PROPSEGMENT: A Large-Scale Corpus for Proposition-Level Segmentation and Entailment Recognition

Sihao Chen*[1,2] **Senaka Buthpitiya**[1] **Alex Fabrikant**[1] **Dan Roth**[2] **Tal Schuster**[1]

[1]Google Research [2]University of Pennsylvania

{senaka,fabrikant,talschuster}@google.com, {sihaoc,danroth}@cis.upenn.edu

Abstract

The widely studied task of Natural Language Inference (NLI) requires a system to recognize whether one piece of text is textually entailed by another, i.e. whether the *entirety* of its meaning can be inferred from the other. In current NLI datasets and models, textual entailment relations are typically defined on the sentence- or paragraph-level. However, even a simple sentence often contains multiple *propositions*, i.e. distinct units of *meaning* conveyed by the sentence. As these propositions can carry different truth values in the context of a given premise, we argue for the need to recognize the textual entailment relation of each proposition in a sentence individually.

We propose PROPSEGMENT, a corpus of over 45K propositions annotated by expert human raters. Our dataset structure aligns with the tasks of (1) segmenting sentences within a document to the set of propositions, and (2) classifying the entailment relation of each proposition with respect to a different yet topically-aligned document, i.e. documents describing the same event or entity. We establish strong baselines for the segmentation and entailment tasks. Through case studies on summary hallucination detection and document-level NLI, we demonstrate that our conceptual framework is potentially useful for understanding and explaining the compositionality of NLI labels.

1 Introduction

Natural Language Inference (NLI), or Recognizing Textual Entailment (RTE), is the task of determining whether the meaning of one text expression can be inferred from another (Dagan and Glickman, 2004). Given two pieces of text (P, H), we say the premise P *entails* the hypothesis H if the *entirety* of H's meaning can be most likely inferred true after a human reads P. If some units of meaning in H are contradicted by, or cannot be determined from P, we describe the relation between the two as *contradiction* or *neutral* (de Marneffe et al., 2008) respectively. This fundamentally challenging natural language understanding task provides a general interface for semantic inference and comparison across different sources of textual information.

Premise Document	
Andrew Warhola, known as Andy Warhol, is an American artist born August 6, 1928 in Pittsburgh, Pennsylvania and died February 22, 1987 in New York. He is one of the main representatives of pop art. Warhol is known the world over for his work as a painter, music producer, author, avant-garde films... (7 more sentences omitted)	
Hypothesis Sentence (*from another document of the same topic*)	
... The Andy Warhol Museum in his hometown, Pittsburgh, Pennsylvania, contains an extensive permanent collection of art. ...	
Propositions	**Entailment Label**
The Andy Warhol Museum in his hometown, Pittsburgh, Pennsylvania, contains an extensive permanent collection of art.	*Neutral*
The Andy Warhol Museum in his hometown, Pittsburgh, Pennsylvania, contains an extensive permanent collection of art.	*Entailment*
The Andy Warhol Museum in his hometown, Pittsburgh, Pennsylvania, contains an extensive permanent collection of art.	*Neutral*

Table 1: An example instance from the PROPSEGMENT dataset with propositions (marked as token subsets highlighted in blue) and their entailment labels.

In reality, most naturally occurring text expressions are composed of a variable number of *propositions*, i.e. distinct units of meaning conveyed by the piece of text. Consider the sentence shown in Table 1: "*The Andy Warhol Museum in his hometown, Pittsburgh, Pennsylvania, contains an extensive permanent collection of art.*" Despite the sentence being relatively compact, it still contains (at least) three propositions, as listed in Table 1. While the entire hypothesis would be classified as *neutral* or *not-entailed* to the premise, one of its proposi-

* Work done as an intern at Google

Findings of the Association for Computational Linguistics: ACL 2023, pages 8874–8893
July 9-14, 2023

tions "*Andy Warhol's hometown is in Pittsburgh, Pennsylvania*" is in fact entailed by the premise, while the premise provides no support for the other two propositions. This phenomenon, namely *partial entailment* (Levy et al., 2013), is a blind spot for existing sentence- or paragraph-level NLI formulations. When a hypothesis is *compositional*, NLI labels coarsely defined on the sentence/paragraph-level cannot express the difference between partial entailment from the non-entailment cases.

This work argues for the need to study and model textual entailment relations on the level of *propositions*. As NLI tasks and applications typically involve different genre of text with variable length and number of propositions (Yin et al., 2021), decomposing textual entailment relation to the propositional level provides a more fine-grained yet accurate description of textual entailment relation between two arbitrary text expressions. Modeling *propositional textual entailment* provides a more unified inference format across NLI tasks, and would potentially improve the generalization capabilities of NLI models, e.g. with respect to the variability in input lengths (Schuster et al., 2022).

We propose PROPSEGMENT, a multi-domain corpus with over 45K human-annotated propositions. [1] We define the tasks of proposition-level segmentation and entailment. Given a hypothesis sentence and a premise document, a system is expected to segment the hypothesis into the set of propositions, and recognize whether each proposition can be inferred from the premise.

Interestingly, we observe that existing notions of proposition adopted by Open Information Extraction (OpenIE) or Semantic Role Labeling (SRL) (Baker et al., 1998; Kingsbury and Palmer, 2002; Meyers et al., 2004) often fail to account for the complete set of propositions in a sentence, partly due to the fact that predicates and arguments in different propositions do not necessarily follow the same granularity (§2). We therefore adopt a more flexible and unified way of representing a proposition as a *subset of tokens* from the input sentence, without explicitly annotating the semantic role or predicate-argument structure within the proposition, as illustrated in Table 1. We discuss the motivation and design desiderata in §2.

We construct PROPSEGMENT by sampling clusters of topically-aligned documents, i.e. documents focusing on the same entity or event, from WIKIPEDIA (Schuster et al., 2022) and the news domains (Gu et al., 2020). We train and instruct expert annotators to identify all propositions exhaustively in a document, and label the textual entailment relation of each proposition with respect to another document in the cluster, viewed as the premise.

We discuss the modeling challenges, and establish strong baselines for the segmentation and entailment tasks. We demonstrate the utility of our dataset and models through downstream use case studies on summary hallucination detection (Maynez et al., 2020), and DocNLI (Yin et al., 2021), through which we show that recognizing and decomposing entailment relations at the proposition-level could provide fine-grained characterization and explanation for NLI-like tasks, especially with long and compositional hypotheses.

In summary, the main contributions in our paper include: (1) Motivating the need to recognize textual entailment relation on proposition level; (2) Introducing the first large-scale dataset for studying proposition-level segmentation and entailment recognition; and (3) Leveraging PROPSEGMENT to train Seq2Seq models as strong baselines for the tasks, and demonstrating their utility in document-level NLI and hallucination detection tasks.

2 Motivations & Design Challenges

Our study concerns the challenges of applying NLI/RTE task formulations and systems in *real-world* downstream applications and settings. As textual entailment describes the relation between the meanings of two text expressions, one natural type of downstream use cases for NLI systems is to identify alignments and discrepancies between the semantic content presented in different documents/sources (Kryscinski et al., 2020; Schuster et al., 2021; Chen et al., 2022).

Our study is motivated by the task of comparing the content of topically-related documents, e.g. news documents covering the same event (Gu et al., 2020), or Wikipedia pages from different languages for similar entities (Schuster et al., 2022). As existing NLI datasets typically define the textual entailment relations at the sentence or paragraph level (Bowman et al., 2015; Williams et al., 2018), NLI systems trained on such resources can only recognize whether or not the entirety of a hypothesis sentence/paragraph is entailed by a premise. However, we estimate that, in these two domains, around

[1] The dataset is available at https://github.com/google-research-datasets/propsegment

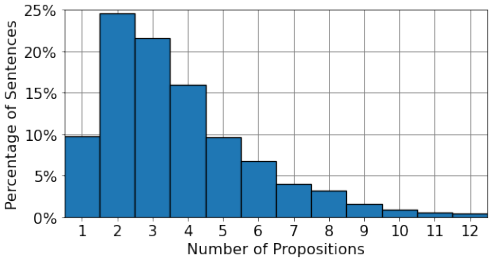

Figure 1: Distribution of proposition counts in sentences with at least one informational propositions from Wikipedia and news in PROPSEGMENT.

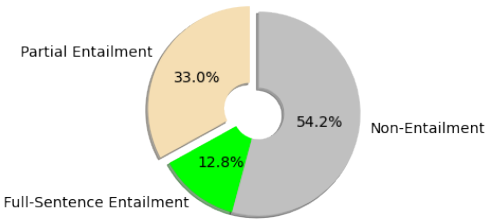

Figure 2: The percentage of sentences with *partial entailment* relation to another topically-related document from Wikipedia or news in PROPSEGMENT. Typically, NLI/RTE datasets do not distinguish partial entailment from the non-entailment categories.

90% of the sentences that convey any informational propositions contain more than one proposition (Figure 1). In the presence of multiple propositions, *partial entailment* (Levy et al., 2013) describes the phenomenon where only a subset of propositions in the hypothesis is entailed by the premise.

Partial entailment is 3× more common than full-sentence entailment. In our corpus, we observe that, given two topically related documents from news or Wikipedia, 46% of sentences in one document have at least some information supported by the other document (Figure 2). But 74% of such sentences are *partially entailed*, with only some propositions supported by the other document. In this sense, a sentence-level NLI model can only detect a quarter of sentences that have meaningful entailment relations. In applications that seek a full understanding of cross-document semantic links, there is thus 4× headroom, a significant blind spot for sentence-level NLI models.

As we observe that most natural sentences are compositional, i.e. contain more than one proposition, we argue for the need to decompose and recognize textual entailment relation at the more granular level of propositions. In other words, instead of assessing the entire hypothesis as one unit in the context of a premise, we propose to evaluate the truth value of each proposition individually, and aggregate for the truth value of the hypothesis.

Current predicate-argument based methods often fail to extract all propositions in a sentence. The linguistic notion of a proposition refers to a single, contextualized unit of meaning conveyed in a sentence. In the NLP community, propositions are usually represented by the predicate-argument structure of a sentence. For example, resources like FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005), NomBank (Meyers et al., 2004), among others, represent a proposition by a predicate (verbal, nominal, etc.), with arguments filling its thematic proto-roles. Such resources facilitate the development of SRL systems (Palmer et al., 2010) for proposition extraction, with a closed, predefined set of proto-roles. To increase the coverage of propositions extracted, OpenIE formulations (Etzioni et al., 2008; Del Corro and Gemulla, 2013; Cui et al., 2018) were proposed to forgo the limits on fixed semantic roles and account for both explicit and implicit predicates. However, we observe that OpenIE systems often fail to account for the complete set of propositions in a sentence. In many cases, e.g. the *Andy Warhol's hometown* example in Table 1, arguments of a proposition might not follow the same granularity as the ones in the sentence, e.g. *Andy Warhol* vs *Andy Warhol Museum*. Also, as OpenIE triples are still defined on direct predicate-argument relations, they often fail to produce a *decontextualized* (Choi et al., 2021) view of a proposition. For example, an OpenIE system would recognize the possessive relation "he has a hometown", but fail to resolve the references of *he* → *Andy Warhol*, and *hometown* → *Pittsburgh*.

Furthermore, Gashteovski et al. (2020) and Fatahi Bayat et al. (2022) observe that *neural* OpenIE systems tend to extract long arguments that could potentially be decomposed into more compact propositions. For textual entailment, we argue for the need to extract the complete set of propositions in their most *compact* form, due to the fact that their truth value could vary individually.

To illustrate the difference between OpenIE and our approach, we offer a list of example propositions from our proposed PROPSEGMENT dataset, and compared them to extractions from rule-based and neural OpenIE systems, in Appendix D.

3 PROPSEGMENT Dataset

We propose PROPSEGMENT, a large-scale dataset featuring clusters of topically similar news and

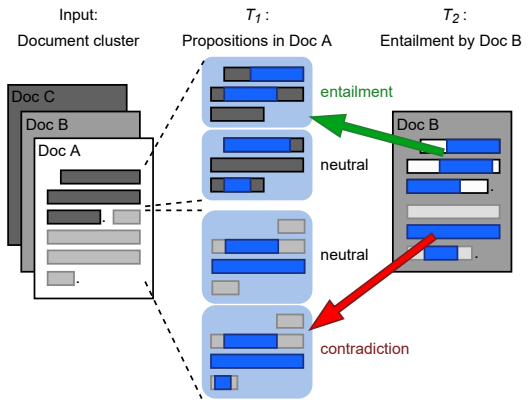

Figure 3: Given a cluster of related documents, T_1 asks for each sentence of each document to be segmented into propositions, represented as subsets of a sentence's tokens. T_2 asks to classify the entailment relation {*entails, neutral, contradicts*} of each proposition in document A w.r.t. another document B from the same cluster; Our annotations also feature a single proposition in B that best supports each *entails* or *contradicts* label.

Wikipedia documents, with human annotated propositions and entailment labels.

3.1 Task Definitions

We formulate the task of recognizing propositional textual entailment into two sub-tasks (Fig. 3). Given a hypothesis sentence and a premise document, a system is expected to (1) identify all the propositions within the hypothesis sentence, and (2) classify the textual entailment relation of each proposition with respect to the premise document.

T_1: Propositional Segmentation Given a sentence S with tokens $[t_0, t_1, ..., t_l]$ from a document D, a system is expected to identify the set of propositions $\mathcal{P} \subseteq 2^S$, where each proposition $p \in \mathcal{P}$ is represented by a unique subset of tokens in sentence S. In other words, each proposition can be represented in sequence labeling format, per the example from Table 1. Each proposition is expected (1) to correspond to a distinct fact that a reader learns directly from reading the given sentence, (2) include all tokens within the sentence that are relevant to learning this fact, and (3) to not be equivalent to a conjunction of other propositions. We opt for this format as it does not require explicit annotation of the predicate-argument structure. This allows for more expressive power for propositions with implied or implicit predicates (Stern and Dagan, 2014). Also, representing each proposition as a separate sequence could effectively account for cases with shared predicate or arguments spans, and make evaluation more readily accessible.

Since the propositions, as we demonstrated earlier, do not necessarily have a unique and identifiable predicate word in the sentence, the typical inference strategy, e.g. in SRL or OpenIE, which first extracts the set of predicates, and then identifies the arguments with respect to each predicate would not work in this case. For this reason, given an input sentence, we expect a model on the task to directly output *all* propositions. In such *one-to-set* prediction setting, the output propositions of the model are evaluated as an unordered set.

T_2: Propositional Entailment Given a hypothesis proposition p from document D_{hyp} and a whole premise document D_{prem}, a system is expected to classify whether the premise entails the proposition, i.e. if the information conveyed by the proposition would be inferred true from the premise.

3.2 Dataset Construction

We sample 250 document clusters from both the Wiki Clusters (Schuster et al., 2022) and NewSHead (Gu et al., 2020) datasets. Each cluster contains the first 10 sentences of three documents, either news articles on the same event, or Wikipedia pages in different languages (machine-translated into English) of the same entity. For each sentence, we train and instruct three human raters to annotate the set of propositions, each of which represented by a unique subset of tokens from the sentence. Conceptually, we instruct raters to include all the words that (1) pertain to the content of a proposition, and (2) are explicitly present in the sentence. For example, if there does not exist a predicate word for a proposition in the sentence, then only include the corresponding arguments. Referents present within the sentence are included in addition to pronominal and nominal references. We provide a more detailed description of our rater guidelines and how propositions are defined with respect to various linguistic phenomena in Appendix B.

Given the three sets of propositions from the three raters for a sentence, we reconcile and select one of the three raters' responses with the highest number of propositions that the other raters also annotate. Since the exact selection of tokens used to mark a proposition may vary across different raters, we allow for fuzziness when measuring the match between two propositions. Following FitzGerald et al. (2018) and Roit et al. (2020), we use Jaccard similarity, i.e. intersection over union of the two

Item	Wikipedia			News			Full Dataset		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
News Clusters	210	15	24	210	15	25	420	30	49
Documents	630	45	72	630	45	75	1260	90	147
Sentences	4990	376	532	4923	348	596	9913	724	1128
Propositions	21191	1597	2380	17015	1344	2023	38206	2941	4403
Prop.→Doc. Label #	14083	1057	4729	11369	948	4008	25452	2005	8737
Entail Label %	34.70	33.24	34.85	20.27	19.98	20.13	28.26	26.99	28.19

Table 2: Notable Statistics from the PropSegmEnt dataset.

sets of selected tokens, to measure the similarity between two propositions. We say two propositions match if their Jaccard similarity is greater or equal to a threshold $\theta = 0.8$, and align two raters' responses using unweighted bipartite matching between propositions satisfying the Jaccard threshold.

Next, for all propositions in a document, we sample one other document from the document cluster as premise, and ask three raters to label the textual entailment relation between each proposition and the premise, i.e. one of {*Entailment, Neutral, Contradiction*}. We take the majority vote from the three as the gold entailment label. Interestingly, we observe that only 0.2% of all annotated labels from the rater are "*contradictions*". We speculate that the low presence of contraditions can in part be attributed to the difficulty in establishing reference determinacy (Bowman et al., 2015) between the premise and hypothesis. We discuss more details in Appendix C. For this reason, we choose to only consider two-way label ({*Entailment, Non-Entailment*}) for the entailment task evaluation.

We create the train/dev/test splits based on clusters, so that documents in each cluster exclusively belong to only one of the splits. Overall, the dataset features 1497 documents with ∼45K propositions with entailment labels; More statistics in Table 2.

3.3 Inter-Rater Agreement

For the propositional segmentation task (T_1), as the inter-rater agreement involves set-to-set comparison between the propositions annotated by a pair of raters, we report two different metrics.

First, between each pair of raters, we use the same Jaccard similarity with $\theta = 0.8$ and find the matched set of propositions between the raters with bipartite matching for each example. We measure the coverage of the matched set by either rater with F_1 score. We observe 0.57 F_1 among all raters. As comparison, we use the same metric for model evaluation and human performance estimation, as we will discuss in § 5.1. In addition, we measure the token-level agreement on the matched set of propositions among raters with Fleiss' kappa (Fleiss, 1971), i.e. whether raters agree on whether each token should be included in a proposition or not. We observed $\kappa = 0.63$, which indicates moderate to substantial agreement among raters.

For the entailment task, (T_2), we observe Fleiss' kappa = 0.84 across three-way {*Entailment, Neutral, Contradiction*} labels.

4 Baseline Methods

4.1 Propositional Segmentation Baselines

The key challenge with the proposition extraction task lies within the one-to-set structured prediction setting. Our one-to-set prediction format is similar to QA-driven semantic parsing such as QA-SRL (He et al., 2015; Klein et al., 2022), as both involve generating a variable number of units of semantic content under no particular order between them. As in propositions, there might not necessarily be a unique and identifiable predicate word associated with each proposition, extracting predicates first (e.g. as a sequence tagging task), and later individually produce one proposition for each predicate would not be a sufficient solution in this case.

For this particular one-to-set problem setup, We introduce two classes of baseline models.

Seq2Seq: T5 (Raffel et al., 2020) When formatting a output set as a sequence, Seq2Seq models have been found to be a strong method for tasks with set outputs, as they employ chain-rules to efficiently model the joint probability of outputs (Vinyals et al., 2016). The obvious caveat for representing set outputs as sequences is that we need an ordering for the outputs. Having a consistent ordering helps seq2seq model learn to maintain the output set structure (Vinyals et al., 2016), and the best ordering scheme is often both model- and task-specific (Klein et al., 2022). In our experiments, we observe that sorting the propositions by the appearance order of the tokens in the sentence, i.e.

Task/Setting	Model	Jaccard $\theta = 0.8$			Exact Match		
		Precision	Recall	F1	Precision	Recall	F1
T_1: Propositional Segmentation	BERT-Base	33.77	33.53	33.65	14.33	14.60	14.47
	BERT-Large	34.97	33.42	34.17	14.61	14.16	14.38
	T5-Base	54.96	51.93	53.41	32.87	31.54	32.19
	T5-Base *w/ Entail.*	53.54	51.50	52.50	31.61	30.67	31.13
	T5-Large	**55.95**	**55.05**	**55.50**	**32.40**	**32.16**	**32.28**
	T5-Large *w/ Entail.*	**56.27**	**55.50**	**55.89**	**31.94**	**32.11**	**32.02**
	Human Performance	69.63	64.69	67.07	44.86	42.93	43.87
		Performance (2-way Class.)			Per-Label F_1 (3-way Class.)		
		Accuracy	Balanced Accuracy		Entail.	Neutral	Contra.
T_2: Propositional Entailment	*Always Entails.*	27.89	50.00		43.62	0.00	0.00
	Always Neutral	72.10	50.00		0.00	83.54	0.00
	T5-Base	85.17	81.44		73.32	89.68	11.21
	T5-Large	**91.38**	**89.75**		**84.78**	**93.98**	**20.34**
	Human Performance	90.20	88.31		-	-	-

Table 3: Performance of the baseline models on the full (WIKI + NEWS) test set. Due to the low presence of contradictions ($32/8643 = 0.4\%$ of test), F_1 for contradiction does not reflect statistically significant improvement.

positions of the foremost tokens of each proposition in the sentence, yields the best performance.

We start from the pretrained T5 1.1 checkpoints from the T5x library (Roberts et al., 2022). Given a sentence input, we finetune the T5 model to output the propositions in a single sequence. For each input sentence, we sort the output propositions using the aforementioned ordering scheme, and join them by a special token `[TARGET]`. The spans of tokens included in each proposition is surrounded by special tokens `[M]` and `[/M]`. For instance, *" `[M]`Alice `[/M]` and Bob `[M]`went to the Zoo `[/M]`. `[TARGET]` Alice and `[M]`Bob went to the Zoo. `[/M]` "*. In addition, we evaluate the setting where the model is also given the premise document D_{prem}, and learns to output the entailment label along with each proposition (T5 *w/ Entail.* in Table 3).

Encoder+Tagger: BERT (Devlin et al., 2019) For comparison, we provide a simpler baseline that does not model joint probability of the output propositions. On top of the last layer an encoder model, i.e. BERT, we add k linear layers that each correspond to one output proposition. Given an input sentence, the i^{th} linear layer produces a binary $(0/1)$ label per token, indicating whether the token is in the i^{th} proposition or not. k is set to be a sufficiently large number, e.g. $k = 20$ in our experiments. We use the label of the `[CLS]` token of the i^{th} linear layer to indicate whether the i^{th} proposition should exist in the output. For such, we follow the same ordering of the output propositions as in the seq2seq (T5) baseline setup.

4.2 Propositional Entailment Baselines

We formulate the task as a sequence labeling problem, and finetune T5 model as our baseline. The inputs consist of the hypothesis proposition p with its document context D_{hyp}, plus the premise document D_{prem}. The output is one of the three-way labels {*Entailment, Neutral, Contradiction*}. Due to low presence of contradictions, we merge the *neutral* and *contradiction* outputs from the model as *non-entailments* during evaluation. To ensure that the model has access to the essential context information, our task input also include the document D_{hyp} of the hypothesis proposition p, so that model has a decontextualized view of p when inferring its textual entailment relation with D_{prem}.

5 Experiments and Results

5.1 Evaluation Metrics

Propositional Segmentation We measure the precision and recall between the set of predicted and gold propositions for a given sentence. As the set of gold propositions do not follow any particular ordering, we first produce a bipartite matching between them using the Hungarian algorithm (Kuhn, 1955). We again use the Jaccard similarity over $\theta = 0.8$ as a fuzzy match between two propositions (§ 3.2). We also use exact match, an even more restrictive measure where two propositions match if and only if they have the exact same tokens. We report the macro-averaged precision and recall over sentences in the test set.

Propositional Entailment We report the baseline performance under two-way classification re-

Train Domain	Test Domain ($P/R/F_1$ *w/ Jaccard* $\theta = 0.8$) WIKI	NEWS
WIKI	53.95/53.16/53.56	44.93/44.95/44.94
NEWS	45.21/43.65/44.42	49.58/47.81/48.68

Table 4: Cross-domain (i.e. train on NEWS → test on WIKI, and train on WIKI → test on NEWS) generalization results of T5-large on the segmentation (T_1) task.

sults in accuracy. We also report the balanced accuracy, i.e. average of true positive and true negative rate, due to label imbalance (Table 2). To understand the per-label performance, we also report the F_1 score w.r.t. each of the three-way label.

5.2 Baseline Results

Table 3 shows the evaluation results for the segmentation (T_1) and entailment task (T_2) respectively.

For the segmentation task (T_1), the seq2seq T5 model setup yields superior performance compared to the simpler encoder+tagger BERT setup. As the encoder+tagger setup predicts each proposition individually, and does not attend on other propositions during inference, we observe that the model predicts repeated/redundant propositions in $> 20\%$ of the input sentences. In the seq2seq T5 setup, the repetition rate is $< 1\%$. For both setups, we remove the redundant outputs as a post processing step. We also evaluate the multi-task setup (i.e. T5 *w/ Entail.* in Table 3) where the model jointly learns the entailment label with each proposition, and observe no significant improvements. For the entailment task (T_2), we see that T5-Large yields the best overall performance. We observe that the performance with respect to the *entailment* label is lower compared to the *neutral* label.

For both tasks, we estimate the averaged human expert performance by comparing annotations from three of the authors to ground truth on 50 randomly sampled examples from the dataset. We observe that for the segmentation task T_1, we observe that the human performance increases after reconciling and selecting the ground truth response ($0.57 \rightarrow 0.67$ F_1). We see that there remains a sizable gap between the best model, T5-Large, and human performance. On the entailment task T_2, T5-Large exceeds human performance, which is not uncommon among language inference tasks of similar classification formats (Wang et al., 2019).

Document: The incident happened near Dr Gray's Hospital shortly after 10:00. The man was taken to the hospital with what police said were serious but not life-threatening injuries. The A96 was closed in the area for several hours, but it has since reopened.

Summary w/ human labeled hallucinated spans:
A man has been taken to hospital following a one-vehicle crash on the A96 in Aberdeenshire.

Predicted propositions (blue) and entailment labels
#1: A man has been taken to hospital following a one-vehicle crash on the A96 in Aberdeenshire. ✔
#2: A man has been taken to hospital following a one-vehicle crash on the A96 in Aberdeenshire. ✗
#3: A man has been taken to hospital following a one-vehicle crash on the A96 in Aberdeenshire. ✗
#4: A man has been taken to hospital following a one-vehicle crash on the A96 in Aberdeenshire. ✗

Predicted hallucinated spans (union of ✗- union of ✔)
A man has been taken to hospital following a one-vehicle crash on the A96 in Aberdeenshire.

Table 5: An example model generated summary on the XSum dataset, with human-annotated hallucination spans from Maynez et al. (2020). We show that we can infer the hallucinated spans from the set of four propositions and their entailment labels (*entail*=✔, *not-entail*=✗), predicted by our T5-Large models. More examples can be found in Appendix E

5.3 Cross-Domain Generalization

On the propositional segmentation (T_1) task, we evaluate the how the best baseline model generalizes across the Wikipedia (`Wiki`) and `News` domains. Table 4 shows the results of T5-Large models finetuned on data from each domain, and evaluated on the test split of both domains.

When applying a model trained on `Wiki`, we see a larger drop in performance when tested on `News`, as the `News` domain features more syntactic and stylistic variations compared to the `Wiki` domain.

6 Analysis and Discussion

We exemplify the utilities of our propositional segmentation and entailment framework, which we refer to as PropNLI, through the lens of two downstream use cases, e.g. summary hallucination detection (§ 6.1), and document-level NLI w/ variable-length hypotheses (§ 6.2).

6.1 Application: Hallucination Detection

We first look at the task of summary hallucination detection, i.e. given a summary of a source document, identify whether the summary's content is *faithful* to the document. Naturally the task can be represented as a NLI problem, and NLI systems

Method	Hallu. Class.	Span Detection					
		Faith. Tokens			*Hallu. Tokens*		
	B. Acc.	*P*	*R*	F_1	*P*	*R*	F_1
PropNLI	**.62**	.78	.50	.61	.64	.71	.67
MNLI	.59	.96	.17	.30	.56	.88	.68

Table 6: Zero-shot performance of PropNLI vs. T5-Large MNLI model on hallcination identification and span detection tasks from Maynez et al. (2020).

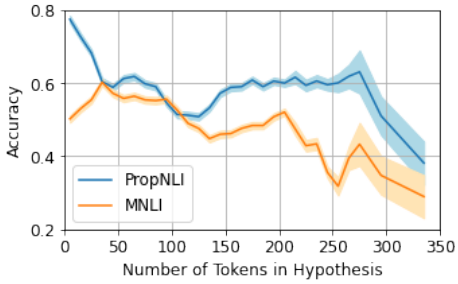

Figure 4: Zero-shot performance of T5-large MNLI model compared to our PropNLI T5-large models (i.e. segmentation '→ entailment → aggregation) with respect to varying *hypothesis length* in DocNLI dev set. The shaded region shows 95% confidence interval.

have been shown effective on the task (Kryscinski et al., 2020; Chen et al., 2021). As summaries can be long and compositional, recognizing partial entailment, and identifying which part(s) of a summary is hallucinated becomes important (Goyal and Durrett, 2020; Laban et al., 2022).

To show that PropNLI can be used for hallucination detection, we experiment on the model generated summaries on the XSum dataset (Narayan et al., 2018), where Maynez et al. (2020) provide human annotations of the sets of hallucinated spans (if they exist) in the summaries. Table 5 illustrates our idea. If a proposition in a summary is *entailed* by the document, then all spans covered by the proposition are faithful. Otherwise, *some* spans would likely contain *hallucinated* information.

Following such intuitions, we first evaluate the performance of our method in zero-shot settings as a hallucination classifier , i.e. binary classification for whether a summary is hallucinated or not. For baseline comparison, we use a T5-large model finetuned on MNLI (Williams et al., 2018) to classify a full summary as entailed (→ *faithful*) or not (→ *hallucinated*). As ˜89% of the summaries annotated by Maynez et al. (2020) are hallucinated, we again adopt balanced accuracy (§ 5.1) as the metric. On 2500 examples, our method achieved 61.68% balanced accuracy, while MNLI achieved 58.79%.

Next, we study whether the entailment labels of propositions can be composed to detect hallucinated spans in a summary. As in Table 5, we take the union of the spans in *non-entailed* propositions, and exclude the spans that has appeared in *entailed* propositions. The intuition is that the hallucinated information likely only exists in the non-entailed propositions , but not the entailed ones.

We evaluate hallucinated span detection as a token classification task. For each summary, we evaluate the precision and recall of the *faithful* and *hallucinated* set of predicted tokens respectively against the human-labeled ground truth set. We report the macro-averaged precision, recall and F_1 score over all 2,500 summaries. We compare our method to a T5-Large model finetuned on MNLI, where we label all tokens as *faithful* if the summary is predicted to be *entailed*, and all tokens as *hallucinated* otherwise. We report the performance with respect to each of the two labels in Table 6. As the MNLI model don't distinguish partial entailment from non-entailment cases, it predicts more tokens to be hallucinated, and thus having low precision and high recall on the hallucinated tokens, and vice versa. On the other hand, we observe our model can be used to detect the nuance between faithful and hallucinated tokens with good and more balanced performance for both cases. Table 5 shows one example summary and PropNLI's predictions, and we include more examples in Appendix E.

6.2 Proposition-Level → Sentence/Paragraph-Level Entailment

We would like to see whether proposition-level entailment labels can potentially be *composed* to explain sentence/paragraph-level NLI predictions.

Given a hypothesis sentence/paragraph and a premise, our PropNLI framework takes three steps. First we segment the hypothesis into propositions. For each proposition, we infer its entailment relation with the premise. In cases where multiple propositions exist in the hypothesis, the proposition-level entailment labels can be aggregated to obtain the entailment label for the *entire* hypothesis, similar to ideas presented in Stacey et al. (2022). As a starting point, we assume logical conjunction as the aggregation function, and hypothesize that this will offer a more fine-grained and explainable way of conducting NLI inference.

To demonstrate the utility of the idea, we conduct a case study on DocNLI (Yin et al., 2021), which features premise and hypothesis of differ-

ent length, and so varying number and compositions of propositions. We take the baseline T5-Large segmentation and entailment models respectively, and use logical conjunction to aggregate the proposition-level entailment prediction. We compare PropNLI in a zero-shot setting against the T5-Large MNLI model. The MNLI model takes the entire hypothesis and premise and input without any segmentation or decomposition.

The results are shown in Figure 4. We take the development set of DocNLI and split examples into buckets according to number of tokens in the hypothesis. We examine the zero-shot performance of the PropNLI setup versus the finetuned MNLI model. We observe that with shorter hypotheses (< 100 tokens), the two setups demonstrated similar performance, as the hypothesis length is similar to the distribution of MNLI training set (avg. 21.73 tokens ± 30.70). As the length of the hypothesis increases, the performance of MNLI model starts to drop, while PropNLI's performance remains relatively stable. Such observations suggest the potential of using the PropNLI framework to describe the textual entailment relations between a pair of premise and hypothesis in a more precise and fine-grained manner. In the realistic case where input hypotheses are compositional, the PROPSEGMENT present an opportunity for developing more generalizable NLI models and solutions.

7 Conclusion

In this paper, we presented PROPSEGMENT, the first large-scale dataset for studying proposition-level segmentation and entailment. We demonstrate that segmenting a text expression into propositions, i.e. atomic units of meanings, and assessing their truth values would provide a finer-grained characterization of the textual entailment relation between two pieces of text. Beyond NLI/RTE tasks, we hypothesize that proposition-level segmentation might be helpful in similar ways for other text classification tasks as well. We hope that PROPSEGMENT will serve as a starting point, and pave a path for research forward along the line.

Limitations

Since the PROPSEGMENT dataset feature entailment labels for *all* propositions in a document, the label distribution are naturally imbalanced, which would potentially pose challenge for modeling. We observe low presence of contradiction examples in our dataset construction process, which could be a limiting factor for the utility of the dataset. Unlike previous NLI datasets (Bowman et al., 2015; Williams et al., 2018), we speculate that reference determinacy, i.e. whether the hypothesis and premise refer to the same scenario at the same time, cannot be certainly guaranteed and safely assumed in our case, which in part leads to low presence of contradictions during annotation. We offer a detailed discussion on the implications of reference determinacy and contradictions in Appendix C. We leave the exploration on *natural* contradictions for future work.

As the annotation complexity and cost scales quadratically w.r.t. the number of propositions in a document, we truncate the documents in PROPSEGMENT to the first ten sentences of the original document.

Ethical Considerations

In the proposition-level entailment task (T_2), the inference of the entailment relation between a premise document and a hypothesis proposition uses the *assumption* that the premise document is true. The assumption is common to NLI datasets (Dagan et al., 2005; Bowman et al., 2015; Williams et al., 2018), and is necessary for the task's structure. With the documents in PROPSEGMENT, we make the assumption only for the experimental purpose of T_2, and make no claim about the actual veracity of the premise documents.

Acknowledgements

We thank Michael Collins, Corinna Cortes, Paul Haahr, Ilya Kornakov, Ivan Kuznetsov, Annie Louis, Don Metzler, Jeremiah Milbauer, Pavel Nalivayko, Fernando Pereira, Sandeep Tata, Yi Tay, Andrew Tomkins, and Victor Zaytsev for insightful discussions, suggestions, and support. We are grateful to the annotators for their work in creating PROPSEGMENT.

References

Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. TensorFlow: a system for Large-Scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Lei Cui, Furu Wei, and Ming Zhou. 2018. Neural open information extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 407–413, Melbourne, Australia. Association for Computational Linguistics.

I. Dagan and O. Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. In *Learning Methods for Text Understanding and Mining*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio. 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*, pages 355–366.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Farima Fatahi Bayat, Nikita Bhutani, and H. Jagadish. 2022. CompactIE: Compact facts in open information extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 900–910, Seattle, United States. Association for Computational Linguistics.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale QA-SRL parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060, Melbourne, Australia. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Kiril Gashteovski, Rainer Gemulla, Bhushan Kotnis, Sven Hertling, and Christian Meilicke. 2020. On aligning OpenIE extractions with knowledge bases: A case study. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 143–154, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, You Wu, Cong Yu, Daniel Finnie, Hongkun Yu, Jiaqi Zhai, and Nicholas Zukoski. 2020. Generating representative headlines for news stories. In *Proceedings of The Web Conference 2020*, pages 1773–1784.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Paul R Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *LREC*, pages 1989–1993.

Ayal Klein, Eran Hirsch, Ron Eliav, Valentina Pyatkin, Avi Caciularu, and Ido Dagan. 2022. QASem parsing: Text-to-text modeling of QA-based semantics. *arXiv preprint arXiv:2205.11413*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Omer Levy, Torsten Zesch, Ido Dagan, and Iryna Gurevych. 2013. Recognizing partial textual entailment. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 451–455, Sofia, Bulgaria. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank project: An interim report. In *Proceedings of the workshop frontiers in corpus annotation at hlt-naacl 2004*, pages 24–31.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, et al. 2022. Scaling up models and data with t5x and seqio. *arXiv preprint arXiv:2203.17189*.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. Association for Computational Linguistics.

Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. 2022. Stretching sentence-pair NLI models to reason over long documents and clusters. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 394–412, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, and Marek Rei. 2022. Logical reasoning with span predictions: Span-level logical atoms for interpretable and robust nli models. In *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

Asher Stern and Ido Dagan. 2014. Recognizing implied predicate-argument relationships in textual inference. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 739–744, Baltimore, Maryland. Association for Computational Linguistics.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. 2016. Order matters: Sequence to sequence for sets. In *Proceedings of the International Conference on Learning Representations*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform

for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

A Model Implementation

T5 We use T5 1.1 checkpoints from the T5x library (Roberts et al., 2022), with Flaxformer[2] implementation. For all sizes of T5 model and all tasks, we finetune the model for three epoch, with $1e-3$ learning rate, 0.1 dropout rate, batch size of 128. We train the models on 16 TPU v3 slices.

BERT We use the BERT English uncased models from Tensorflow (Abadi et al., 2016), in large (24 layers, 16 attention heads, 1024 max sequence length) and base (12 layers, 12 attention heads, 768 max sequence length) sizes. For both sizes, we finetune the model for five epoch, with $1e-5$ learning rate, 0.1 dropout rate, batch size of 16. We train the models on 8 TPU v3 slices.

B Annotation Guidelines

B.1 Segmentation annotation guidelines

There is no unequivocally unique definition for precisely how to segment an English sentence in the context of a document into propositions defined as token subsets, due to a variety of complex language phenomena. Our raters were instructed to follow the following overall guidelines for the segmentation task:

1. Each proposition is expected to correspond to a distinct fact that a reader learns directly from reading the given sentence.
   (a) The raters are instructed to focus on the text's most literal *denotation*, rather than drawing further inferences from the text based on world knowledge, external knowledge, or common sense.
   (b) The raters are instructed to consider *factivity*, marking only propositions that, in their judgement, the author intends the reader to take as factual from reading the sentence.
   (c) With regard to quotes, raters are asked to estimate the author's intent, including the proposition quoted when the reader is expected to take it as factual, and/or the proposition of the quote itself having been uttered if the reader is expected to learn that a speaker uttered that quote.
   (d) The raters are instructed to omit text that are clearly non-factual, such as rhetorical

[2] https://github.com/google/flaxformer

flourishes or first-person account of an article author's emotional response to the topic. This rule is specific to the news and Wikipedia domains, since in other domains of prose, first-person emotions may well be part of the intended informational payload.

2. Each proposition should include all tokens within the sentence that are relevant to learning this fact.
   (a) Specifically, the raters are asked to include any tokens in the same sentence that are antecedents of pronouns or other endophora in the proposition, or relevant bridging references.
   (b) Raters are asked to ignore punctuation, spacing, and word inflections when selecting tokens, though a number of other minutiae, such as whether to include articles, are left unspecified in the rater instructions.

3. Choose the simplest possible propositions, so that no proposition is equivalent to a conjunction of the other propositions, and so that the union of all of the sentence's proposition gives us all the information a reader learns from the sentence.

The raters are also asked to omit propositions from any text that doesn't constitute well-formed sentences, typically arising from parsing errors or from colloquialisms.

Note that the resulting subsets of tokens do not, generally, constitute well-formed English sentences when concatenated directly, but can, in our ad hoc trials, easily be reconstituted into stand-alone sentences by a human reader.

B.2 Entailment annotation guidelines

For the propositional entailment task, our instructions are somewhat similar to the RTE task (Dagan and Glickman, 2004), but specialized to the proposition level.

The raters are asked to read the premise document and decide whether a specific hypothesis proposition is entailed by it, contradicted, or neither. In the first two cases, the raters are asked to mark a proposition in the premise document that most closely supports the hypothesis proposition, using the same definition of proposition as above. The interface nudges the raters to select one of the propositions marked by the segmentation rater, but allows the entailment rater to create a new proposition as well. Note that the choice of a specific supporting proposition is sometimes not well defined.

To judge entailment, the raters are asked "from reading just the premise document, do we learn that the hypothesis proposition is true, learn that it's false, or neither?" More specifically, the raters are asked:

1. To consider the full document of the hypothesis as the context of the hypothesis proposition, and the full premise document.

2. To allow straightforward entailment based on "common sense or widely-held world knowledge", but otherwise avoid entailment labels whenever "significant analysis" (any complex reasoning, specialized knowledge, or subjective judgement) is required to align the two texts.

3. To assume that the two documents were written in the same coarse spatiotemporal context — same geographical area, and the same week.

Raters have the option of marking that they don't understand the premise and/or the hypothesis and skipping the question.

C Reference Determinacy and Contradictions

The PROPSEGMENT dataset is constructed in document-to-document comparison settings. Even though the document clusters are sampled so that documents in a cluster target the same event or event, the documents typically have different focus. Besides the factual information, which are mostly consistent across documents, the focus or specific perspective of each document varies largely, which is in part why we observe very few contradictions.

Apart from such, We speculate that the low presence of contradictions can also be in part attributed to the difficulty in establishing reference determinacy, i.e. whether the entities and events described in a hypothesis can be assumed to refer to the same ones or happening at the same point in the premise. To illustrate the importance of this, consider the following example from SNLI (Bowman et al., 2015).

> *Premise*: A black race car starts up in front of a crowd of people.

Hypothesis: A man is driving down a lonely road.

In SNLI, reference determinacy is assumed to be true. In other words, the human raters assume that the scenario described in the premise and hypothesis happens in the same context at the same time point. Therefore, the example pair is labeled as contradiction, as "lonely road" contradicts "a crowd of people" if we assume both happen on the same road. Without such assumption, the example would likely be labeled as *neutral*, since there is no extra context that would indicate the two events happen in the same context.

In reality, reference determinacy is often difficult to establish with certainty. Unlike existing NLI/RTE datasets (Dagan et al., 2005; Bowman et al., 2015; Williams et al., 2018), in the creation process of PROPSEGMENT, we do not assume reference determinacy between the hypothesis proposition and premise document, but rather relay the judgement to human raters by reading context information presented in the documents. We observe that it is often hard to tell if a specific proposition within a document can establish reference determinacy with the other document, unless the proposition describes a property that is stationary with respect to time. For this reason, most contradictions, among the few that exist in our dataset, are factual statements. Here is an example from the development split.

Premise: ... The team was founded in 1946 as a founding member of the All-America Football Conference (AAFC) and joined the NFL in 1949 when the leagues merged..
Hypothesis: The 49ers have been members of the NFL since the AAFC and National Football League (NFL) merged in 1950...

We view the lack of contradictions as a potential limitation for the dataset for practical purposes. We argue for the need to circumscribe the exact definition of contradiction (from the practical perspective) when reference determinacy cannot be simply assumed. We leave this part for future work.

D Example Propositions From OpenIE vs. PROPSEGMENT

To illustrate the difference between how we define propositions in PROPSEGMENT, versus OpenIE formulations, we include a few examples sentences with propositions in PROPSEGMENT in Table 7 and 8, and compare propositions extracted with ClausIE, a rule-based OpenIE model (Del Corro and Gemulla, 2013), and a neural Bi-LSTM model from Stanovsky et al. (2018).

E XSum Hallucination Detection - Examples

Table 9 and 10 show two example documents, with propositions and the inferred hallucinated spans in model-generated and gold summaries by our PropNLI model. We compare the predictions to the annotations of hallucinated span provided by Maynez et al. (2020).

Sentence: The 82nd NFL Draft took place from April 27-29, 2017 in Philadelphia.
PROPSEGMENT
#1: The 82nd NFL Draft took place from April 27-29, 2017 in Philadelphia.
#2: The 82nd NFL Draft took place from April 27-29, 2017 in Philadelphia.
ClausIE
#1: (The 82nd NFL Draft, took place, from April 27-29, 2017 in Philadelphia)
#2: (The 82nd NFL Draft, took place, from April 27-29, 2017)
Neural Bi-LSTM OIE *(Splitting each modifier, i.e. ARGM)*
#1: (The 82nd NFL Draft, took, place, from April 27-29, 2017)
#2: (The 82nd NFL Draft, took, place, in Philadelphia)

Sentence: She has also appeared in films such as Little Women (1994), The Hours (2002), Self Defense (1997), Les Miserables (1998) and Orson Welles y yo (2009).
PROPSEGMENT
#1: She has also appeared in films such as Little Women (1994), The Hours (2002), Self Defense (1997), Les Miserables (1998) and Orson Welles y yo (2009).
#2: She has also appeared in films such as Little Women (1994), The Hours (2002), Self Defense (1997), Les Miserables (1998) and Orson Welles y yo (2009).
#3: She has also appeared in films such as Little Women (1994), The Hours (2002), Self Defense (1997), Les Miserables (1998) and Orson Welles y yo (2009).
#4: She has also appeared in films such as Little Women (1994), The Hours (2002), Self Defense (1997), Les Miserables (1998) and Orson Welles y yo (2009).
#5: She has also appeared in films such as Little Women (1994), The Hours (2002), Self Defense (1997), Les Miserables (1998) and Orson Welles y yo (2009).
#6: She has also appeared in films such as Little Women (1994), The Hours (2002), Self Defense (1997), Les Miserables (1998) and Orson Welles y yo (2009).
#7: She has also appeared in films such as Little Women (1994), The Hours (2002), Self Defense (1997), Les Miserables (1998) and Orson Welles y yo (2009).
#8: She has also appeared in films such as Little Women (1994), The Hours (2002), Self Defense (1997), Les Miserables (1998) and Orson Welles y yo (2009).
#9: She has also appeared in films such as Little Women (1994), The Hours (2002), Self Defense (1997), Les Miserables (1998) and Orson Welles y yo (2009).
#10: She has also appeared in films such as Little Women (1994), The Hours (2002), Self Defense (1997), Les Miserables (1998) and Orson Welles y yo (2009).
ClausIE
#1: (She, has appeared, in films such as Little Women also)
#2: (She, has appeared, in films such as The Hours also)
#3: (She, has appeared, in films such as Self Defense also)
#4: (She, has appeared, in films such as Les Miserables also)
#5: (She, has appeared, in films such as Orson Welles y yo also)
#6: (She, has appeared, in films such as Little Women)
#7: (She, has appeared, in films such as The Hours)
#8: (She, has appeared, in films such as Self Defense)
#9: (She, has appeared, in films such as Les Miserables)
#10: (She, has appeared, in films such as Orson Welles y yo)
#11: (Little Women, is, 1994)
#12: (The Hours, is, 1994)
#13: (Self Defense, is, 1994)
#14: (Les Miserables, is, 1994)
#15: (Orson Welles y yo, is, 1994)
#16: (The Hours, is, 2002)
#17: (Self Defense, is, 1997)
#18: (Les Miserables, is, 1998)
#19: (Orson Welles y yo, is, 2009)
Neural Bi-LSTM OIE
#1: (She, appeared, in films such as Little Women (1994), The Hours (2002), Self Defense (1997), Les Miserables (1998) and Orson Welles y yo (2009))

Table 7: Comparison of propositions in PROPSEGMENT with extractions with ClausIE (Del Corro and Gemulla, 2013), and the neural Bi-LSTM OIE model from Stanovsky et al. (2018).

Sentence: The Andy Warhol Museum in his hometown, Pittsburgh, Pennsylvania, contains an extensive permanent collection of art.
PROPSEGMENT
#1: The Andy Warhol Museum in his hometown, Pittsburgh, Pennsylvania, contains an extensive permanent collection of art.
#2: The Andy Warhol Museum in his hometown, Pittsburgh, Pennsylvania, contains an extensive permanent collection of art.
#3: The Andy Warhol Museum in his hometown, Pittsburgh, Pennsylvania, contains an extensive permanent collection of art.
ClausIE
#1: (his, has, hometown)
#2: (his hometown, is, Pittsburgh Pennsylvania)
#3: (The Andy Warhol Museum in his hometown, contains, an extensive permanent collection of art)
Neural Bi-LSTM OIE
#1: (The Andy Warhol Museum in his hometown Pittsburgh Pennsylvania, contains, an extensive permanent collection of art)

Sentence: The Cleveland Cavaliers got the first choice in the lottery, which was used on 20-year-old forward Anthony Bennett, a freshman from the University of Nevada.
PROPSEGMENT
#1: The Cleveland Cavaliers got the first choice in the lottery, which was used on 20-year-old forward Anthony Bennett, a freshman from the University of Nevada.
#2: The Cleveland Cavaliers got the first choice in the lottery, which was used on 20-year-old forward Anthony Bennett, a freshman from the University of Nevada.
#3: The Cleveland Cavaliers got the first choice in the lottery, which was used on 20-year-old forward Anthony Bennett, a freshman from the University of Nevada.
#4: The Cleveland Cavaliers got the first choice in the lottery, which was used on 20-year-old forward Anthony Bennett, a freshman from the University of Nevada.
#5: The Cleveland Cavaliers got the first choice in the lottery, which was used on 20-year-old forward Anthony Bennett, a freshman from the University of Nevada.
#6: The Cleveland Cavaliers got the first choice in the lottery, which was used on 20-year-old forward Anthony Bennett, a freshman from the University of Nevada.
ClausIE
#1: (The Cleveland Cavaliers, got, the first choice in the lottery)
#2: (the lottery, was used, on 20-year-old forward Anthony Bennett)
#3: (Anthony Bennett, is, a freshman from the University of Nevada)
Neural Bi-LSTM OIE
#1: (The Cleveland Cavaliers, got, the first choice in the lottery, which was used on 20-year-old forward Anthony Bennett, a freshman from the University of Nevada.)
#2: (the lottery, was used, on 20-year-old forward Anthony Bennett, a freshman from the University of Nevada.)

Table 8: (Cont.) Comparison of propositions in PROPSEGMENT with extractions with ClausIE (Del Corro and Gemulla, 2013), and the neural Bi-LSTM OIE model from Stanovsky et al. (2018).

Document: The incident happened near Dr Gray's Hospital shortly after 10:00. The man was taken to the hospital with what police said were serious but not life-threatening injuries. The A96 was closed in the area for several hours, but it has since reopened.

Summary from `BertS2S`
A man has been taken to hospital following a one-vehicle crash on the A96 in Aberdeenshire.
Predicted propositions (blue) and entailment labels
#1: A man has been taken to hospital following a one-vehicle crash on the A96 in Aberdeenshire. ✔
#2: A man has been taken to hospital following a one-vehicle crash on the A96 in Aberdeenshire. ✗
#3: A man has been taken to hospital following a one-vehicle crash on the A96 in Aberdeenshire. ✗
#4: A man has been taken to hospital following a one-vehicle crash on the A96 in Aberdeenshire. ✗
Predicted hallucinated spans (union of ✗- union of ✔)
A man has been taken to hospital following a one-vehicle crash on the A96 in Aberdeenshire.

Summary from `TConvS2S`
a man has been taken to hospital after being hit by a car in Moray.
Predicted propositions (blue) and entailment labels
#1: a man has been taken to hospital after being hit by a car in Moray. ✔
#2: a man has been taken to hospital after being hit by a car in Moray. ✗
Predicted hallucinated spans (union of ✗- union of ✔)
a man has been taken to hospital after being hit by a car in Moray.

Gold Summary from the XSum dataset
A cyclist has suffered serious head injuries after a collision with a car in Elgin.
Predicted propositions (blue) and entailment labels
#1: A cyclist has suffered serious head injuries after a collision with a car in Elgin. ✗
#2: A cyclist has suffered serious head injuries after a collision with a car in Elgin. ✗
#3: A cyclist has suffered serious head injuries after a collision with a car in Elgin. ✗
Predicted hallucinated spans (union of ✗- union of ✔)
A cyclist has suffered serious head injuries after a collision with a car in Elgin.

Summary from `PTGen`
A man has been taken to hospital after being hit by a car in the A96 area of Glasgow.
Predicted propositions (blue) and entailment labels
#1: A man has been taken to hospital after being hit by a car in the A96 area of Glasgow. ✔
#2: A man has been taken to hospital after being hit by a car in the A96 area of Glasgow. ✗
#3: A man has been taken to hospital after being hit by a car in the A96 area of Glasgow. ✗
Predicted hallucinated spans (union of ✗- union of ✔)
A man has been taken to hospital after being hit by a car in the A96 area of Glasgow

---

Summary from `TranS2S`
A man has been taken to hospital after a two-vehicle crash on the A96 in County Antrim.
Predicted propositions (blue) and entailment labels
#1: A man has been taken to hospital after a two-vehicle crash on the A96 in County Antrim. ✔
#2: A man has been taken to hospital after a two-vehicle crash on the A96 in County Antrim. ✗
#3: A man has been taken to hospital after a two-vehicle crash on the A96 in County Antrim. ✗
#4: A man has been taken to hospital after a two-vehicle crash on the A96 in County Antrim. ✗
Predicted hallucinated spans (union of ✗- union of ✔)
A man has been taken to hospital after a two-vehicle crash on the A96 in County Antrim.

Table 9: More example of model generated summaries on the XSum dataset, with human-annotated hallucination spans from Maynez et al. (2020). For each document, Maynez et al. (2020) provide summaries and hallucination annotations from 5 different summarization systems. We randomly sample documents and show our model's predictions for all 5 summaries here.

Document: Dervite, 28, made 14 appearances last season to help Wanderers finish second in League One and secure promotion. The French centre-back joined Bolton from Charlton in 2014 and has made 83 appearances in all competitions. "Dorian was a bit of a forgotten man last year but came in and made an excellent contribution towards the end of the campaign," manager Phil Parkinson told the club website. Dervite follows David Wheater, Gary Madine and Jem Karacan in signing new contracts with Bolton, following their promotion to the Championship.

Summary from `BertS2S`
Bolton defender Dorian Dervite has signed a new two-year contract with the championship club.
Predicted propositions (blue) and entailment labels
#1: Bolton defender Dorian Dervite has signed a new two-year contract with the championship club. ✔
#2: Bolton defender Dorian Dervite has signed a new two-year contract with the championship club. ✗
Predicted hallucinated spans (union of ✗- union of ✔)
Bolton defender Dorian Dervite has signed a new two-year contract with the championship club.

Summary from `TConvS2S`
Bolton Wanderers have signed defender Dorian Dervite from bolton wanderers for an undisclosed fee.
Predicted propositions (blue) and entailment labels
#1: Bolton Wanderers have signed defender Dorian Dervite from bolton wanderers for an undisclosed fee. ✗
#2: Bolton Wanderers have signed defender Dorian Dervite from bolton wanderers for an undisclosed fee. ✗
#3: Bolton Wanderers have signed defender Dorian Dervite from bolton wanderers for an undisclosed fee. ✗
#4: Bolton Wanderers have signed defender Dorian Dervite from bolton wanderers for an undisclosed fee. ✔
Predicted hallucinated spans (union of ✗- union of ✔)
Bolton Wanderers have signed defender Dorian Dervite from bolton wanderers for an undisclosed fee.

Gold Summary from the XSum dataset
Defender Dorian Dervite has signed a new one-year contract with Bolton.
Predicted propositions (blue) and entailment labels
#1: Defender Dorian Dervite has signed a new one-year contract with Bolton ✔
#2: Defender Dorian Dervite has signed a new one-year contract with Bolton. ✗
Predicted hallucinated spans (union of ✗- union of ✔)
Defender Dorian Dervite has signed a new one-year contract with Bolton.

Summary from `PTGen`
Bolton Wanderers defender Dorian Dervite has signed a new three-and-a-half-year contract with the league one club until the end of the 2018-19 season.
Predicted propositions (blue) and entailment labels
#1: Bolton Wanderers defender Dorian Dervite has signed a new three-and-a-half-year contract with the league one club until the end of the 2018-19 season. ✔
#2: Bolton Wanderers defender Dorian Dervite has signed a new three-and-a-half-year contract with the league one club until the end of the 2018-19 season. ✗
#3: Bolton Wanderers defender Dorian Dervite has signed a new three-and-a-half-year contract with the league one club until the end of the 2018-19 season. ✗
Predicted hallucinated spans (union of ✗- union of ✔)
Bolton Wanderers defender Dorian Dervite has signed a new three-and-a-half-year contract with the league one club until the end of the 2018-19 season.

Summary from `TranS2S`
Bolton Wanderers midfielder Gary Wheat has signed a new one-year contract with the championship side.
Predicted propositions (blue) and entailment labels
#1: Bolton Wanderers midfielder Gary Wheat has signed a new one-year contract with the championship side. ✗
#2: Bolton Wanderers midfielder Gary Wheat has signed a new one-year contract with the championship side. ✗
Predicted hallucinated spans (union of ✗- union of ✔)
Bolton Wanderers midfielder Gary Wheat has signed a new one-year contract with the championship side.

Table 10: (Cont.) More example of model generated summaries on the XSum dataset, with human-annotated hallucination spans from Maynez et al. (2020).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes, in the "Limitations" section

☑ A2. Did you discuss any potential risks of your work?
Yes, in the "Ethical Considerations" section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and introduction (section 1) summarize the main contributions of the paper

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

section 3

☑ B1. Did you cite the creators of artifacts you used?
section 3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We mention that it will be released upon publication in Ethical Considerations

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
section 3

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix A

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix B

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Cannot disclose due to legal reasons (proprietary information).

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Cannot disclose due to legal reasons (proprietary information).

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Cannot disclose due to legal reasons (proprietary information).

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Cannot disclose due to legal reasons (proprietary information).