Stretching Sentence-pair NLI Models to Reason over Long Documents and Clusters

Tal Schuster,[1] **Sihao Chen,**[1,2] **Senaka Buthpitiya,**[1] **Alex Fabrikant,**[1] **Donald Metzler**[1]

[1]Google Research [2]University of Pennsylvania
{talschuster,sihaoc,senaka,fabrikant,metzler}@google.com

Abstract

Natural Language Inference (NLI) has been extensively studied by the NLP community as a framework for estimating the semantic relation between sentence pairs. While early work identified certain biases in NLI models, recent advancements in modeling and datasets demonstrated promising performance. In this work, we further explore the direct zero-shot applicability of NLI models to real applications, beyond the sentence-pair setting they were trained on. First, we analyze the robustness of these models to longer and out-of-domain inputs. Then, we develop new aggregation methods to allow operating over full documents, reaching state-of-the-art performance on the ContractNLI dataset. Interestingly, we find NLI scores to provide strong retrieval signals, leading to more relevant evidence extractions compared to common similarity-based methods. Finally, we go further and investigate whole document clusters to identify both discrepancies and consensus among sources. In a test case, we find real inconsistencies between Wikipedia pages in different languages about the same topic.[1]

1 Introduction

Natural Language Inference (NLI) involves automatically determining whether the meaning of one piece of text (i.e., hypothesis) can be inferred from another (i.e., the premise) (Dagan et al., 2006). This formulation is relatively simple, enabling large-scale data annotation (e.g., Bowman et al., 2015; Williams et al., 2018), yet imposes complex semantic reasoning challenges (e.g., background knowledge, commonsense), leading to a useful training and evaluation NLP framework (Mishra et al., 2021; Sainz et al., 2022; Vu et al., 2021).

Formally, in NLI, we are interested in learning a function $f : (X_{\text{hyp}} \times X_{\text{prem}}) \rightarrow \mathcal{Y}$ that predicts the relation Y between the provided premise X_{prem} and the examined hypothesis X_{hyp}, where $\mathcal{Y} = \{\text{entailment}, \text{neutral}, \text{contradiction}\}$.[2] In most NLI datasets, X_{hyp} and X_{prem} are short texts consisting of one or few sentences, allowing current Large Language Models (LLMs) with limited input length to process the two with cross-attention. In practice, however, many systems require operating over long texts such as full documents or even collections of documents without knowing a priori which parts are most relevant.

Consider the example in Figure 1. The system is trying to reason over a collection of documents and find statements that they all agree upon (consesnus), or alternatively, find potential disagreements across documents. Instead of having a clear hypothesis-premise pair, each statement across all documents is an hypothesis of interest that should be evaluated against all other documents as the premise.

In this paper, we focus on these realistic scenarios and present retrieve-and-classify methods for inferring over long and out-of-distribution (OOD) inputs in a zero-shot fashion. As this setting emphasizes the need for a robust sentence-pair NLI model as a backbone, we train on multiple datasets, including adversarial and contrastive ones (Nie et al., 2020; Schuster et al., 2021) to increase the model's robustness and avoid dataset-specific biases (Gururangan et al., 2018; McCoy et al., 2019; Poliak et al., 2018; Schuster et al., 2019).

Our proposed pipelines go beyond the length and format supported by most LMs and include new retrieval, aggregation and classification solutions—all based on the same classifier. Thereby, we continue the line of work on evaluating the robustness of such models and their ability to truly capture semantic relations. Moreover, long inputs highlight the commonly overlooked yet practically important

[1]Released Wikipedia translated clusters dataset: https://github.com/google-research-datasets/wiki-translated-clusters-nli

[2]This formulation also fits the task of fact-checking a claim against given evidence (Thorne et al., 2018), therefore henceforth we use the NLI terminology for both tasks.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 394–412
December 7-11, 2022

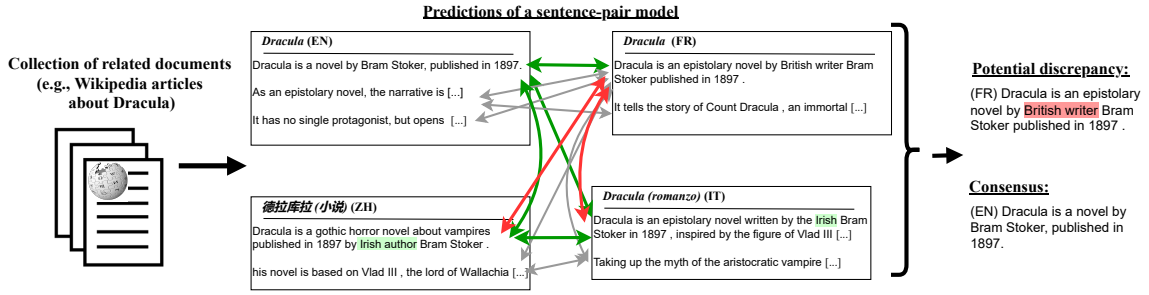

Figure 1: Illustration of our procedure for flagging potential discrepancies in document clusters (§3.4). This is a real-world example from Wikipedia's translated articles in different languages about the novel Dracula (as of Feb. 2022).[3] Our model identified the French Wikipedia to disagree with other articles on the nationality of the author.

challenge of specifying "neutral" relations. For example, mistaking certain part of the premise as entailing instead of neutral can overshadow another segment that contradicts the hypothesis.

We first evaluate the zero-shot performance of our multi-task NLI model on X_{prem}s that are longer than the examples seen in training, but short enough to allow supervised LMs with similar length constraints to perform well (Yin et al., 2021). In the zero-shot setting, we find that models generalize beyond the training distribution, but drop in performance for very long X_{prem}s (> 400 tokens).

Then, we turn to focus on the scenario of having a full document as X_{prem} (Koreeda and Manning, 2021). A typical approach—the default behavior of most LMs—would truncate the end of the document beyond some predefined length (e.g., 512 tokens). However, this might remove and ignore important information. Instead, we hypothesize that a good NLI model should be able to separate the wheat from the chaff and distinguish neutral spans towards X_{hyp} from informative ones. To this end, we develop a solely NLI-based retrieve-and-classify approach that outperforms similarity-based retrievers and whole-document classifiers.

Finally, we go further and demonstrate the utility of our model for reasoning over entire clusters of related documents. Our proposed procedure, illustrated in Figure 1, ranks all of the cluster's spans by their entailment relations with spans from other documents. Testing our approach on Wikipedia introductions on the same topic in different languages, we successfully identify claims that are unique to one version and contradicted by others.

In summary, this work stretches sentence-pair NLI models to new practical capabilities and demonstrates their direct utility in real-world applications. Our main contributions include:

- A multi-task sentence-level NLI model with strong zero-shot and supervised performance for both evidence retrieval and classification.
- Simple and effective retrieve-and-classify methods to extend sentence-pair semantic classifiers and outperform whole-document models.
- A new entailment task and dataset that requires inference over clusters of documents (§4.1).[4]
- Demonstrating the utility of our approach to reveal real and simulated discrepancies in Wikipedia pages by automatically comparing with content from multiple translated articles.

2 Model and Definitions

Our pipelines build on entailment scores for hypothesis-premise pairs. To predict these scores, we train a sentence-pair NLI model that we use as the backbone for all methods. Specifically, we pick the T5 encoder-decoder architecture as it has been shown to perform well in multi-task and transfer settings (Aribandi et al., 2022; Raffel et al., 2020). See Appendix A for more technical details.

Definitions. As T5 is a seq-to-seq model, we train it over the training set $(X_{\text{hyp}}, X_{\text{prem}}, Y) \sim \mathcal{D}_{\text{train}}$ by feeding the following format to the encoder: "*entailment:* X_{hyp} *[SEP]* X_{prem}." The decoder's goal is to generate a single character 'e', 'n', or 'c', representing the three classes in $\mathcal{Y}$: entailment, neutral, and contradiction, respectively. Thereafter, when making a prediction on a new pair $(X_{\text{hyp}}, X_{\text{prem}}) \in \mathcal{D}_{\text{test}}$, we encode the input and measure the decoder's score s_y for each of the three classes. Finally, we normalize the three scores with a softmax operator: $p_y = \text{Softmax}(s_e, s_n, s_c)[y]$. Note that these scores should not be directly treated as class probabilities as they are not calibrated,

[3] https://fr.wikipedia.org/w/index.php?title=Dracula&oldid=190970820

[4] We intend to release this new dataset upon publication.

Train dataset	Hypothesis length	Premise length	Train pairs
MNLI	13.23 (7−20)	27.57 (9−50)	392,702
SNLI	9.50 (5−15)	16.86 (9−27)	550,152
ANLI	13.32 (7−21)	79.95 (53−111)	162,865
FEVER	12.50 (8−18)	47.98 (21−82)	178,059
VitaminC	18.18 (10−29)	43.03 (19−72)	370,653

Table 1: Training datasets of SENTLI. We report the average length of the tokenized (for T5) hypothesis and premise, and the 10th-90th percentiles in parentheses.

especially when evaluating out-of-domain inputs. However, as we will show in our experiments, they can be readily tuned or leveraged as a valuable signal for textual semantic relations.

The SENTLI model. To train our model, we use the following sentence-pair datasets: SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), ANLI (Nie et al., 2020), FEVER (Thorne et al., 2018), and VitaminC (Schuster et al., 2021). For FEVER, we use the sentence-pair version from VitaminC with retrieved evidence to neutral claims.

This multi-task training is important for improving the robustness of the model by leveraging a large amount of diverse data. Also, including adversarial (ANLI) and contrastive (VitaminC) examples was shown to prevent the model from relying on hypothesis-only biases. Table 1 shows the input length statistics of the datasets according to the English sentence-piece tokenizer of T5. Having concise and short inputs generally makes the task less ambiguous—as claim in question is clear—and focused on the textual entailment component.

We denote our T5 mutli-task NLI model trained on all 5 datasets SENTLI. Our application-specific pipelines will differ in the way that they make use of SENTLI's predictions. We also experiment with models trained only on MNLI without multi-tasking (-M.T), only on the three *NLI datasets (-F.V), or with ContractNLI (Koreeda and Manning, 2021) examples.

3 Beyond Sentence-level Inference

We now assume that our target application requires the evaluation of hypotheses against *long texts*. We also assume that we have limited or no training data for this domain, and focus on *zero-shot* transfer. Formally, we assume that X_{prem} is a document consisting of n sentences $S_{1:n}$ and we don't know which part of the document is most relevant for verifying or rejecting X_{hyp}. In this case, it is common to not only classify the truthfulness of the given statement, but to also point to the exact evidence in the document that led to this conclusion. This is a crucial requirement, both benefiting the interpretability and trustworthiness of the model (e.g., avoiding hypothesis-only bias), and saving human time needed for manual prediction verification.

3.1 Naïve premise truncation

A naïve design choice for such applications would be to simply use a similar cross-attention model and provide as much as possible from the input text. As Transformer models are trained with a defined maximum input length limit (typically 512 tokens), this approach has obvious limitations. Yet, in some applications we can assume where the relevant information is likely to be and remove the rest. For example, Yin et al. (2021) found a model that truncates the input to even outperform a model that supports long inputs on DocNLI.

Nevertheless, this approach is unlikely to suit very long inputs as the complexity of Transformers grows quadratically with the input length (Tay et al., 2021; Vaswani et al., 2017). Also, this approach doesn't directly support the important interpretability requirement discussed above, as it is unclear which part of the long document led the model to its prediction.

3.2 Retrieve-and-classify over long premises

Instead, we opt to break the long premise into individual sentences and make pointwise predictions against the hypothesis before aggregating them to the final classification. This approach can readily extend to any document length without modifications to the core NLI model. Also, in zero-shot transfer, this allows better alignment with the sentence-pair training distribution. While it requires n inference runs of the NLI model instead of a single pass, the cost increases linearly with n, unlike the quadratic effect of increasing the input length. Also, these inference passes can be computed in parallel with batches.

This pointwise approach, however, has some limitations. First, it requires a robust sentence-pair model that can separate neutral sentences from relevant ones. Second, it doesn't immediately support multi-hop inference over multiple sentences (Jiang et al., 2020). This can be partially alleviated with preprocessing techniques (e.g., Choi et al., 2021). In practice, we don't observe this limitation in our explored applications as most sentences in these domains are sufficiently self-contained (see Figure 1).

Next, we discuss methods for performing the two key steps of the retrieve-and-classify approach.

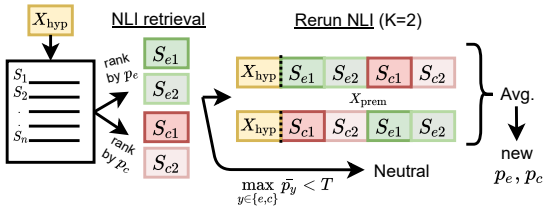

Figure 2: NLI-based retrieve-and-rerank concatenates the K spans with the strongest 'entail' score, and the K with the strongest 'contradict' score for reranking, as long as some span's non-neutral score exceeds T.

3.2.1 Retrieval

Given the long multi-sentenced premise, we would like to identify which sentences are most helpful for accepting or rejecting the hypothesis. We focus on methods supporting data-sparse target domains.

Similarity-based retrieval. Most unsupervised retrievers use a similarity function (e.g., inner product) over sparse or dense representations to compare X_{hyp} and each of the premise sentences.

NLI-based retrieval. We introduce an alternative retrieval approach where we use the non-neutral scores of an NLI model to determine the usefulness of a sentence for classifying the hypothesis. This is motivated by the fact that two neutral sentences can be very similar by many measures, but uninformative for our purpose, such as, e.g., X_{prem} ="*We discuss later whether* X_{hyp}".

Here, each sentence ends up having two ranking scores with respect to the hypothesis: for entailment and for contradiction. As we see next, this granularity is useful for downstream steps such as reranking or reasoning over clusters.

3.2.2 Classification

Following the evidence retrieval, we define two methods for constructing the final prediction:

Retrieve-and-Predict. Assuming NLI-based retrieval, we can simply reuse the same scores. We pick the span with the strongest score for each label, $\bar{p_y} = \max_{i\in[1,n]}(p_{y,i})$, and then predict by the highest NLI score:

$$\text{prediction} = \begin{cases} \arg\max_y \bar{p_y} & \text{if } \max_{y\in\{e,c\}} \bar{p_y} > T, \\ \text{'neutral'} & \text{otherwise.} \end{cases}$$

We simply set $T = 0.5$ for the zero-shot setting, but it can be potentially tuned. Effectively, the prediction is determined by the span with the strongest sentiment towards the hypothesis.

Retrieve-and-Rerank. Instead of directly using the retrieval scores, we rerun the same NLI model on the original hypothesis, and a concatenation of the top-K spans retrieved for both non-neutral labels. For symmetry, we average over two instances: the first concatenating the top-entailing spans in score order, then the top-contradicting spans; and the second instance switching the entailing and contradicting spans. Figure 2 illustrates this process with $K = 2$.

This reranking allows the NLI model to directly contrast the spans that are most entailing with the spans that are most contradicting toward the hypothesis. The resultant multi-sentence premise is longer than the training distribution. Yet, we find SENTLI to generalize well to slightly longer but focused premises.

Algorithm 1 Find factual discrepancies in a cluster.

Input: Cluster of documents $\mathcal{D}$ with spans $\mathcal{S}$, and NLI model.
Output: Sorted spans by discrepancy likelihood.

```
ω_{i,j} ← 0   ∀ span j in document i
for D_i ∈ D do
    for S_j ∈ D_i do
        X_hyp ← S_{i,j};   Ω ← {}
        for D_k ∈ D \ D_i do
            Γ ← {}
            for S_{k,l} ∈ D_k do
                p_c ← SENTLI(X_hyp, S_{k,l})[c]
                Γ ← Γ ∪ {p_c}
            Ω ← Ω ∪ max(Γ)
        ω_{i,j} ← mean(Ω)
return S sorted by respective ω
```

3.3 Multi-sentence hypotheses

NLI models could also be useful in scoring texts that are longer than a single focused statement. Recently, Laban et al. (2022) used NLI scores to predict the faithfulness of generated summaries with respect to their source. Here, we can break both the premise (e.g., the source) and the hypothesis (the summary) into spans, retrieve-and-classify (with or without reranking) for each hypothesis span, and aggregate. We consider two methods for aggregating the scores of the hypothesis spans:

Soft aggregation. Following Laban et al. (2022), we take the average entailment scores across spans. When using our reranking method, we take the shifted difference between the scores: $p_e - p_c + T$.

Hard aggregation. We take the minimum entailment score across spans, effectively requiring all of the hypothesis spans to be strongly supported.

3.4 Reasoning over multi-document clusters

So far, we dealt with evaluating a single, short or long hypothesis against a long premise. In some applications, however, the user might not know

Eval dataset	Hypothesis length	Premise length	Docs per premise	Sents per doc	Sentence length
DocNLI (-ANLI) (Yin et al., 2021)	98.91 (52-144)	530.24 (70-1312)	1	17.53 (3 - 43)	30.47 (12-51)
ContractNLI (Koreeda and Manning, 2021)	19.35 (11-27)	2408.75 (939-4409)	1	79.63 (36 - 128)	30.28 (1-74)
SUMMAC (Laban et al., 2022)	62.63 (22-117)	678.52 (240-1234)	1	22.16 (8-41)	30.38 (11-50)
Wiki Clusters (Section 4.1)	34.83 (17-58)	2498.39 (999-4152)	9.86 (10-10)	7.79 (2-16)	32.52 (13-56)

Table 2: Evaluation datasets and the average tokenized lengths (with 10/90th percentiles), number of sentences per premise document, and length of each sentence. The premise in our Wiki clusters consists of multiple documents.

which hypothesis to check, but rather would like to query over their corpus to identify the most extreme ones. For example, consider a collection of news articles on the same topic written by different sources. Typical questions to ask about this corpus could be: *"is there any claim made by one article that other articles disagree with?"*, or other queries like *"what is the most controversial claim?"* or *"is there consensus on some claims in the corpus?"*

Answering such questions goes beyond the typical NLI setting and requires understanding a complex many-to-many relation between the documents, involving multiple alignment and reasoning challenges. Therefore, any solution with low signal-to-noise ratio is likely to fail.

Using our robust SENTLI model, we introduce an algorithm for identifying such claims. Algorithm 1 ranks all of the cluster's spans by discrepancy likelihood.[5] Each span is compared against all other spans from all documents. The score is determined by the most contradicting pairing from each document and averaged across the cluster. While this procedure requires many calls to the NLI model (quadratic in number and size of documents), they are independent and can easily be batched and parallelized.

4 Evaluation Tasks and Datasets

We evaluate our methods on the following 3 benchmarks that contain 9 datasets from different domains. In addition, we create a new of its kind dataset with clusters of related documents (§4.1). The statistics are summarized in Table 2.

DocNLI (Yin et al., 2021) includes long hypotheses and premises, mostly from the news domain. We remove ANLI since it was included in our training data.[6] Despite the length, a model with input limit of 512 tokens can generally perform well.

DocNLI uses only two classes, "entail" vs. "not entail". We discuss different zero-shot conversion techniques from 3-way models to binary classification in Appendix B.

ContractNLI (Koreeda and Manning, 2021) has NLI examples in the legal domain. Each hypothesis is short and focused, but the premise is a long document (80 sentences on average). A model with input limit of 512 tokens performs poorly here.

SUMMAC (Laban et al., 2022) is a benchmark for predicting the factual consistency of summaries with their source. We follow the zero-shot setting here, but for fair comparison with Laban et al. (2022), we also tune the threshold on the validation set for each dataset, and report the results on the test set. In early exploration we found naive threshold settings to be competitive as well.

4.1 Wikipedia clusters evaluation dataset

In addition, we create a new dataset for exploring *inference over collections of related articles*. Specifically, we collect clusters of introductions to popular Wikipedia articles on the same topic written in up to 11 different languages, machine translated to English. See App. E for more details.[7]

Each version of Wikipedia is managed by a different community, leading to occasional disagreements or mistakes (IV et al., 2021; Vrandecic, 2020), or even the risk of version-specific conspiracy theories.[8] Therefore, automatically comparing and contrasting the information from different articles could be very helpful. We examine both synthetic corruptions and real discrepancies.

Corrupted articles. We simulate a corruption to the English version of each article by inserting a local edit to one of the sentences. The task is to use the other articles of that cluster to identify the sentence that was changed. While it is possible that all other articles don't mention any information about the specific corrupted fact, thanks to the popularity of the chosen articles and languages we find that mostly at least one of the articles includes sufficient information to refute the corrupted sentence.

To create the corruptions, we use edits from the

[5] When looking for consensus, p_c is replaced by p_e.

[6] The ANLI examples cover only 1.2% of the DocNLI test set, so the difference is minimal. SENTLI and SENTLI $_{\text{tuned}}$ get .350 and .410 F_1 scores on the full test set, respectively.

[7] The Wikipedia clusters data is available at: https://github.com/google-research-datasets/wiki-translated-clusters-nli

[8] www.bbc.com/news/technology-59325128

Model	Dev.	Test
Random	.198	.199
supervised:		
Longformer-base* (Yin et al., 2021)	.462	.444
Roberta-large (Yin et al., 2021)	.631	.613
T5-large	.642	.618
zero-shot:		
SENTLI (no setnence split)	.341	.345
SENTLI $_{\text{tuned}}$ (no setnence split)	.409	.408

Table 3: F_1(E) scores on the DOCNLI(-ANLI) binary classification dataset. The zero-shot shot predictions are based on a threshold T on the entailment score which is either set to 0.5 or *tuned* over 0.2% of the dev set. *Longformer's scores are over the full DocNLI.

test set of the VitaminC dataset (Schuster et al., 2021) that express opposite relations towards a mutual claim. In total, we create 824 instances based on 144 different topics. In each instance, we corrupt a single sentence from one of the English articles, and provide the 10 related articles from other languages to help identify which fact was changed.

We note that SENTLI observed Wikipedia sentences (from other articles) in the training mixture. However, they were only used as the premise, whereas here they also represent the hypothesis.

Real discrepancies. We look for discrepancies in-the-wild, searching in current Wikipedia. Here, we don't know which article, if any, might include a discrepancy. Therefore, in this setting, we focus on qualitative evaluation and explore whether we can rank all spans from all articles to identify real discrepancies, or consensus.

5 Experiments

We evaluate our inference pipelines against supervised models from DocNLI and ContractNLI, and the zero-shot SUMMAC model, adopting the main evaluation metrics from each paper. We also train T5-large supervised models to directly compare with SENTLI's zero-shot performance.

DocNLI (Yin et al., 2021) used a RoBERTa-large (Liu et al., 2019) model with input limit of 512 tokens and a Longformer-base (Beltagy et al., 2020) model. ContractNLI introduced the SpanNLI (Koreeda and Manning, 2021) model to process long documents that they train to jointly identify key spans and to make the final verdict. SUMMAC (Laban et al., 2022) used a BERT-large (Devlin et al., 2019) model that was also trained on multiple NLI datasets.

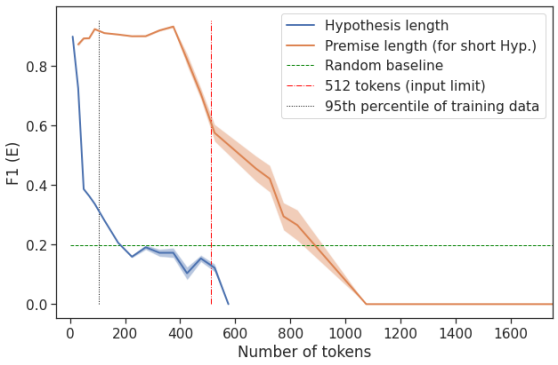

Figure 3: Effect of input length of DocNLI examples when naively using zero-shot SENTLI **without sentence splitting** (simply providing the whole document as a single premise). The blue line shows F_1 score as a function of hypothesis length ($|X_{\text{hyp}}|$). The orange line bins examples by the premise length ($|X_{\text{prem}}|$), focusing only on short hypotheses (where $|X_{\text{hyp}}| \leq 20$). For short hypotheses, the performance stays high even beyond the training distribution, but sharply drops around the input length limit of the model.

5.1 Zero-shot transfer to new domains

DocNLI. Table 3 summarizes the results on DocNLI. SENTLI performs much better than a random baseline and is competitive with some supervised models, indicating promising transfer potential.

We also examine the effect of the input length on the performance in Figure 3. First, we see that the performance is highly affected by the length of the hypothesis. Yin et al. (2021) observed a similar trend even with supervised models. We conjecture that this is due to the natural increase in ambiguity with the hypothesis' length, as it is more likely to include multiple claims that could be questioned.

Second, we look at the performance as a function of the premise length. To focus on examples where the hypothesis is well defined, we only consider cases with a short hypothesis of no more than 20 tokens (total of 5,932 cases from DocNLI(-ANLI) test set). As the orange line shows, SENTLI performs well on these cases even when the premise is much longer than the inputs that the model was trained on, demonstrating promising potential for zero-shot applications. However, the performance significantly drops when reaching the input length limit, requiring us to truncate the premise.

ContractNLI. We test zero-shot transfer to the legal domain of ContractNLI. To disentangle the effect of input length, we first examine an oracle retriever setting where the premise includes only the few relevant sentences. Table 4 summarizes the results. Surprisingly, we find our zero-shot

	Model	S.P.	F_1(C)	F_1(E)	AVG
	Majority vote‡		.239	.645	.442
supervised	SpanNLI-base‡		.657	.816	.736
	SpanNLI-large‡		.620	.806	.713
	T5-large	v	**.815**	.971	.893
	SENTLI +ContractNLI	v	.813	**.978**	**.895**
0-shot	SENTLI -M.T	v	.616	.882	.749
	SENTLI -F.V	v	.616	.869	.742
	SENTLI	v	**.661**	**.904**	**.782**

Table 4: ContractNLI results with Oracle evidence spans (excluding neutral examples). Sentence-pair models (S.P.), even in zero-shot setting, outperform the SpanNLI model that was trained on long inputs. ‡Results from Koreeda and Manning (2021).

Model	S.P.	P@R.8	F_1(C)	F_1(E)	AVG
Majority vote‡		-	.083	.428	.256
supervised:					
SpanNLI-base‡		.663	.287	.765	.526
SpanNLI-large‡		**.793**	.357	**.834**	.595
T5-large	v	.575	.512	.691	.601
SENTLI +Cont.NLI	v	.580	.521	.754	.637
+ Rerank$_{(K=1)}$	v	"	**.537**	.741	**.639**
+ Rerank$_{(K=5)}$	v	"	.520	.719	.619
zero-shot:					
TF-IDF‡	v	.057	-	-	-
BERT emb.	v	.083	-	-	-
SentenceT5	v	.311	-	-	-
SENTLI -M.T	v	.397	.261	.551	.406
SENTLI -F.V	v	.397	.247	.594	.420
SENTLI	v	**.412**	.257	.573	.415
+ Rerank$_{(K=1)}$	v	"	.363	**.659**	.511
+ Rerank$_{(K=5)}$	v	"	**.404**	.652	**.528**

Table 5: Evidence retrieval and classification results of both supervised and zero-shot models on ContractNLI test set. Within sentence-pair (S.P.) models, NLI-based retrieval is more precise than similarity retrievers. Supervised S.P. models outperform the joint SpanNLI model in the final classification task thanks to better F_1(C). ‡Results from Koreeda and Manning (2021).

sentence-pair models outperform the supervised SpanNLI models. SENTLI, trained on all five NLI datasets performs the best, demonstrating strong NLI capabilities on this new domain.

5.2 NLI vs. similarity-based retrieval

We use the span-level annotations of ContractNLI to evaluate span retrieval over the premise for NLI. We find SENTLI's NLI scores to provide a very strong retrieval signal, ranking one of the annotated spans at the top 61% of the time. A random baseline, in comparison, achieves less than 1%.

To compare with similarity-based retrievers, we adopt the TF-IDF baseline from Koreeda and Manning (2021), and also extract unsupervised sentence embeddings with BERT-large (Devlin et al., 2019) and SentenceT5-large (Ni et al., 2021).

Table 5 shows the precision of each retriever at recall 0.8 (P@R.8). Unsupervised NLI-based retrieval outperforms similarity-based retrievers. As discussed in §3.2.1, we conjecture that this is because some sentences in the document could be very similar to the hypothesis, but neutral towards it. Table C.1 shows an example of such case. Within the supervised methods, SpanNLI performs better, perhaps due to only including single random spans for neutral cases in the sentence-pair training data, not utilizing the full document.

5.3 Retrieve-and-classify

DocNLI. We test if our sentence-pair method can improve the low performance over long premises in DocNLI. We focus on examples with a short hypothesis (up to 20 tokens) and long premise (more than 512 tokens). Naively applying SENTLI on these cases without sentence splitting leads to a low 0.21 F_1 score. Using our retrieve-and-predict approach, the score increases up to 0.41. Reranking doesn't seem to improve in this case, possibly due to the dataset's two-way classification format.

ContractNLI. Table 5 shows the main results on ContractNLI comparing sentence-pair zero-shot models with both kinds of supervised models. Surprisingly, the zero-shot sentence-pair models are competitive with the supervised SpanNLI model, even outperforming them in F_1(C). The supervised SENTLI, trained also with ContractNLI, performs best overall. Even though its retrieval performance is still behind SpanNLI (§5.2), its final verdict on the hypothesis is better.

Reranking significantly improves the performance of zero-shot SENTLI, increasing the average score by up to 27%. We observe better performance with larger context ($K = 5$). For the supervised SENTLI model, reranking provides only marginal gains, and increasing the context is not beneficial.

Overall, we see that sentence-pair models obtain very strong classification performance, even in a zero-shot setting (Table 4), while also providing descent retrieval capabilities. We hypothesize that the gap in the retrieval performance could be due to randomly sampling neutral spans instead of utilizing the full document. Augmenting the training set with more and better neutral examples could further close this gap.

SUMMAC. For evaluating on SUMMAC, we first adopt the zero-shot method of Laban et al. (2022)

(SUMMAC$_{ZS}$) and only replace the backbone NLI model with SENTLI. This improves the performance by absolute 1.2 points (see Table D.1). This overall improvement comes despite a drop of almost 10 points on the Polytope dataset. This could be due to SUMMAC's labeling function that treats summaries with any added or omitted content, compared to the reference, as "factual inconsistent". These errors relate more to the summary quality rather than its correctness and therefore, we do not expect zero-shot NLI models to catch them.

Reranking (K=1) is effective for most datasets. The best overall performance is achieved by reranking and hard aggregating the hypothesis sentences, improving over soft aggregation in 4 out of the 6 datasets, and allowing an overall 0.2 points gain.

5.4 Factual discrepancies in clusters

Table 6 reports the results on our **Corrupted Wiki Clusters** dataset (§4.1). In addition to our main method, we tried to reverse rank the spans by entailment score, but find it to perform even worst than random. We find this to be caused by pairings that support unmodified facts in the corrupted sentence. When ranking by contradiction, SENTLI performs the best and successfully flags 68% of the corruptions as its top prediction.

Exploring popular **Real** Wikipedia articles, without any simulated edits our known discrepancies, our method quickly identified existing inconsistencies. We attach examples in Appendix F and discuss them briefly here. As depicted in Figure 1, we find the French Wikipedia to disagree with other versions on the nationality of Bram Stoker. SENTLI ranked this sentence highest among the 53 sentences of that cluster. Investigating the page's history, this claim was introduced by an edit[9] in Jan. 2017 and remained unchanged for over 5 years.

Interestingly, when looking for consensus and ranking sentences by agreement, SENTLI returns the English version that avoids stating any nationality: *"Dracula is a novel by Bram Stoker, published in 1897."* Intuitively, shorter statements have higher chance of obtaining consensus.

In another example, we look at the articles about "Big Ben". The top discrepancy prediction was a sentence from the Chinese version that was likely mistranslated due to multiple segmentation options. While this does not necessarily reveal a mistake in the original document, it shows the potential of this approach for flagging translation mistakes when related sources are available. The second ranked sentence identifies a statement from the Swedish page regarding the monument's official name. Upon manual verification, even though articles discuss several names that were changed over time, none seem to directly support that claim.

[9] fr.wikipedia.org/w/index.php?title=Dracula

Model	Accuracy@K		
	K=1	5	10
Random	17.11	46.53	73.68
Reversed ranking by entailment:			
SENTLI	1.46	17.35	40.29
Ranking by contradiction:			
SENTLI $_{-M.T}$	35.32	67.60	83.37
SENTLI $_{-F.V}$	38.59	70.39	83.13
SENTLI	**68.20**	**89.08**	**95.15**

Table 6: Wikipedia corruption detection results by different ranking methods/ models. Accuracy of including the corrupted sentence within the top K predictions.

6 Related Work

As mentioned in the introduction, the NLI task (Dagan et al., 2006, 2013), sometimes called Recognizing Textual Entailment (RTE), was extensively studied by the NLP community over the past several years as a semantic reasoning benchmark (see Poliak, 2020; Storks et al., 2019, for surveys). The field of fact verification (Vlachos and Riedel, 2014) also recently gained increased attention (Bekoulis et al., 2021; Kotonya and Toni, 2020; Guo et al., 2022; Zeng et al., 2021), sharing similar pair-wise semantic inference challenges, together with evidence retrieval. While both tasks were found to be vulnerable to idiosyncrasies (Gururangan et al., 2018; McCoy et al., 2019; Poliak et al., 2018; Schuster et al., 2019), methods and datasets for reducing the bias were proposed (Belinkov et al., 2019; Karimi Mahabadi et al., 2020; Shah et al., 2020; Utama et al., 2020, 2021; Wu et al., 2022).

Recently, NLI-style models were expanded for concrete purposes beyond benchmarking. For example, showing promising potential in verifying the factual correctness of dialog (Gupta et al., 2021; Honovich et al., 2021), summarization (Chen et al., 2021b; Eyal et al., 2019; Fabbri et al., 2021b; Laban et al., 2022), and QA (Bulian et al., 2022; Chen et al., 2021a; Mishra et al., 2021) systems. The NLI format was also found helpful for general self-training (Vu et al., 2021). Here, we focus on real-world direct applications such as automatic con-

tract analysis (Koreeda and Manning, 2021) and identifying discrepancies in document collections.

In parallel work, Utama et al. (2022) improve NLI models for evaluating summaries by generating in-domain data with automatic perturbations to simulate contradictions. We observe similar improvements when training the backbone NLI models on in-domain data for ContractNLI (supervised vs. zero-shot setting). Any additional improvements to the NLI model, aimed towards the target domain, are likely to further improve the reasoning capabilities of the retrieve-and-classify pipeline.

7 Conclusion

We present a comprehensive study on the performance of sentence-pair NLI models in real world applications that often involve both a shift in domain and long texts. Our findings indicate the readiness of these models to provide meaningful signal on the semantic relation between texts that can be easily aggregated towards practical gains. To demonstrate this, we also defined a new zero-shot entailment-focused task over clusters of related documents. Our multi-task sentence-pair NLI model (SENTLI) successfully flags spans that stand out due to their claims.

Ultimately, this study should help practitioners interested in applying NLI-style inference in real-world applications to design the best model for their target task. Our results suggest that if the hypothesis is short and the premise fully fits in the input limit of the model, a regular cross-attention classifier is likely to perform well in terms of accuracy as it is able to contextualize sentences in the premise. However, if we want to interpret the prediction by identifying the exact piece from the premise that led to the predicted conclusion, breaking the premise into segments could be useful. Furthermore, if the premise is too long to fit in the model's receptive field, then breaking the premise into segments and aggregating with our proposed techniques (retrieve-and-classify and reranking) is beneficial. Finally, when we don't have a well defined hypothesis, one might still want to automatically reason over a pair or collection of documents and identify any statements that stand out. In this case, our reasoning over clusters methodology shows how to use strong sentence-pair classifiers to obtain useful signals that are then aggregated to highlight specific claims.

It's important to note that when designing the methods for this work, we preferred simplicity over performance in order to directly study the quality of the SENTLI's scores. Yet, we achieve high zero-shot performance and even reach state-of-the-art on ContractNLI. We hope that this work will motivate future research on further expanding these methods, for example by decontextualizing the premise, supporting multi-hop reasoning, expanding the context with sliding windows instead of sentence splitting, or hypothesis fragmentation.

Acknowledgements

We thank Shashi Narayan and Simon Baumgartner for valuable feedback on the writing. We also thank Sumit Sanghai, Annie Louis, Jiaming Luo, Roee Aharoni, Yi Tay, Kai Hui, Jai Gupta, Vinh Tran, and Dara Bahri for helpful conversations and feedback.

Limitations

As mentioned in the conclusion section and along the paper, our experiments focus on exploring and leveraging the direct signal from sentence-pair NLI models. In this work, we did not employ more advanced techniques to process the data such as contextualizing the premise or fragmenting the hypothesis. We leave such studies on further improving the downstream performance to future work. Also, we train and evaluate our NLI models on English inputs, and don't explore morphologically richer languages here.

In our Wikipedia clusters experiments, we translate all pages to English. This translation process might introduce some mistakes. However, when examining several samples we find the translation quality to be overall high. Also, capturing translation mistakes with SENTLI is another potentially interesting application of our setup.

Finally, our zero-shot evaluations on DocNLI and ContractNLI are out-of-domain, but the Wikipedia Cluster experiments are partly in-domain as the training data includes premises from Wikipedia. Yet, different from the training, the hypothesis is also a Wikipedia sentence. Also, we use the VitaminC test set to avoid potential overlaps with pages that the model saw on training.

Ethical Considerations

We emphasize that our method and experimentation on identifying disagreements between documents is focused on the research question of whether our models can capture the required signal from the

text. When highlighting such cases, we do not claim by any means to state anything regarding the truthfulness of any of the statements. Rather, we examine the question regarding the usefulness of ranking the statements by their perceived agreement with each other.

References

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2022. Ext5: Towards extreme multi-task scaling for transfer learning. In *International Conference on Learning Representations*.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *ACM Comput. Surv.*, 55(1).

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. On adversarial removal of hypothesis-only bias in natural language inference. *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

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.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Boerschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation.

Jifan Chen, Eunsol Choi, and Greg Durrett. 2021a. Can NLI models verify QA systems' predictions? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3841–3854, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021b. 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.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

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.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. 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 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021a. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021b. Qafacteval: Improved qa-based factual consistency evaluation for summarization.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Zhijiang Guo, M. Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Dialfact: A benchmark for fact-checking in dialogue.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Robert L. Logan IV, Alexandre Passos, Sameer Singh, and Ming-Wei Chang. 2021. Fruit: Faithfully reflecting updated information in text.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. Looking beyond sentence-level natural language inference for question answering and text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1336, Online. Association for Computational Linguistics.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of*

the First Workshop on Evaluation and Comparison of NLP Systems, pages 92–109, Online. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

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(140):1–67.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling up models and data with `t5x` and `seqio`.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. 2022. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning.

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.

Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Darsh Shah, Tal Schuster, and Regina Barzilay. 2020. Automatic fact-guided sentence modification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8791–8798.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Prasetya Utama, Joshua Bambrick, Nafise Moosavi, and Iryna Gurevych. 2022. Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2763–2776, Seattle, United States. Association for Computational Linguistics.

Prasetya Utama, Nafise Sadat Moosavi, Victor Sanh, and Iryna Gurevych. 2021. Avoiding inference heuristics in few-shot prompt-based finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9063–9074, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Denny Vrandecic. 2020. Architecture for a multilingual wikipedia. *CoRR*, abs/2004.04733.

Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. 2021. STraTA: Self-training with task augmentation for better few-shot learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5715–5731, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets.

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.

Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.

A Implementation details for SENTLI

We use the T5X framework (Roberts et al., 2022) to finetune the T5.1.1-Large model for 500K steps and pick the best performning checkpoint on the evaluation splits of the source datasets. We use a batch size of 128 with a balanced sampling across the training datasets to account for their different sizes.

B Zero-shot binary classification

DocNLI uses only two classes, "entail" vs. "not entail", by merging the "neutral" and "contradiction" definitions. The "not entail" instances are created with both rule-based and LM-based local perturbations over the positive pairs. This results in a slightly different task definition than NLI since even altered (i.e., "fake") texts can still factually agree with the hypothesis (Schuster et al., 2020). Yin et al. (2021) account for this by augmenting the training set, but zero-shot NLI models might be affected by this provenance-based rather than factual-based partition of the test set.

Since NLI models are commonly trained with three target labels, adjusting to a two-way classification requires some modifications. Being a zero-shot setting, we cannot train the model's internal representations to adjust to this new label space. Instead, we can define an aggregation method. We experiment with the following variants:

1. **Entailment threshold**: predicting "entail" if $p_e > T$, else "not entail".

2. **Contradiction threshold**: predicting "not entail" if $p_c > T$, else "entail".

3. **Binary softmax**: recompute the softmax without s_n, and predict "entail" if $\text{Softmax}(s_e, s_c)[e] > T$, else "not entail".

T is a decision threshold that we can either set to some arbitrary value such as 0.5, or calibrate it on a small set of labeled data.

Table B.1 presents the performance of the three aggregation methods with or without tuning T on 500 random examples for the development set (with 0.05 intervals). Thresholding on p_e performs best and gives higher precision compared to the binary softmax that discards the 'neutral' score. Yet, all three options perform well, with different trade-offs between precision and recall, motivating the use of simple heuristics in the absence of supervised data for the target task. In the following experiments, we use the 'e' threshold method. Tuning T significantly improves the F_1 score by sacrificing recall for precision. We find the optimal threshold to be 0.95, meaning that we predict 'entail' only when the model is highly confident in this relation.

C Example of NLI vs. similarity-based retrieval

Table C.1 shows a retrieval example from the ContractNLI dataset, using either SentenceT5 (similarity-based) or SENTLI (NLI-based) to retrieve spans that might support or refute the candidate hypothesis.

D SUMMAC Evaluation

The SUMMAC benchmark includes six datasets: CGS (Falke et al., 2019), XSF (Maynez et al., 2020), Polytope (Huang et al., 2020), FactCC (Kryscinski et al., 2020), SummEval (Fabbri et al., 2021a), Frank (Pagnoni et al., 2021).

Table D.1 reports the results on this benchmark. See §5.3 for discussion.

E Additional Details on the Wiki Clusters Dataset

As described in Section 4.1, we created a Wikipedia-based dataset with clusters of similar articles written by different communities in multiple languages. Below, we provide additional details on the process.

We first collect the 5000 most popular accessed pages in the English Wikipedia, according to the ranking of November 2021.[10] We take a cleaned version (without links) of the introduction of each page (the text coming before the content table). Then, to create the cluster for each article, we use each article's language links and collect similar introductions from the 10 (non-EN) languages that with most admins as of November 2021.[11] Finally, we use Google Translate API to translate all articles to English.

We manually examine a random subset of the clusters and find very few translation mistakes and that in most clusters there are at least several non-English introductions with sufficient length and content. This is mostly thanks to our choice of

[10] https://bit.ly/Wiki_popular_pages

[11] By https://en.wikipedia.org/wiki/List_of_Wikipedias#Edition_details: DE, FR, IT, PL, RU, SV, ZH, ES, PT, UK.

Aggregation	T	0.2% of DocNLI Dev. Prec.	Recall	F_1(E)	DocNLI Dev. Prec.	Recall	F_1(E)	DocNLI Test Prec.	Recall	F_1(E)
'e' threshold	0.5	.198	.746	.313	.216	.810	.341	.220	.800	.345
	tuned	.319	.610	**.419**	.312	.594	**.409**	.313	.589	**.408**
'c' threshold	0.5	.187	.966	.313	.182	.969	.306	.182	.969	.306
	tuned	.244	.915	.386	.230	.872	.364	.230	.872	.364
bin. softmax	0.5	.195	.966	.324	.192	.949	.319	.193	.949	.321
	tuned	.273	.847	.413	.263	.806	.397	.263	.807	.397

Table B.1: Different aggregation methods for converting predictions of a three-way NLI model to a binary label space. Aggregating by the score of the entailment class performs best. Tuning the threshold on random 500 examples (0.2% of full Dev.) further improves the precision and F_1 scores, compared to the naive 0.5 baseline.

Hypothesis:	Confidential Information shall only include technical information. *(gold label = contradiction)*
ST5 top-1	5.1.2. use Confidential Information only for the Project;
top-2	5.3.2. The disclosure of Confidential Information to Recipient or its Representatives shall not give Recipient or its Representatives any licence or other rights in relation to that Confidential Information [...]
SENTLI argmax p_c	3.5. "Confidential information" means any information of whatever form relating to the Project or Discloser or any of its Affiliates or Clients, supplied or made available by Discloser or on its behalf to recipient [...]
argmax p_e	You the subject-matter expert

Table C.1: ContractNLI evidence retrieval example. The top two retrievals of the SentenceT5 (ST5) model relate to the hypothesis (discussing confidential information), but are do not refute or support it. Alternatively, retrieving by NLI scores is highly informative as the sentence with max p_c clearly contradicts the hypothesis.

popular pages and languages. We also tried to randomly sample English articles but find many of the versions in other languages to be missing or have a single sentence.

As mentioned in Section 4.1, we designed a controlled experiment where we simulate local corruptions to the English version, and use the information from the other versions to predict which sentence was changed. To make the corruptions realistic and not obvious, we rely on the Wikipedia edits collection from the VitaminC test set (Schuster et al., 2021). These edits include both real revisions from Wikipedia history logs, and synthetic edits created by annotators to modify certain facts. To ensure that we only use edits that present a factual modification, we only take pairs of evidence from the same page that are paired with the same claim, but express an opposing relation (one supports it and the other refutes it).

To match the edits from VitaminC with our current version, we use word-level (also splitting hyphens) Jaccard similarity. First, looking from small edits, we only take the VitaminC edit instances with similarity greater than 0.25 between the "before" (x_b) and "after" (x_a) sentences. Then, for each of the edits, we look for the sentence x from the current article that has the highest Jaccard similarity with either the x_b or x_a sentence. If no sentence has greater than 0.2 similarity, we skip this edit. Finally, to decide which of the two sentences is coherent with the current version, we pick the one with the higher Jaccard similarity with x. Accordingly, we assume that the other sentence represent a factual modification to the current article, and therefore create a local discrepancy by replacing it with x.

Following this process, we obtain a total of 824 local corruptions to 144 different articles. See Table E.1 for an example.

F Examples of discrepancies and consensus in Wikipedia

We present the retrievals of our method for identifying discrepancies and consensus in document clusters (§4.1) when applied on the "Dracula" (Tables F.1-F.2), "Big Ben" (Tables F.3-F.4), and "Cameron Boyce" (Tables F.5-F.6) pages. We observe that in general shorter sentences with consensus tend to be short and concise. This is intuitive as longer sentence are more likely to include claims that are missing from other documents.

Method	CGS	XSF	poly	factC	sumEv	frank	AVG
SUMMAC$_{ZS}$	70.4	58.4	**62.0**	83.8	78.7	79.0	72.1
SENTLI (soft)	79.3	59.3	52.4	**89.5**	77.2	**82.1**	73.3
+ Rerank (soft)	79.6	62.7	52.8	86.1	78.5	80.4	73.3
+ Rerank (hard)	**80.5**	**64.2**	55.1	83.3	**79.7**	78.4	**73.5**

Table D.1: SUMMAC zero-shot balanced accuracy. The hypothesis aggregation method (§3.3) is in parenthesis.

Original	Mars is the site of Olympus Mons, the largest volcano and highest known mountain on any planet in the Solar System, and of Valles Marineris, one of the largest canyons in the Solar System.
Corruption	Mars is the site of Olympus Mons, the largest volcano and second-highest known mountain in the Solar System, but far away from Valles Marineris, one of the largest canyons in the Solar System.
Examples of related sentences from other articles with helpful information for identifying the corruption:	
FR	The highest volcano in the Solar System, Olympus Mons (which is a shield volcano), and the largest canyon, Valles Marineris, are found on Mars.
IT	Among the most noteworthy geological formations of Mars are: Olympus Mons, or Mount Olympus, the largest volcano in the solar system (27 km high); the Valles Marineris, a long canyon considerably larger than the terrestrial ones; and a huge crater on the northern hemisphere, about 40% wide of the entire Martian surface.
SV	During large parts of Mars' history, long-lasting volcanic eruptions occurred which, among other things, created Olympus Mons, the highest mountain in the solar system.

Table E.1: Example of a *simulated* corruption in the English Wikipedia about Mars from our corrupted articles dataset (§4.1). We are given the introduction of the English article, consisting of 35 sentences, where one sentence, the "original", was replaced with the "corruption" one. The goal is to successfully identify which sentence was corrupted by leveraging information from the other 10 related articles (each with 18 sentences on average). We only present the most relevant sentences here, but the model has to read through the whole cluster.

*Dracula: searching for **discrepancies**.*

FR	Dracula is an epistolary novel by British writer Bram Stoker published in 1897.
Top sentence from each document by disagreement with the candidate:	
EN	A small group, led by Abraham Van Helsing, hunt Dracula and, in the end, kill him.
DE	Dracula is a novel by Irish writer Bram Stoker published in 1897.
IT	Dracula is an epistolary novel written by Irish Bram Stoker in 1897, inspired by the figure of Vlad III, prince of Wallachia, and is one of the last examples of Gothic novels.
PL	Dracula - a 19th-century Gothic novel by the Irish writer Bram Stoker, depicting the fight of a group of volunteers with the vampire Dracula.
RU	Dracula is a novel by the Irish writer Bram Stoker, first published in 1897.
PT	Dracula (Dracula) is an 1897 gothic horror novel written by Irish author Bram Stoker, starring the vampire Count Dracula.
ES	Dracula is a novel published in 1897 by the Irishman Bram Stoker, as a result of which his antagonist character, Count Dracula, became the quintessential Western vampire archetype, becoming considered the most famous vampire.
ZH	"Dracula" is a gothic horror novel based on vampires published in 1897 by Irish writer Bram Stoker.
SV	Dracula is a horror novel from 1897 by the Irish author Bram Stoker, in which the main antagonist is the vampire Count Dracula.
UK	Dracula is a novel by Irish writer Bram Stoker, first published in 1897.

Table F.1: The sentence with highest discrepancy score (shown at the top) among all 53 sentences from "Dracula" articles. Beneath, we show the sentence from each Wikipedia version that had the highest disagreement score with the candidate. This example is also illustrated in Figure 1.

*Dracula: searching for **consensus**.*

EN	Dracula is a novel by Bram Stoker, published in 1897.
Top sentence from each document by agreement with the candidate:	
DE	Dracula is a novel by Irish writer Bram Stoker published in 1897.
FR	Dracula is an epistolary novel by British writer Bram Stoker published in 1897.
IT	Dracula is an epistolary novel written by Irish Bram Stoker in 1897, inspired by the figure of Vlad III, prince of Wallachia, and is one of the last examples of Gothic novels.
PL	Dracula - a 19th-century Gothic novel by the Irish writer Bram Stoker, depicting the fight of a group of volunteers with the vampire Dracula.
RU	Dracula is a novel by the Irish writer Bram Stoker, first published in 1897.
PT	Dracula (Dracula) is an 1897 gothic horror novel written by Irish author Bram Stoker, starring the vampire Count Dracula.
ES	Dracula is a novel published in 1897 by the Irishman Bram Stoker, as a result of which his antagonist character, Count Dracula, became the quintessential Western vampire archetype, becoming considered the most famous vampire.
ZH	"Dracula" is a gothic horror novel based on vampires published in 1897 by Irish writer Bram Stoker.
SV	Dracula is a horror novel from 1897 by the Irish author Bram Stoker, in which the main antagonist is the vampire Count Dracula.
UK	Dracula is a novel by Irish writer Bram Stoker, first published in 1897.

Table F.2: The sentence with the most consensus (shown at the top) among all 53 sentences from "Dracula" articles. Beneath, we show the sentence from each Wikipedia version that had the highest agreement score with the candidate.

*Big Ben: searching for **discrepancies**.*

SV	Big Ben is officially called the Great Bell of Westminster and strikes every hour in the tower clock with the official name Great Clock of Westminster.
Top sentence from each document by disagreement with the candidate:	
EN	The official name of the tower in which Big Ben is located was originally the Clock Tower, but it was renamed Elizabeth Tower in 2012, to mark the Diamond Jubilee of Elizabeth II, Queen of the United Kingdom.
DE	The tower has been officially called Elizabeth Tower since September 2012.
FR	Previously, it was simply called the Clock Tower.
IT	This bell tower rings every quarter of an hour.
PL	On September 12, 2012, the tower was officially named Elizabeth Tower in honor of Elizabeth II's 60-year reign.
RU	The official name of the tower since 2012 is the Elizabeth Tower, one of the most recognizable symbols of Great Britain, often used in souvenirs, advertisements, and movies.
PT	The official name of the tower in which Big Ben is located was originally Clock Tower, but it was renamed Elizabeth Tower in 2012 to mark Queen Elizabeth II's Diamond Jubilee.
ES	Its official name was Clock Tower, until on June 26, 2012, in honor of Queen Elizabeth II's Diamond Jubilee, it was decided that the tower would be renamed Elizabeth Tower.
ZH	Big Ben (English: Big Ben, or translated as Big Ben) is a big newspaper clock located at the north end of the Palace of Westminster in London.
UK	The official name of the tower since 2012 - Elizabeth Tower (English Elizabeth Tower).

Table F.3: The sentence with second highest discrepancy score (shown at the top) among all 72 sentences from "Big Ben" articles. Beneath, we show the sentence from each Wikipedia version that had the highest disagreement score with the candidate.

*Big Ben: searching for **consensus**.*

UK	The official name of the tower since 2012 - Elizabeth Tower (English Elizabeth Tower).
Top sentence from each document by agreement with the candidate:	
EN	The official name of the tower in which Big Ben is located was originally the Clock Tower, but it was renamed Elizabeth Tower in 2012, to mark the Diamond Jubilee of Elizabeth II, Queen of the United Kingdom.
DE	The tower has been officially called Elizabeth Tower since September 2012.
FR	The tower was renamed on the occasion of the Diamond Jubilee of Elizabeth II in 2012.
IT	Known as the Clock Tower, the name was officially changed to Elizabeth Tower on the occasion of Elizabeth II's Diamond Jubilee in June 2012.
PL	On September 12, 2012, the tower was officially named Elizabeth Tower in honor of Elizabeth II's 60-year reign.
RU	The official name of the tower since 2012 is the Elizabeth Tower, one of the most recognizable symbols of Great Britain, often used in souvenirs, advertisements, and movies.
PT	The official name of the tower in which Big Ben is located was originally Clock Tower, but it was renamed Elizabeth Tower in 2012 to mark Queen Elizabeth II's Diamond Jubilee.
ES	Its official name was Clock Tower, until on June 26, 2012, in honor of Queen Elizabeth II's Diamond Jubilee, it was decided that the tower would be renamed Elizabeth Tower.
ZH	Big Ben (English: Big Ben, or translated as Big Ben) is a big newspaper clock located at the north end of the Palace of Westminster in London.
SV	The tower, which has been called Elizabeth Tower since 2012 and where the clock hangs, is 96.3 meters high.

Table F.4: The sentence with the most consensus (shown at the top) among all 72 sentences from "Big Ben" articles. Beneath, we show the sentence from each Wikipedia version that had the highest agreement score with the candidate.

*Cameron Boyce: searching for **discrepancies**.*

UK	He died of an epileptic seizure on the night of July 7, 2019 at 2:35 p.m.
Top sentence from each document by disagreement with the candidate:	
ZH	On July 6, 2019, Boyce died of epileptic seizures at the age of 20.
RU	Cameron Mica Boyce (born May 28, 1999 - July 6, 2019) - American actor and dancer, best known for his leading roles in the comedy series Jesse (2011-2015) and Gamer's Diary (2015-2017) , as well as in the series of films "Descendants" (2015-2019).
EN	Cameron Mica Boyce (May 28, 1999 – July 6, 2019) was an American actor.
PL	Cameron Mica "Cam" Boyce (born May 28, 1999 in Los Angeles, died July 6, 2019 therein) - American actor, dancer and model.
FR	Cameron Boyce is an American actor, dancer, singer and model, born May 28, 1999 in Los Angeles (California) and died July 6, 2019 in the same city.
ES	Cameron Mica Boyce (Los Angeles, California; May 28, 1999-July 6, 2019) was an American actor known primarily for his roles in the feature films Descendants, Descendants 2, Descendants 3, as well as for his role of Luke Ross on the Disney Channel series Jessie.
PT	Cameron Boyce (Los Angeles, May 28, 1999 – Los Angeles, July 6, 2019) was an American actor, singer, dancer and voice actor, known for starring in films such as Mirrors and appearing in Eagle Eye, both of 2008.
SV	Cameron Boyce, born May 28, 1999 in Los Angeles, died July 6, 2019 in Los Angeles, was an American actor.
DE	Cameron Boyce (born May 28, 1999 in Los Angeles, California - July 6, 2019) was an American actor and child actor.
IT	Cameron Mica Boyce (Los Angeles, May 28, 1999 - Los Angeles, July 6, 2019) was an American actor and dancer.

Table F.5: The sentence with highest discrepancy score (shown at the top) among all 38 sentences from "Cameron Boyce" articles. Beneath, we show the sentence from each Wikipedia version that had the highest disagreement score with the candidate.

Cameron Boyce: searching for ***consensus****.*

EN	Cameron Mica Boyce (May 28, 1999 – July 6, 2019) was an American actor.
Top sentence from each document by agreement with the candidate:	
PL	Cameron Mica "Cam" Boyce (born May 28, 1999 in Los Angeles, died July 6, 2019 therein) - American actor, dancer and model.
RU	Cameron Mica Boyce (born May 28, 1999 - July 6, 2019) - American actor and dancer, best known for his leading roles in the comedy series Jesse (2011-2015) and Gamer's Diary (2015-2017) , as well as in the series of films "Descendants" (2015-2019).
ES	Cameron Mica Boyce (Los Angeles, California; May 28, 1999-July 6, 2019) was an American actor known primarily for his roles in the feature films Descendants, Descendants 2, Descendants 3, as well as for his role of Luke Ross on the Disney Channel series Jessie.
IT	Cameron Mica Boyce (Los Angeles, May 28, 1999 - Los Angeles, July 6, 2019) was an American actor and dancer.
ZH	Cameron Mica Boyce (English: Cameron Mica Boyce, May 28, 1999-July 6, 2019) was an American actor and dancer.
PT	Cameron is best known for playing Luke Ross in Jessie, a series that aired on the Disney Channel, and for his role in the film Descendants, having also played the role of Conor in Gamer's Guide to Pretty Much Everything.
SV	Boyce was known for his roles in films such as Mirrors, Eagle Eye, Grown Ups and Grown Ups 2.
FR	Cameron Boyce is an American actor, dancer, singer and model, born May 28, 1999 in Los Angeles (California) and died July 6, 2019 in the same city.
DE	Cameron Boyce (born May 28, 1999 in Los Angeles, California - July 6, 2019) was an American actor and child actor.
UK	Cameron Boyce is an American actor and dancer best known for his roles in the feature films "Mirrors," "Hook," "Classmates," "Heirs," "Heirs 2" and the Disney series Jesse.

Table F.6: The sentence with the most consensus (shown at the top) among all 38 sentences from "Cameron Boyce" articles. Beneath, we show the sentence from each Wikipedia version that had the highest agreement score with the candidate.