holds between its child nodes. It reflects the semantic context of the associated propositions $prop \in PROP$. In that way, the coherence structure of the input is preserved.

Finally, let $E \in CL$, with $CL \in \{$*core, context*$\}$, be the set of constituency labels. A $c \in CL$ represents a labeled edge that connects two nodes $V \in LPT$. It enables the distinction between core information and less relevant contextual information. In that way, hierarchical structures between the split propositions $prop \in PROP$ are captured. Figure 1 shows the LPT that is generated by our TS approach on an example sentence.

2.3 Transformation Algorithm

Algorithm 1 Transform into Semantic Hierarchy

Input: complex source sentence str
Output: linked proposition tree $tree$

```
function INITIALIZE(str)
    new_leaves ← source sentence str
    new_node ← create a new parent node for new_leaves
    new_node.labels ← None
    new_node.rel ← ROOT
    linked proposition tree tree ← initialize with new_node
    return tree
end function

procedure TRAVERSETREE(tree)
    ▷ Process leaves (i.e. propositions) from left to right
    for leaf in tree.leaves do
        ▷ Check transformation rules in fixed order
        for rule in TRANSFORM_RULES do
            if match then
                ▷ (a) Sentence splitting
                simplified_propositions ← decompose leaf into a
                set of simplified propositions
                new_leaves ← convert simplified_propositions
                into leaf nodes
                ▷ (b) Constituency Type Classification
                new_node ← create a new parent node for new_leaves
                new_node.labels ← link each leaf in new_leaves to
                new_node and label each edge with the leaf's constituency
                type c ∈ CL
                ▷ (c) Rhetorical Relation Identification
                cue_phrase ← extract cue phrase from leaf.parse_tree
                new_node.rel ∈ REL ← match cue_phrase against a
                predefined set of rhetorical cue words
                ▷ Update Tree
                tree.replace(leaf, new_node)
                ▷ Recursion
                TRAVERSETREE(tree)
            end if
        end for
    end for
    return tree
end procedure
```

[3] Details on the underlying linguistic principles, supporting the systematicity and universality of the developed transformation patterns, can be found in Niklaus (2022), p. 92–97.

[4] An example of a transformation rule is provided in Table 5 in Section A. For reproducibility purposes, the full set of patterns is presented in Niklaus (2022), p. 111–141.

[5] A simple sentence comprises exactly one clause.

[6] In addition, a specified set of phrasal elements were extracted. The reader may refer to Section A for more details.

The transformation algorithm of our approach (see Algorithm 1) takes a natural language sentence as input and applies the transformation patterns to recursively transform it into a semantic hierarchy of minimal propositions, represented as an LPT.

Initialization In the initialization step (see lines 1-8 of Algorithm 1), the linked proposition tree LPT is instantiated with the source sentence. It is represented as a single leaf node that has an unlabeled edge to the root node.

Tree Traversal Next, the LPT is recursively traversed, splitting up the input in a top-down approach (9-37). Starting from the root node, the leaves are processed in depth-first order. For every leaf (11), we check if its phrasal parse tree matches one of the transformation patterns (13). The rules are applied in a fixed order that was empirically determined. The first pattern that matches the proposition's parse tree is executed (14). For instance, the first rule that matches the source sentence from Fig. 1 is the pattern shown in Table 5.

(a) Sentence Splitting In a first step, the current proposition is decomposed into a set of shorter utterances that present a more regular structure (16-17). This is achieved through disembedding clausal or phrasal components and converting them into stand-alone sentences. Accordingly, the transformation rule encodes both the split point and the rephrasing procedure for reconstructing grammatically sound sentences.[7] Each split will result in two sentences with a simpler syntax. They are represented as leaf nodes in the LPT (18-19) (see subtask (a) in Figure 3). To establish a semantic hierarchy between the split spans, two further subtasks are carried out, as described below.

(b) Constituency Type Classification To set up a contextual hierarchy between the split sentences, the transformation rule determines the constituency type $c \in CL$ of the leaf nodes that were created in the previous step (21-24). To differentiate between *core* sentences that contain the key message of the input and *contextual* sentences that provide additional information about it, the transformation pattern encodes a simple syntax-based method. Based on the assumption that subordinations commonly express background information, simplified propositions resulting from subordinate clausal or phrasal elements are classified as context sentences, while those emerging from their superordinate counterparts are labelled as core sentences. Coordinations, too, are flagged as core sentences, as they are of equal status and typically depict the main information of the input (see subtask b in Figure 3).[8]

(c) Rhetorical Relation Identification To preserve the semantic relationship between the simplified propositions, we classify the rhetorical relation $rel \in REL$ that holds between them. For this purpose, we utilize a predefined list of rhetorical cue words adapted from the work of Taboada and Das (2013).[9] To infer the type of rhetorical relation, the transformation pattern first extracts the cue phrase of the given sentence (26). It is then used as a lexical feature for classifying the semantic relationship that connects the split propositions (27-28). For example, the rule in Table 5 specifies that the phrase *"although"* is the cue word in the source sentence of Figure 1, which is mapped to a "Contrast" relationship according to the findings in Taboada and Das (2013) (see subtask c in Figure 3).

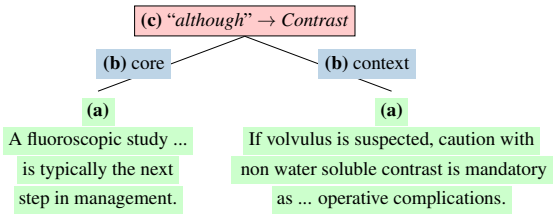

Figure 3: Semantic hierarchy after the first transformation pass. **(Subtask a)** The source sentence is split up and rephrased into a set of syntactically simplified sentences. **(Subtask b)** Then, the split sentences are connected with information about their constituency type to establish a contextual hierarchy between them. **(Subtask c)** Finally, by identifying and classifying the rhetorical relation that holds between the simplified sentences, their semantic relationship is preserved.

Recursion Next, the LPT is updated by replacing the leaf node that was processed in this run

[7] Table 4 in Section A provides an overview of the linguistic constructs that are tackled by our approach. Note that this subtask is presented in detail in Niklaus et al. (2019b). Therefore, we focus on subtasks b and c in this work.

[8] This approach relates to the concept of nuclearity in RST. In RST, each text span is specified as either a nucleus or a satellite. The *nucleus* span embodies the central piece of information and is comparable to what we denote a core sentence, whereas the role of the *satellite* is to further specify the nucleus, corresponding to a context sentence in our case.

[9] The full list of cue phrases that serve as lexical features for the identification of rhetorical relations in our approach, as well as the corresponding relations to which they are mapped, is provided in Section B.

with the newly generated subtree (30). It is composed of the simplified propositions, their semantic relationship $rel \in REL$ and constituency labels $c \in CL$. Figure 3 depicts the result of the first transformation pass on the example sentence from Figure 1. The resulting leaf nodes are then recursively simplified in a top-down fashion (32).

Termination The algorithm terminates when no more rule matches the set of simplified propositions $prop \in PROP$ in the leaf nodes. It outputs the source sentence's LPT (36), representing its semantic hierarchy of minimal semantic units. In that way, the input is transformed into a set of hierarchically ordered and semantically interconnected sentences that present a simplified syntax. Figure 1 shows the final LPT of our example sentence.

3 Evaluation

3.1 Experimental Setup

3.1.1 Automatic Metrics

We evaluate the constituency type classification and rhetorical relation identification steps by mapping the simplified sentences that were generated in the sentence splitting subtask to the Elementary Discourse Units (EDUs) of the RST-DT corpus (`https://t1p.de/n6t9`). This dataset is a collection of 385 Wall Street Journal articles annotated with rhetorical relations based on the RST framework (Mann and Thompson, 1988). For matching simplified sentences generated by our TS approach to the annotations of the RST-DT corpus, we compare each split sentence to all the EDUs of the corresponding input sentence. For each pair, we search for the longest contiguous matching subsequence. Next, based on the size of the matched sequences, a similarity score between the two input strings is calculated. Each pair whose similarity score surpasses an empirically determined threshold of 0.65 is considered a match.

Constituency Type Classification To determine whether the hierarchical relationship that is assigned by our TS framework between a pair of simplified sentences is correct, we check if the hierarchy of its contextual layers corresponds to the nuclearity of the aligned text fragments of the RST-DT. For this purpose, we make use of the nuclearity status encoded in the annotations of this dataset. In addition, we compare the performance of our TS approach with that of a set of widely used sentence-level discourse parsers on this task.

Rhetorical Relation Identification To assess the performance of the rhetorical relation identification step, we determine the distribution of the relation types allocated by our TS approach when operating on the 7,284 input sentences of the RST-DT and compare it to the distribution of the manually annotated rhetorical relations of this corpus. Moreover, we examine for each matching sentence pair whether the rhetorical relation assigned by our TS framework equates the relation that connects the corresponding EDUs in the RST-DT dataset. For this purpose, we apply the more coarse-grained classification scheme from Taboada and Das (2013), who group the full set of 78 rhetorical relations that are used in the RST-DT corpus into 19 classes of relations that share rhetorical meaning. Finally, we analyze the performance of our framework on the relation labeling task in comparison to a number of discourse parser baselines.

3.1.2 Manual Analysis

To get a deeper insight into the accuracy of the semantic hierarchy established between the split components, the automatic evaluation described above is complemented by a manual analysis. Three human judges independently of each other assessed each decomposed sentence according to the following four criteria: (i) **Limitation to core information**: Is the simplified output limited to core information of the input sentence? *(yes - no - malformed)*; (ii) **Soundness of the contextual proposition**: Does the simplified sentence express a meaningful context fact? *(yes - no)*; (iii) **Correctness of the context allocation**: Is the contextual sentence assigned to the parent sentence to which it refers? *(yes - no)*; and (iv) **Properness of the identified semantic relationship**: Is the contextual sentence linked to its parent sentence via the correct semantic relation? *(yes - no - unspecified)*. The first three categories of our analysis address the correctness of the constituency type classification task, while the latter targets the rhetorical relation identification step. The annotation task was carried out on a random sample of 100 sentences from the OIE2016 Open IE benchmark (Stanovsky and Dagan, 2016).

3.2 Results

3.2.1 Automatic Metrics

Using the matching function described in Section 3.1.1, we obtained 1,827 matched sentence pairs, i.e. 11.74% of the pairs of simplified sentences were successfully mapped to a counterpart

of EDUs from the RST-DT. The relatively low number of matches can be attributed to the fact that the text spans we compare have very different features.[10] As we are primarily interested in determining whether the constituency and relation labels that are assigned by our approach are correct, we will focus on precision in the following.[11]

Constituency Type Classification In 88.88% of the matched sentence pairs, the hierarchical relationship that is allocated between a pair of simplified sentences by our reference TS implementation DISSIM corresponds to the nuclearity status of the aligned EDUs from RST-DT, i.e. in case of a nucleus-nucleus relationship in RST-DT, both output sentences from DISSIM are assigned to the same context layer, while in case of a nucleus-satellite relationship the sentence mapped to the nucleus EDU is allocated to the context layer *cl*, whereas the sentence mapped to the satellite span is assigned to the subordinate context layer *cl*+1. The majority of the cases where our TS approach assigns a hierarchical relationship that differs from the nuclearity in the RST-DT corpus can be attributed to relative clauses.

	nuclearity	relation
DPLP (Ji and Eisenstein, 2014)	71.1	61.8
Feng and Hirst (2014)	71.0	58.2
2-Stage Parser (Wang et al., 2017)	72.4	59.7
Lin et al. (2019)	**91.3**	**81.7**
SPADE (Soricut and Marcu, 2003)	56.1	44.9
HILDA (Hernault et al., 2010)	59.7	48.2
PAR-S (Joty et al., 2015)	75.2	66.1
Lin et al. (2019)	(86.4)*	(77.5)*
DISSIM	**88.9**	**69.5**

Table 1: Precision of DISSIM and the discourse parser baselines, as reported by their authors. (*) In case of automatic discourse segmentation, for Lin et al. (2019) the F_1-score is available only.

Table 1 displays the precision that the discourse parser baselines achieve on the 991 sentences of the RST-DT test set in distinguishing between nucleus and satellite spans *("nuclearity")*. For the approaches in the upper part of the table, the authors report the systems' performance when using gold EDU segmentation, while for those in the lower part the performance is indicated based on automatic segmentation, i.e. when they are fed the output of their respective discourse segmenter. Since our framework makes use of the simplified sentences that were generated in the previous step when setting up the semantic hierarchy, it is better comparable to the latter group. The figures show that in this case our approach outperforms all other systems in the constituency type classification task by a large margin of 13.7% at a minimum.[12]

Rhetorical Relation Identification Table 2 displays the frequency distribution of the 19 classes of rhetorical relations that were specified in Taboada and Das (2013). The ten most frequently occurring classes make up for 89.45% of the relations that are present in the dataset. We decided to limit ourselves to these classes in the evaluation of the rhetorical relation identification step, with two exceptions. First, we did not take into account the "Topic-change" and "Same-unit" classes. Second, we merged the two highly related classes of "Cause" and "Explanation" into a single category.

RHET. RELATION	COUNT	PERCENT.	PRECISION
Elaboration	7,675	25.65%	0.5550
Joint	7,116	23.78%	0.6673
Attribution	2,984	9.97%	0.9601
Same-unit	2,788	9.32%	—
Contrast	1,522	5.09%	0.7421
Topic-change	1,315	4.39%	—
Explanation	966	3.21%	0.7037
Cause	754	2.52%	
Temporal	964	3.22%	0.7895
Background	897	2.30%	0.4459
			avg.: 0.6948

Evaluation (2.0%), Enablement (1.8%), Comparison (1.5%), Textual organization (1.2%), Condition (1.1%), Topic-comment (0.9%), Manner-means (0.7%), Summary (0.7%), Span (0.0%)

Table 2: Frequency distribution of the 19 classes of relations from Taboada and Das (2013) and the precision of DISSIM's rhetorical relation identification step.

The right column in Table 2 displays the precision of our TS approach for each class of rhetorical relation when run over the sentences from RST-DT. The "Attribution" relation reaches by far the highest precision. The remaining relations, too, show decent scores, with a precision of around 70%. The only exception is "Background". The difficulty with this type of relationship is that it signifies a very broad category that is not signalled by discourse markers and therefore hard to detect by our approach (Taboada and Das, 2013). With an average precision of 69.5% in the relation labeling task (see Table 1), our framework again surpasses all

[10] For details, see Section C.

[11] The fraction of labels that are successfully retrieved (i.e. recall) is of minor importance in our setting. In addition, this score might be biased, since a large proportion of EDUs from RST-DT is not mapped to a counterpart of simplified propositions in our experiments. Therefore, we refrain from reporting recall scores.

[12] A very recent approach to intra-sentential sentence parsing was proposed in Lin et al. (2019), achieving an F_1-score of 86.4%. However, the authors do not report its precision.

the discourse parser baselines under consideration when using automatic discourse segmentation.[13]

When comparing the distribution of the rhetorical relations that were identified by our TS approach on the source sentences from the RST-DT (see Figure 4) to that of the manually annotated gold relations displayed in Table 2, it turns out that there is a very high similarity between the two of them. However, it must be noted that in about 20% of the cases, our TS approach is not able to identify a rhetorical relation between a pair of split sentences *("Unknown")*. For the most part, this can be attributed to sentence pairs whose relation is not explicitly stated in the underlying source sentence. As our approach is based on cue phrases, searching for discourse markers that explicitly signal rhetorical relations, it has difficulties in identifying relations that can merely be implied.

3.2.2 Manual Analysis

The results of the human evaluation are displayed in Table 3. The inter-annotator agreement was calculated using Fleiss' κ (Fleiss, 1971). The figures indicate fair to substantial agreement between the three annotators, suggesting that the evaluation scores present a reliable result.

Category	Yes	No	Malf.	Unspec.	κ
Limitation to core information	**68.2%**	20.0%	11.9%	—	0.39
Soundness of the contextual proposition	**83.1%**	16.9%	—	—	0.51
Correctness of the context allocation	**93.2%**	6.8%	—	—	0.41
Properness of the semantic relationship	**69.8%**	7.0%	—	23.2%	0.69

Table 3: Results of the manual analysis.

In more than two out of three cases, the annotators marked the propositions that were classified as core sentences by our TS approach as correct, thus approving that they have a meaningful interpretation and that their content is truly restricted to core information of the underlying source sentence. Only about 12% of the simplified sentences are malformed according to our annotations. The remaining fifth of output core sentences was judged as being misclassified, i.e. they rather contribute less relevant background data than key information of the input. Regarding the soundness of the context propositions, only about 17% of the output propositions that were classified as context sentences were labelled as being inaccurate, while as many as 83% present proper contextual propositions, expressing a meaningful context fact that is asserted by the input and can be properly interpreted. Furthermore, 93% of the context sentences are assigned to their respective parent sentence, whereas only 6% of them are misallocated, according to the annotators' labels. Finally, our evaluation revealed that our TS approach shows a decent performance for the rhetorical relation identification step, too. More than two-thirds of the sentence pairs are classified with the correct rhetorical relation, according to our manual analysis. Only 7% of them are assigned an improper relation. However, in nearly a quarter of the cases, our TS approach was not able to identify a semantic relationship between the given pair of sentences. This can be explained by the fact that for this subtask, our framework follows a rather simplistic approach that is primarily based on cue phrases. Therefore, it fails to identify a semantic relationship whenever none of the specified keywords appears in the underlying input sentence. As a result, our approach provides very precise results. Covering only a small subset of rhetorical relations it lacks in completeness, though.

4 A Lightweight Semantic Representation for Open IE

The fine-grained representation of complex sentences in the form of hierarchically ordered and semantically interconnected propositions may serve as an intermediate representation for downstream tasks. An application area that may benefit greatly from our approach as a preprocessing step is the task of Open IE (Banko et al., 2007). We thus assessed the merits of our proposed discourse-aware TS approach in supporting the extraction of relational tuples from complex assertions in downstream Open IE applications, demonstrating that the semantic hierarchy of minimal propositions benefits them in two dimensions:

(a) The normalized subject-predicate-object syntax of the simplified sentences reduces the complexity of the relation extraction step, resulting in a simplistic canonical predicate-argument structure of the output.

(b) By capturing intra-sentential rhetorical structures and hierarchical relationships between the propositions, it allows for the enrichment

[13] with the exception of Lin et al. (2019)'s parser, for which only the F_1-score is reported by the authors, though. Hence, it is not directly comparable to the other approaches whose performance is analyzed based on their precision.

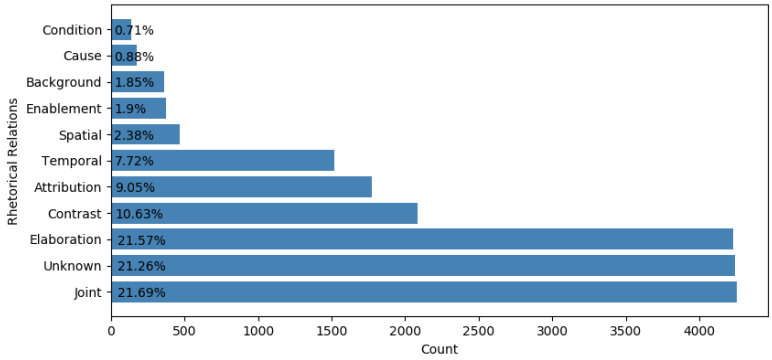

Figure 4: Distribution of the rhetorical relations identified by our TS approach on the RST-DT.

of the extracted relational tuples with additional meta information that supports their interpretability.

In that way, the shallow semantic representation of state-of-the-art Open IE systems is transformed into a canonical context-preserving representation of relational tuples.

```
(1) she;  was confirmed on;          August 6, 2009
(2) He;   nominated Sonia Sotomayor on; May 26
(3) He;   nominated Sonia Sotomayor; 2009
(4) He;   nominated 2009 on;         May 26
(5) Sonia Sotomayor;  be nominated 2009 on;  May 26
(6) He;   nominated 2009;            Sonia Sotomayor
(7) 2009; be nominated Sonia Sotomayor on;May 26
```

Figure 5: OLLIE's extractions (Mausam et al., 2012).

Extraction of Canonical Predicate-Argument Structures Representing normalized monopredicative units, the simplified sentences reduce the complexity of the relation extraction step and inherently support the extraction of canonical predicate-argument structures. Thus, a standardized output scheme is created, where each simplified sentence results in a *normalized (mostly) binary predicate-argument structure*, in which *both the predicate and the argument slots are reduced to their essential components*. In that way, the generation of overly specific predicate and argument phrases, as well as (quasi-)redundant extractions is prevented, as illustrated by the examples in Figures 5 to 7.[14]

Enrichment of the Output with Semantic Information Moreover, our TS approach enables existing Open IE systems to enrich their output with semantic information. The semantic hierarchy can be leveraged to incorporate important contextual information of the extracted relational tuples, thus extending the shallow semantic representation (in the form of isolated predicate-argument structures) of state-of-the-art Open IE systems.[15] First, the semantic hierarchy supports the specification of a hierarchical order between the extracted relational tuples, as it enables to distinguish between *different levels of context* - the lower the allotted layer, the more relevant is the information contained in it. Second, the semantic hierarchy generated by our discourse-aware TS approach can be used to enrich the output of Open IE approaches with additional meta information in terms of rhetorical relations, allowing for the representation of *semantically typed relational tuples*. Thus, the extracted relations are put into a *logical structure that preserves the semantic context of the extractions*, resulting in an output that is more informative and coherent, and thus easier to interpret. See Figure 8 for an example.

Hence, the semantic hierarchy of minimal propositions generated by our discourse-aware TS approach can be leveraged to transform the shallow semantic representation of existing Open IE systems into a novel canonical context-preserving representation of relational tuples. The proposed representation allows for a simplistic unified representation of predicate-argument structures that can easily be enriched with contextual information in terms

[14] In addition, we demonstrated that the precision and recall of state-of-the-art Open IE systems is improved by up to 346% and 52%, respectively, when taking advantage of the split propositions instead of dealing with the complex source sentences (Niklaus et al., 2019b).

[15] Previous work in the area of Open IE has mainly focused on the extraction of isolated relational tuples, ignoring the cohesive nature of texts where important contextual information is spread across clauses or sentences. Consequently, state-of-the-art Open IE approaches are prone to generating a loose arrangement of tuples that lack the expressiveness needed to infer the true meaning of complex assertions.

```
(1) he;     nominated;        Sonia Sotomayor on May 26 2009 to replace David Souter
(2) she;    was confirmed;    on August 6 2009 becoming the first Supreme Court Justice of Hispanic descent
(3) she;    was confirmed;    becoming the first Supreme Court Justice of Hispanic descent
```

Figure 6: Relations extracted by ClausIE (Del Corro and Gemulla, 2013) from the sentence: *"He nominated Sonia Sotomayor on May 26, 2009 to replace David Souter; she was confirmed on August 6, 2009, becoming the first Supreme Court Justice of Hispanic descent."*.

```
(1) #1  0  he;    nominated;    Sonia Sotomayor
(1a)           PURPOSE     to replace David Souter.
(1b)           TEMPORAL    on May 26, 2009.
(2) #2  0  she;    was confirmed;
(2a)           TEMPORAL    on August 6, 2009.
(3) #3  0  she;    was becoming;    the first
           Supreme Court Justice of Hispanic descent
```

Figure 7: Relations extracted by OLLIE and ClausIE when using our TS approach as a preprocessing step.

```
(1) A fluoroscopic study; known; as an upper gastro-
                                    intestinal series
(2) caution with non water soluble contrast; is; man-
                        datory as the usage of barium
(3) as the usage; of barium can impede;   surgical
                                   revision and lead
(4)  ;   to increased;  post operative complications


(5)  #1  0  A fluoroscopic study; is; typically, the
                          next step in management
(5a)              ELABORATION   #2
(5b)              CONTRAST      #3
(6)  #2  1  This; fluoroscopic study is known;    as
                  an upper gastrointestinal series
(7)  #3  0  Caution with non water soluble; is; man-
                                                datory
(7a)              CONTRAST      #1
(7b)              CONDITION     #6
(7c)              BACKGROUND    #4
(7d)              BACKGROUND    #5
(8)  #4  1  The usage of barium; can impede;surgical
                                                revision
(8a)              LIST          #5
(9)  #5  1  The usage of barium;  can lead;   to in-
                creased post operative complications
(9a)              LIST          #4
(10) #6  1  Volvulus;    is suspected;
```

Figure 8: Comparison of the tuples extracted by RnnOIE (Stanovsky et al., 2018) *with* (5 - 10) and *without* (1 - 4) using our TS approach as a preprocessing step.

of intra-sentential rhetorical structures and hierarchical relationships between the extracted tuples, resulting in a set of interrelated semantically typed tuples that preserve the coherence of the output.

5 Related Work

Discourse-level TS The vast majority of structural TS approaches do not take into account discourse-level aspects. However, two notable exceptions have to be mentioned. Siddharthan (2006) was the first to use discourse-aware cues in the simplification process. As opposed to our approach, though, where a semantic relationship is established for each simplified output sentence, only a comparatively low number of sentences is linked by such cue words. Another approach that operates on the level of discourse was proposed by Štajner and Glavaš (2017). It performs a semantically motivated content reduction by maintaining only those parts of a sentence that belong to factual event mentions. Our approach, on the contrary, aims to preserve all the information contained in the source.

Discourse Parsing The challenge of uncovering coherence structures in texts is pursued in the field of Discourse Parsing. It aims to identify discourse relations that hold between textual units in a document (Marcu, 1997). A well-established theory of text structure used in this area is RST. Here, textual coherence is explained by the existence of rhetorical relations that hold between adjacent text spans in a hierarchical structure. Approaches to detect rhetorical structure arrangements in texts range from early rule-based approaches (Marcu, 2000) to supervised data-driven models that were trained on annotated corpora such as the RST-DT (Feng and Hirst, 2014; Li et al., 2014; Lin et al., 2019).[16]

6 Conclusion

We presented a context-preserving TS approach that transforms structurally complex sentences into a hierarchical representation in the form of core sentences and accompanying contexts that are semantically linked by rhetorical relations. In our experiments, we mapped the simplified sentences from our reference implementation DISSIM to the EDUs from RST-DT and showed that we obtain a very high precision of 89% for the constituency type classification and a decent score of 69% on average for the rhetorical relation identification. In the future, we plan to improve the latter step by extending our approach to also capture implicit relationships between the decomposed sentences.

References

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings*

[16] Section D elaborates on why it is not possible to simply use an RST parser for establishing the semantic hierarchy between the decomposed spans.

of the 20th International Joint Conference on Artifical Intelligence, pages 2670–2676, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54:56.

Matthias Cetto, Christina Niklaus, André Freitas, and Siegfried Handschuh. 2018. Graphene: Semantically-linked propositions in open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2300–2311, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: Clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 355–366, New York, NY, USA. Association for Computing Machinery.

Richard Evans and Constantin Orăsan. 2019. Identifying signs of syntactic complexity for rule-based sentence simplification. *Natural Language Engineering*, 25(1):69–119.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521, Baltimore, Maryland. Association for Computational Linguistics.

Daniel Ferrés, Montserrat Marimon, Horacio Saggion, and Ahmed AbuRa'ed. 2016. YATS: yet another text simplifier. In *Natural Language Processing and Information Systems - 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings*, volume 9612 of *Lecture Notes in Computer Science*, pages 335–342. Springer.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Michael Heilman and Noah A. Smith. 2010. Extracting simplified statements for factual question generation. In *Proceedings of the QG2010: The Third Workshop on Question Generation*, pages 11–20.

Hugo Hernault, Helmut Prendinger, Mitsuru Ishizuka, et al. 2010. Hilda: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3).

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069, Doha, Qatar. Association for Computational Linguistics.

Xiang Lin, Shafiq Joty, Prathyusha Jwalapuram, and M Saiful Bari. 2019. A unified linear-time framework for sentence-level discourse parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4200, Florence, Italy. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Daniel Marcu. 1997. The rhetorical parsing of unrestricted natural language texts. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 96–103, Madrid, Spain. Association for Computational Linguistics.

Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts: a surface-based approach. *Computational Linguistics*, 26(3):395–448.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534, Jeju Island, Korea. Association for Computational Linguistics.

Ruslan Mitkov and Horacio Saggion. 2018. Text simplification.

Christina Niklaus. 2022. *From Complex Sentences to a Formal Semantic Representation using Syntactic Text Simplification and Open Information Extraction*. Springer Fachmedien Wiesbaden, Wiesbaden.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2019a. DisSim: A discourse-aware syntactic text simplification framework for English and German. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 504–507, Tokyo, Japan. Association for Computational Linguistics.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2019b. Transforming complex sentences into a semantic hierarchy. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3415–3427, Florence, Italy. Association for Computational Linguistics.

Swarnadeep Saha and Mausam. 2018. Open information extraction from conjunctive sentences. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2288–2299, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Advaith Siddharthan. 2002. An architecture for a text simplification system. In *Language Engineering Conference, 2002. Proceedings*, pages 64–71. IEEE.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Advaith Siddharthan and Angrosh Mandya. 2014. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 722–731, Gothenburg, Sweden. Association for Computational Linguistics.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 228–235.

Sanja Štajner and Goran Glavaš. 2017. Leveraging event-based semantics for automated text simplification. *Expert systems with applications*, 82:383–395.

Sanja Štajner and Maja Popović. 2018. Improving machine translation of English relative clauses with automatic text simplification. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 39–48, Tilburg, the Netherlands. Association for Computational Linguistics.

Gabriel Stanovsky and Ido Dagan. 2016. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2300–2305, Austin, Texas. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Maite Taboada and Debopam Das. 2013. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *D&D*, 4(2):249–281.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.

A Transformation Patterns

One of the fundamental objectives of our discourse-aware TS approach is to decompose complex assertions into a set of self-contained minimal propositions. Table 4 provides an overview of the linguistic constructs that are addressed by our framework in order to achieve this goal, including the number of transformation patterns that were specified for each syntactic phenomenon. Table 5 shows an example of a transformation rule.

B Mapping of Cue Phrases to Rhetorical Relations

Table 6 lists the full set of cue phrases that serve as lexical features for the identification of rhetorical relations when establishing the semantic hierarchy between a pair of split sentences. It further shows to which rhetorical relation each of them is mapped.

In addition, *Spatial* and *Temporal* relationships are identified on the basis of named entities, while *Attribution* relations are detected using a predefined list of verbs of reported speech and cognition (Carlson and Marcu, 2001).

Furthermore, in some cases, the type of relationship that is set between two decomposed spans is selected based on syntactic information. This applies to the following rhetorical relations:

- *Purpose* (in case of adverbial clauses of purpose, lexicalized on the preposition *"to"*),
- *Elaboration* (in case of appositives, adjectival/adverbial phrases, participial phrases without an adverbial connector and relative clauses that are *not* introduced by the relative pronoun *"where"*),
- *Spatial* (in case of relative clauses commencing with the relative pronoun *"where"*) and
- *Temporal* (in case of lead noun phrases).

	CLAUSAL/PHRASAL TYPE	HIERARCHY	# RULES
	Clausal disembedding		
1	Coordinate clauses	coordinate	1
2	Adverbial clauses	subordinate	6
3a	Relative clauses (non-restrictive)	subordinate	5
3b	Relative clauses (restrictive)	subordinate	4
4	Reported speech	subordinate	4
	Phrasal disembedding		
5	Coordinate verb phrases	coordinate	1
6	Coordinate noun phrases	coordinate	2
6	Participial phrases	subordinate	4
8a	Appositions (non-restrictive)	subordinate	1
8b	Appositions (restrictive)	subordinate	1
9	Prepositional phrases	subordinate	3
10	Adjectival and adverbial phrases	subordinate	2
11	Lead NPs	subordinate	1
	Total		35

Table 4: Linguistic constructs addressed by our discourse-aware TS approach DISSIM.

ROOT <<: (S < (NP $.. (VP < +(VP) (**SBAR** <, (*IN* $+ (S < (NP $.. VP))))))))

Table 5: Example of a transformation pattern (for decomposing adverbial clauses). They are specified in terms of Tregex patterns (Levy and Andrew, 2006). A boxed pattern represents the part of a sentence that is extracted from the input and transformed into a new stand-alone sentence. A pattern in bold is deleted from the source. The underlined part is labelled as a context sentence, while the remaining part represents core information. The italic pattern is used as a cue phrase for identifying the rhetorical relation that holds between the decomposed spans.

RHET. RELATION	CUE PHRASES
Contrast	although, but, but now, despite, even though, even when, except when, however, instead, rather, still, though, thus, until recently, while, yet
List	and, in addition, in addition to, moreover
Disjunction	or
Cause	largely because, because, since
Result	as a result, as a result of
Temporal	after, and after, next, then, before, previously
Background	as, now, once, when, with, without
Condition	if, in case, unless, until
Elaboration	more provocatively, even before, for example, further, recently, since, since now, so, so far, where, whereby, whether
Explanation	simply because, because of, indeed, so, so that

Table 6: Mapping of cue phrases to rhetorical relations.

C Evaluation

While the goal of our TS approach is to generate well-formed syntactically simplified sentences, the EDUs in the RST-DT are copied verbatim from the source, resulting in an output of varied length that is usually not grammatically sound. Moreover, in many cases, the EDUs mix multiple semantic units, whereas our approach aims to split the input into atomic components, with each of them expressing a coherent and indivisible proposition.

D Discourse Parsing

The syntactic analysis we propose for establishing the semantic hierarchy between the decomposed spans is bound to the RST discourse markers. However, it is not possible to simply use an RST parser for this task. As illustrated in Figure 9, such a parser does not return grammatically sound sentences. Instead, it segments the input into basic textual units, so-called elementary discourse units (EDUs), which are copied verbatim from the source. In order to reconstruct proper sentences, rephrasing is required. For this purpose, amongst others, referring expressions have to be identified, and phrases have to be rearranged and inflected. Moreover, the textual units resulting from the segmentation process are too coarse-grained for our purpose, since

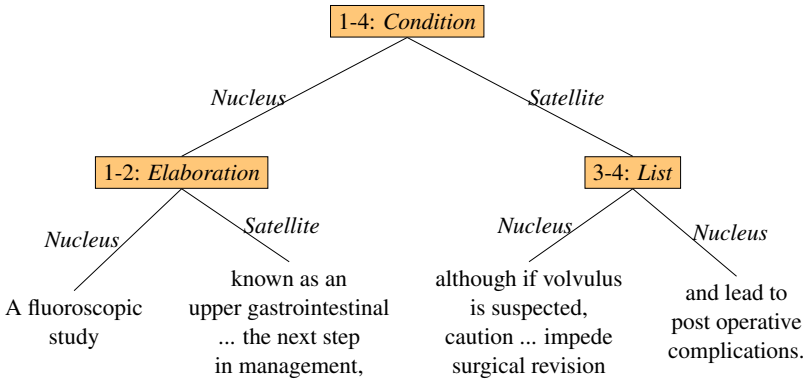

Figure 9: Rhetorical structure tree of our example sentence, generated using the RST parser proposed in Ji and Eisenstein (2014). The leaves correspond to *EDUs*, while each node is characterized by its *nuclearity* and a *rhetorical relation* between adjacent text spans.

RST parsers mostly operate on clausal level. The goal of our approach, though, is to split the input into minimal semantic units, which requires to go down to the phrasal level in order to produce a much more fine-grained output in the form of minimal propositions.