SummaC:
Re-visiting NLI for Summary Consistency

Philippe Laban, Tobias Schnabel, Paul Bennett, Marti Hearst

TACL Paper - ACL 2022 Presentation

What is summary factual consistency?

… and why is it important?

Example Inconsistent Summary

Scientists are studying Mars to learn about the Red Planet and find landing sites for future missions.

One possible site, known as Arcadia Planitia, is covered in strange sinuous features.

The shapes could be signs that the area is actually made of glaciers, which are large masses of slowmoving ice.

Arcadia Planitia is in Mars' northern lowlands.

There are strange shape patterns on Arcadia Planitia.

The shapes could indicate the area might be made of glaciers.

This makes Arcadia Planitia ideal for future missions.

??

Document

Summary

Steps to Tackling Factual Inconsistency

1.

Collect data to measure problem

2.

Build models that can detect errors

3.

Build models that can fix / avoid errors

Steps to Tackling Factual Inconsistency

1.
Collect data to measure problem

2.
Build models that can detect errors

3.
Build models that can fix / avoid errors

Focus of this work

1. Datasets

Creating the SummaC Benchmark

The SummaC benchmark standardizes 6 large summary consistency dataset into a single benchmark.

Model Name	**Size** Valid	**Size** Test	**% Consistent**
CGS (Falke, 2019)	1,281	400	49.8

CGS: Consistency as ranking problem, with pairs of (consistent,inconsistent) sentences.

Creating the SummaC Benchmark

The SummaC benchmark standardizes 6 large summary consistency dataset into a single benchmark.

Model Name	**Size** Valid	**Size** Test	**% Consistent**
CGS (Falke, 2019)	1,281	400	49.8
XSF (Maynez 2020)	1,250	1,250	10.2

XSF: Differentiates between *extrinsic* and *intrinsic* factual errors.
Focus on the XSum dataset, which models struggle with more.

Creating the SummaC Benchmark

The SummaC benchmark standardizes 6 large summary consistency dataset into a single benchmark.

Model Name	Size Valid	Size Test	% Consistent
CGS (Falke, 2019)	1,281	400	49.8
XSF (Maynez 2020)	1,250	1,250	10.2
Polytope (Huang 2020)	634	634	6.6

Polytope: error typology based on MQM, with 5 accuracy errors subtypes.

Creating the SummaC Benchmark

The SummaC benchmark standardizes 6 large summary consistency dataset into a single benchmark.

Model Name	**Size** Valid	**Size** Test	**% Consistent**
CGS (Falke, 2019)	1,281	400	49.8
XSF (Maynez 2020)	1,250	1,250	10.2
Polytope (Huang 2020)	634	634	6.6
FactCC (Kryscinski 2020)	931	503	85.5

FactCC: Annotated by experts rather than crowd-workers. Also contains synthetic training dataset.

Creating the SummaC Benchmark

The SummaC benchmark standardizes 6 large summary consistency dataset into a single benchmark.

Model Name	**Size** Valid	**Size** Test	**% Consistent**
CGS (Falke, 2019)	1,281	400	49.8
XSF (Maynez 2020)	1,250	1,250	10.2
Polytope (Huang 2020)	634	634	6.6
FactCC (Kryscinski 2020)	931	503	85.5
SummEval (Fabbri 2021)	850	850	90.6

<u>SummEval:</u> Extensive effort with annotations for 23 summarizers. Uses 5-pt Likert scale rather than binary tag.

Creating the SummaC Benchmark

The SummaC benchmark standardizes 6 large summary consistency dataset into a single benchmark.

Model Name	**Size** Valid	**Size** Test	**% Consistent**
CGS (Falke, 2019)	1,281	400	49.8
XSF (Maynez 2020)	1,250	1,250	10.2
Polytope (Huang 2020)	634	634	6.6
FactCC (Kryscinski 2020)	931	503	85.5
SummEval (Fabbri 2021)	850	850	90.6
FRANK (Pagnoni 2021)	671	1,575	33.2

FRANK: Introduces error-typology with 7 error types. Analysis both on CNN/DM and XSum.

Objectives of the Benchmark

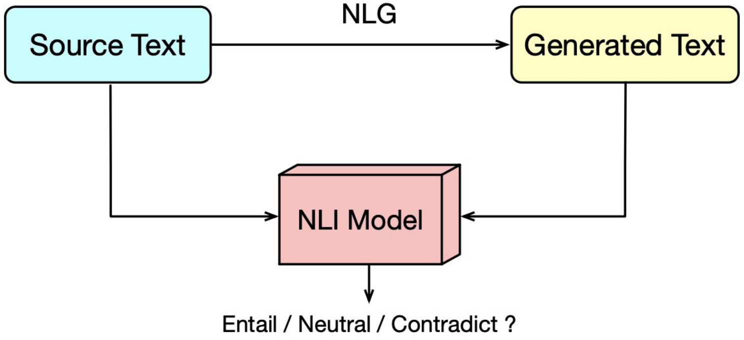
1. **Standardize the task.** By standardizing the task (binary classification) and the metrics (balanced accuracy).

2. **Broader evaluation.** By seeing which models generalize well across datasets / settings.

3. **Ease of access.** The benchmark is available for download publicly to accelerate research.

2. Models

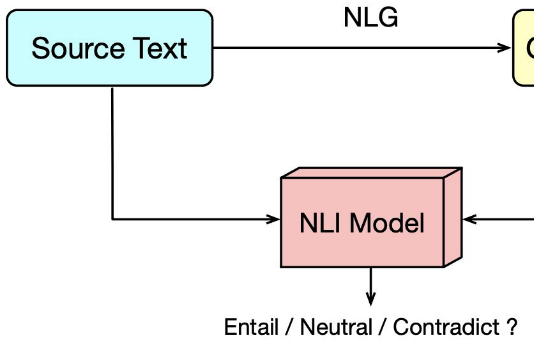
Methods for inconsistency detection

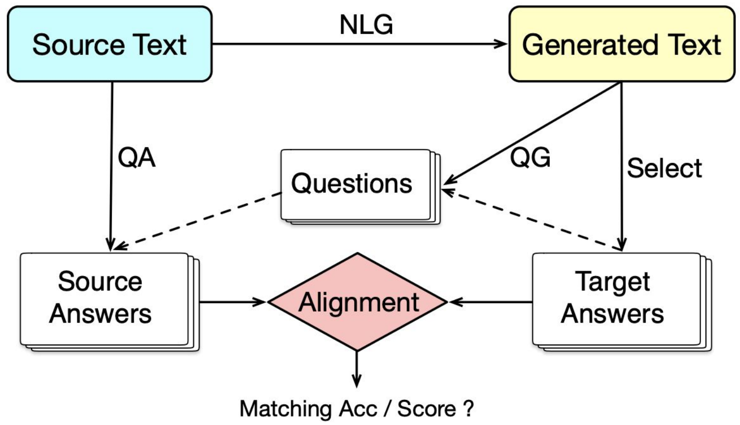


NLI-based

Diagrams from Li, Wei, et al. "Faithfulness in Natural Language Generation: A Systematic Survey of Analysis, Evaluation and Optimization Methods." arXiv preprint arXiv:2203.05227 (2022).

Methods for inconsistency detection





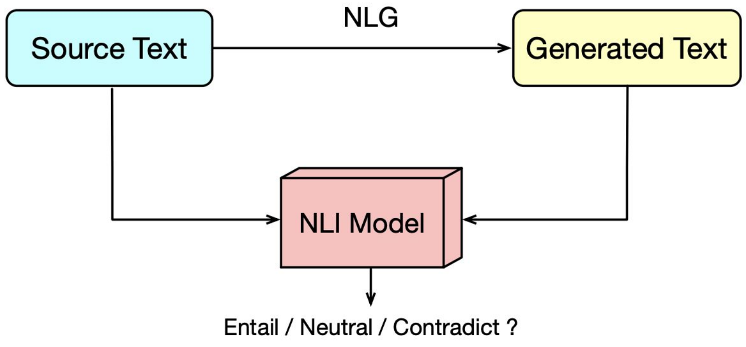
NLI-based

QAG-Based

Diagrams from Li, Wei, et al. "Faithfulness in Natural Language Generation: A Systematic Survey of Analysis, Evaluation and Optimization Methods." arXiv preprint arXiv:2203.05227 (2022).

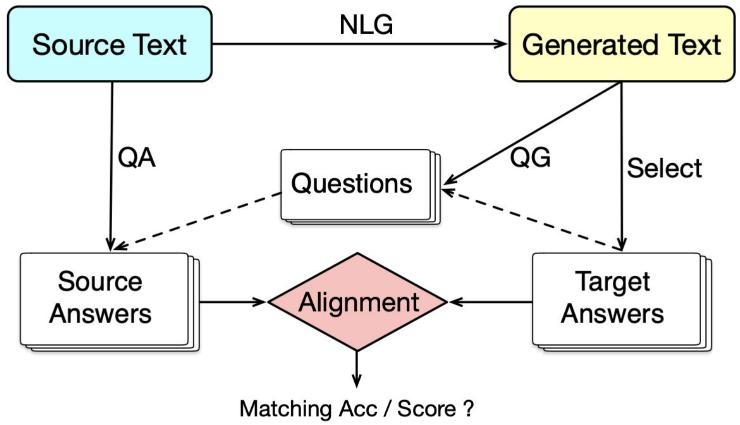
Methods for inconsistency detection

Older methods / Low accuracy



NLI-based

More recent / Higher accuracy

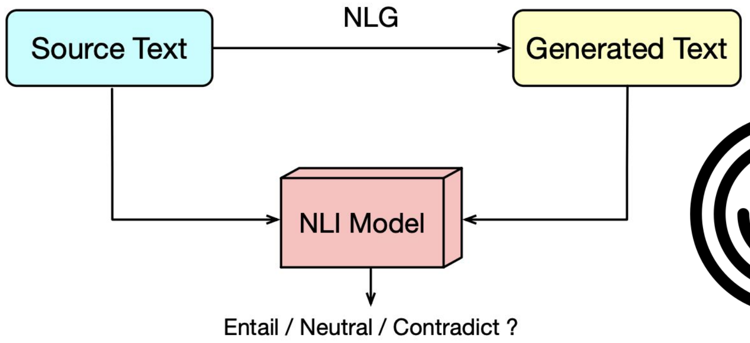


QAG-Based

Diagrams from Li, Wei, et al. "Faithfulness in Natural Language Generation: A Systematic Survey of Analysis, Evaluation and Optimization Methods." arXiv preprint arXiv:2203.05227 (2022).

Methods for inconsistency detection

Older methods / Low accuracy



Let's take a closer look at NLI model performance.

NLI-based

Diagrams from Li, Wei, et al. "Faithfulness in Natural Language Generation: A Systematic Survey of Analysis, Evaluation and Optimization Methods." arXiv preprint arXiv:2203.05227 (2022).

Zooming in on NLI models

NLI (

Scientists are studying Mars to learn about the Red Planet and find landing sites for future missions.

One possible site, known as Arcadia Planitia, is covered in strange sinuous features.

The shapes could be signs that the area is actually made of glaciers, which are large masses of slowmoving ice.

Arcadia Planitia is in Mars' northern lowlands.

,

There are strange shape patterns on Arcadia Planitia.

The shapes could indicate the area might be made of glaciers.

This makes Arcadia Planitia ideal for future missions.

)=0.91

Document **Summary**

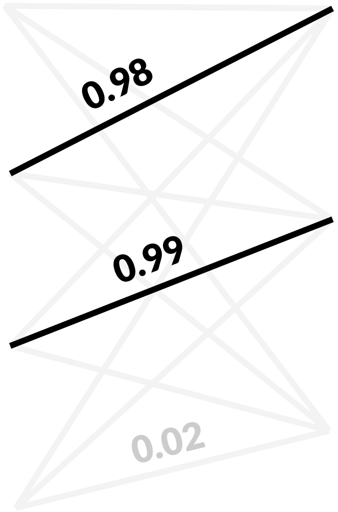
Zooming in on NLI models

Scientists are studying Mars to learn about the Red Planet and find landing sites for future missions.

One possible site, known as Arcadia Planitia, is covered in strange sinuous features.

The shapes could be signs that the area is actually made of glaciers, which are large masses of slowmoving ice.

Arcadia Planitia is in Mars' northern lowlands.



There are strange shape patterns on Arcadia Planitia.

The shapes could indicate the area might be made of glaciers.

This makes Arcadia Planitia ideal for future missions.

NLI (Di, Sj)

Document **Summary**

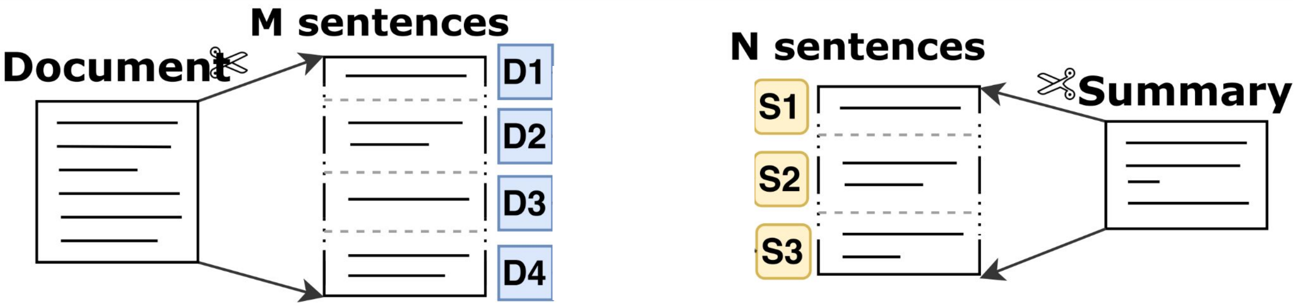
Mismatch in granularity between

NLI datasets
sentence-level

Consistency detection
document-level

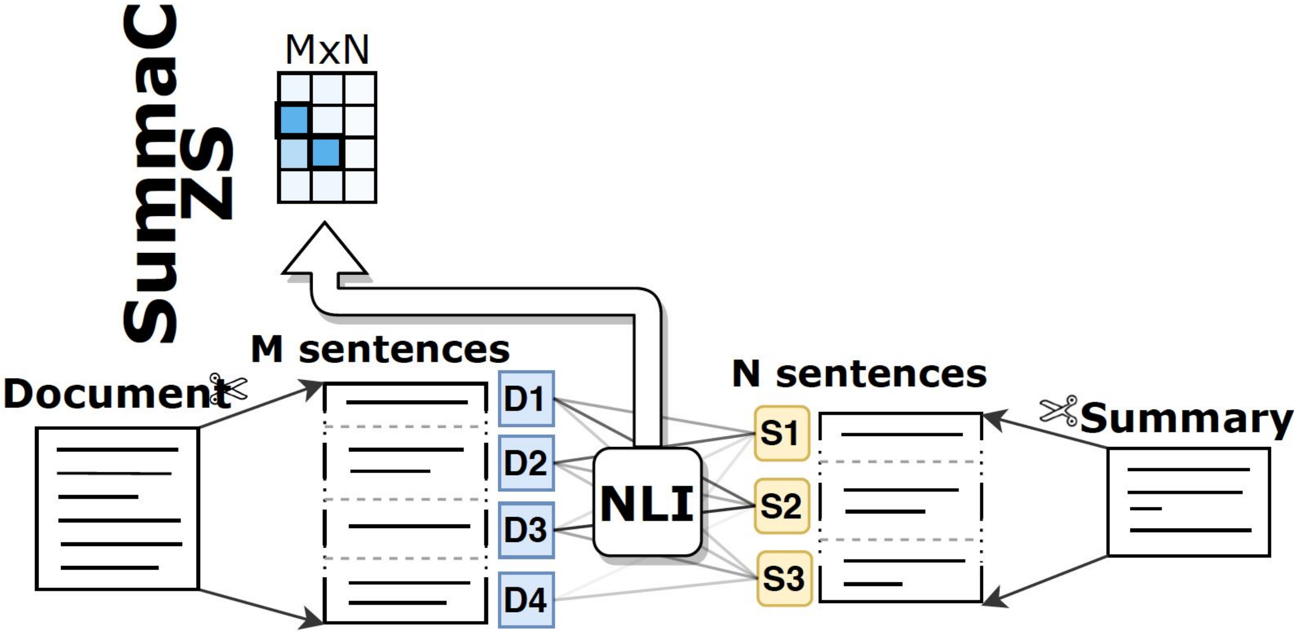
How do we adjust the granularity?

SummaC Zero Shot



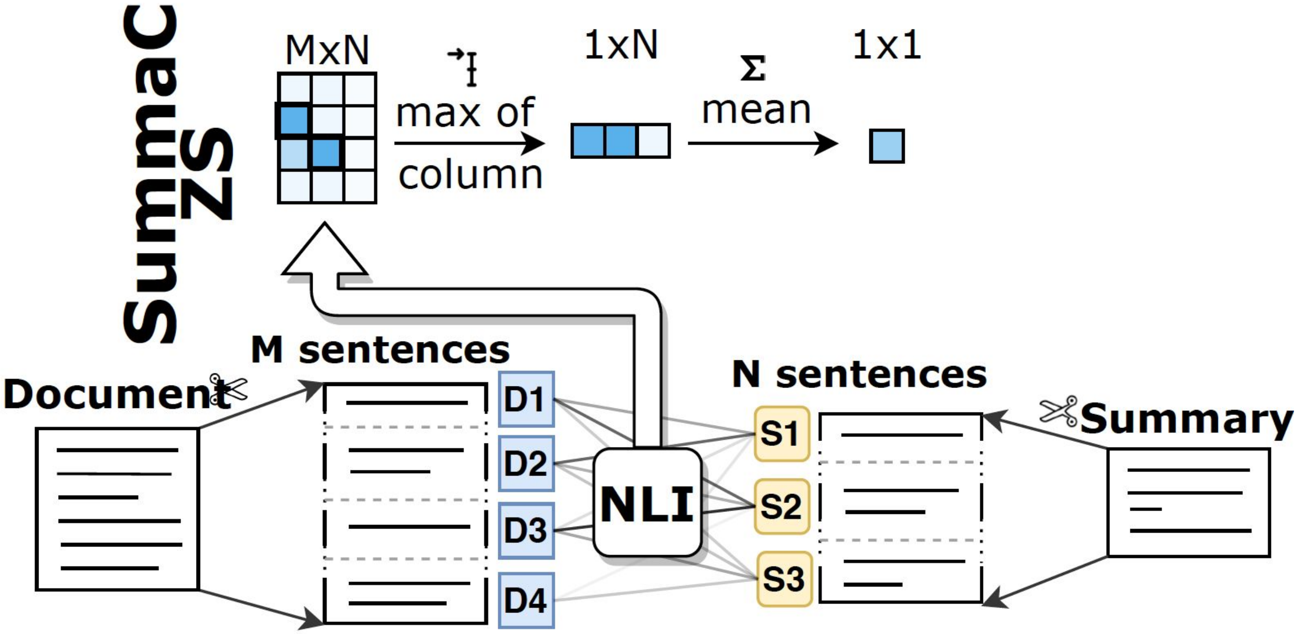
1. Split the document and summary into blocks (sentences, paragraphs, etc.)

SummaC Zero Shot



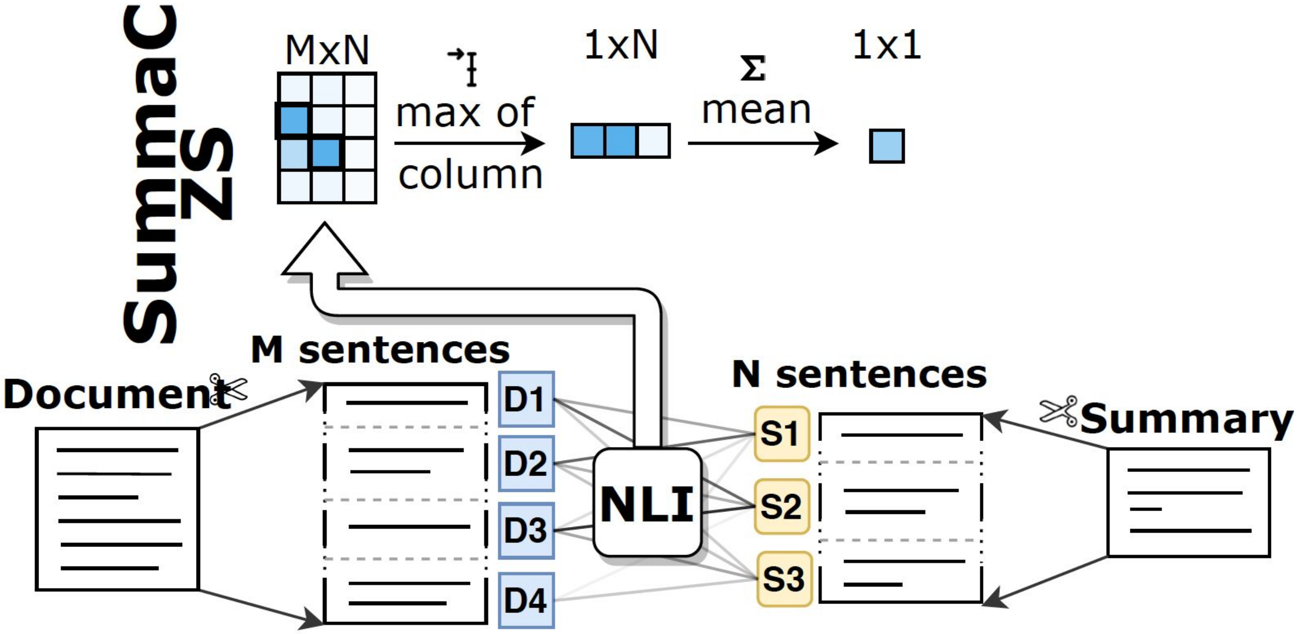
2. Run each (doc, summ) block through NLI model. Form an MxN matrix.

SummaC Zero Shot



3. Take max for each column, and mean over the rows.

SummaC Zero Shot



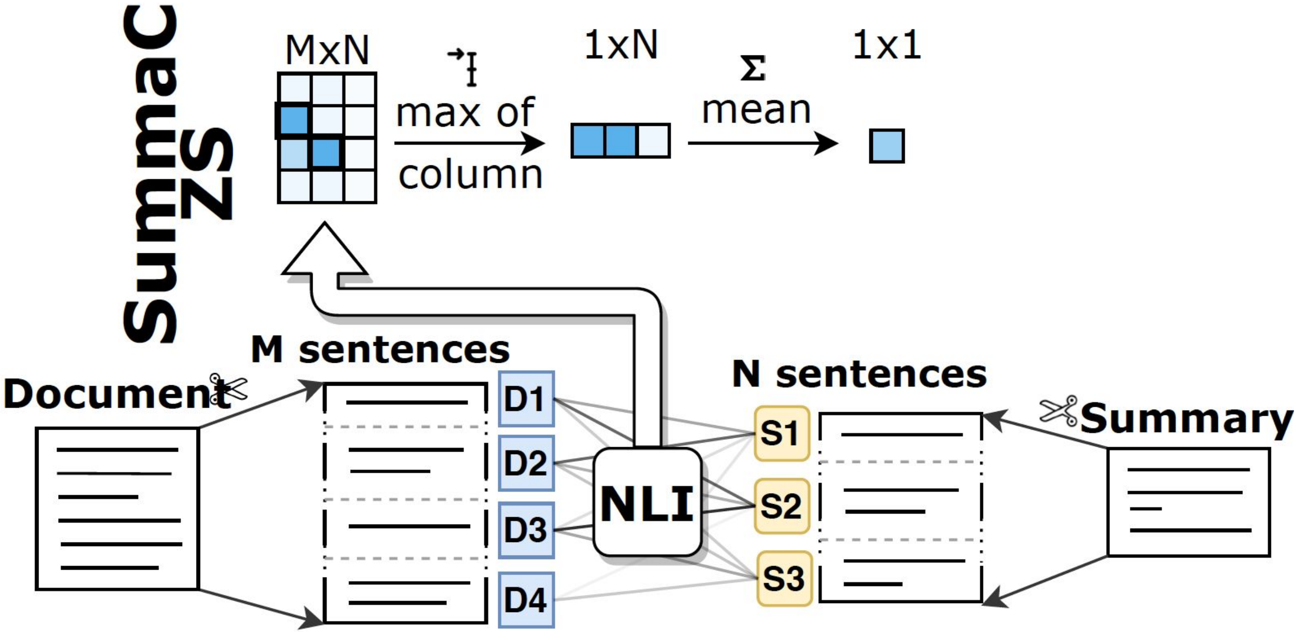
? Two undefined parameters: which NLI model, and which NLI label to use.

SummaC Benchmark Results

Type	Model Name	CGS	XFS	PT	FCC	SE	Fr	Total
Classifier	FactCC-CLS	63.1	57.6	61.0	75.9	60.1	59.4	62.8
Parsing	DAE	63.4	50.8	62.8	75.9	70.3	61.7	64.2
QAG	FEQA	61.0	56.0	57.8	53.6	53.8	69.9	58.7
	QuestEval	62.6	**62.1**	**70.3**	66.6	72.5	**82.1**	69.4
NLI	NLI Doc Level	53.0	57.5	61.0	61.3	66.6	63.6	56.8
	SummaC ZS	**70.4**	58.4	62.0	**83.8**	**78.7**	79.0	**72.1**

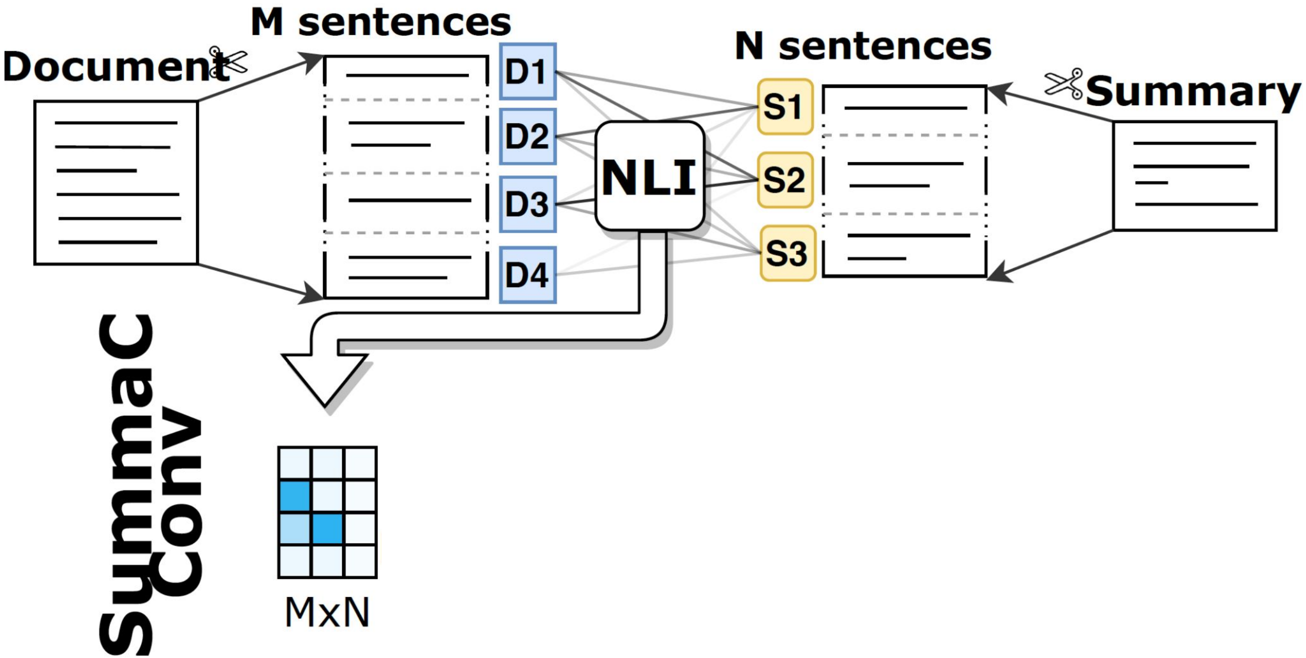
Results of inconsistency detectors on the
SummaC Benchmark (Balanced Accuracy)

SummaC Zero Shot



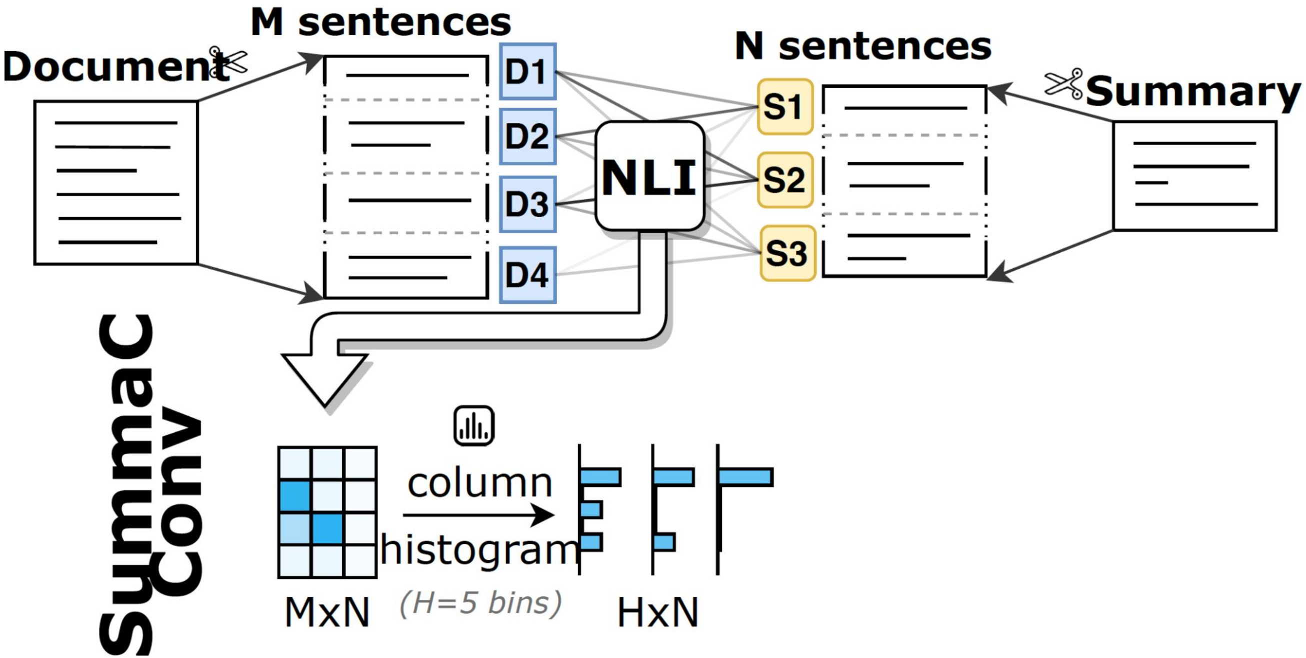
! The max operator is limiting: it removes a lot of information. Can we do better?

SummaC Convolution



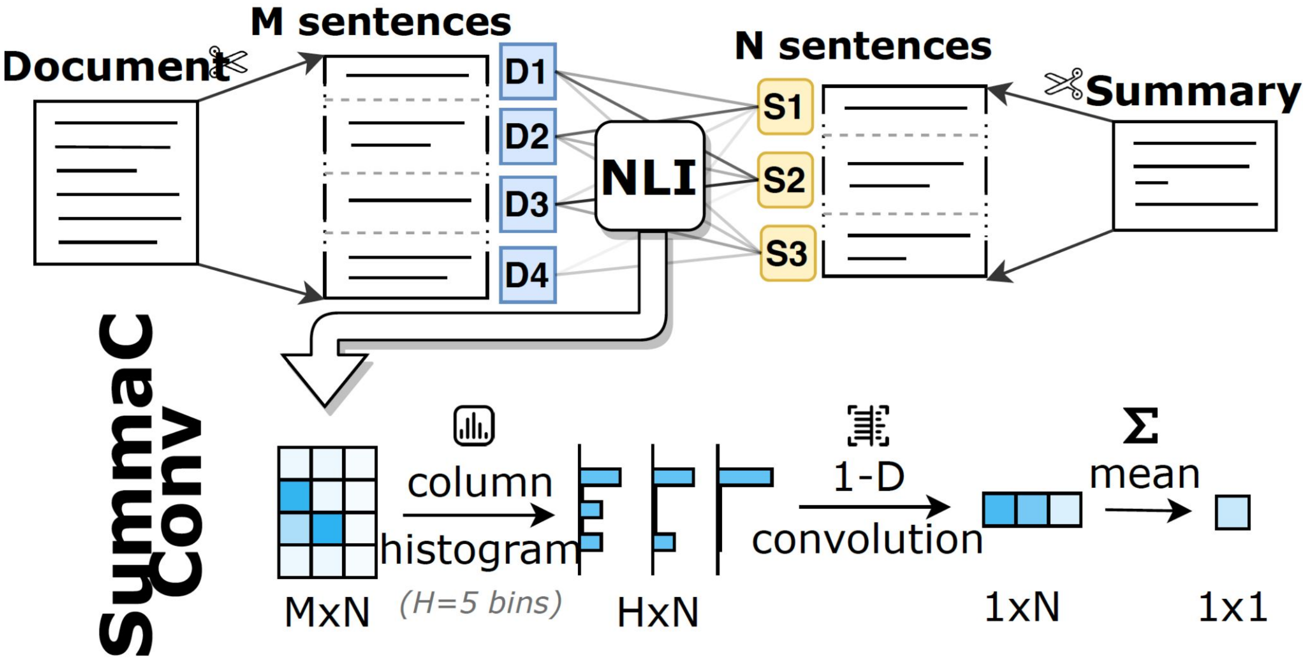
1. Build NLI (MxN) matrix (identical to SummaC ZS).

SummaC Convolution



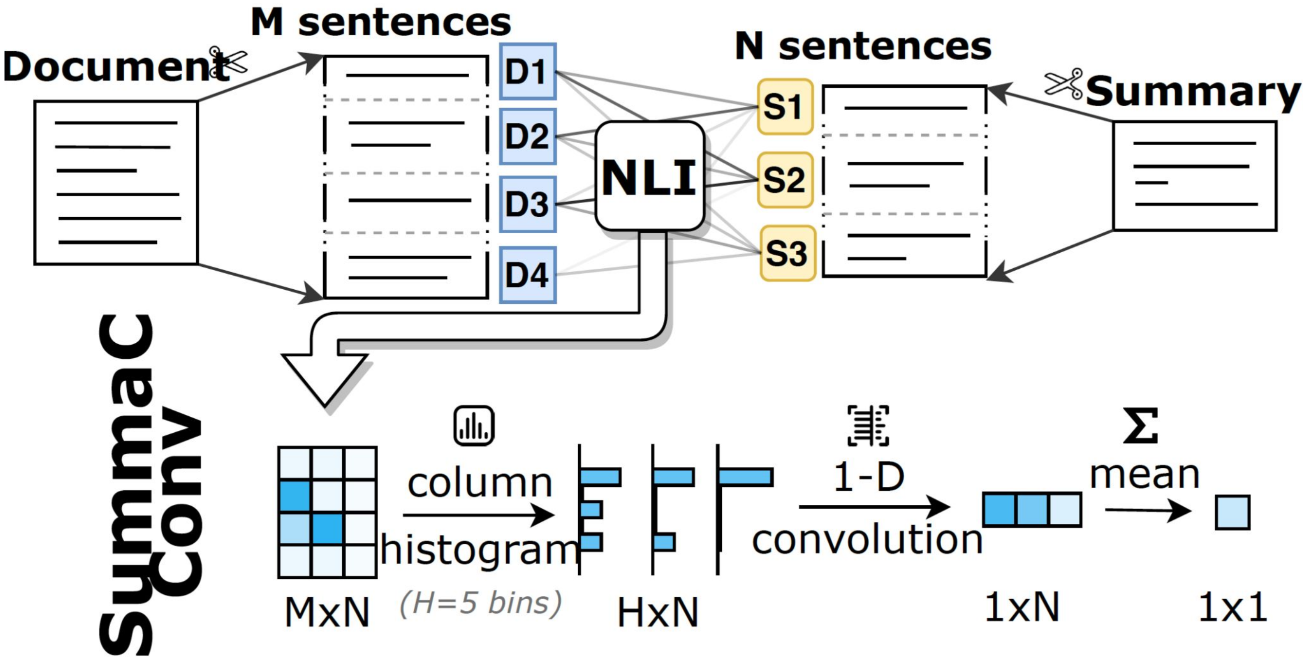
2. For each column, compute a bin (fixed size).

SummaC Convolution



3. Input each histogram into a trained 1-d convolution layer.

SummaC Convolution



4. Take the mean of all convolution outputs.

SummaC Benchmark Results

Type	Model Name	CGS	XFS	PT	FCC	SE	Fr	Total
Classifier	FactCC-CLS	63.1	57.6	61.0	75.9	60.1	59.4	62.8
Parsing	DAE	63.4	50.8	62.8	75.9	70.3	61.7	64.2
QAG	FEQA	61.0	56.0	57.8	53.6	53.8	69.9	58.7
	QuestEval	62.6	62.1	**70.3**	66.6	72.5	**82.1**	69.4
NLI	NLI Doc Level	53.0	57.5	61.0	61.3	66.6	63.6	56.8
	SummaC ZS	**70.4**	58.4	62.0	83.8	78.7	79.0	72.1
	SummaC Conv	64.7	**66.4**	62.7	**89.5**	**81.7**	**81.6**	**74.4**

Results of inconsistency detectors on the SummaC Benchmark (Balanced Accuracy)

NLI Model Selection

		Benchmark Acc.	
Model	**NLI Dataset**	**ZS**	**Conv.**
Decomp Attn	SNLI	56.9	56.4
BERT-Base	SNLI	66.6	64.0
	MNLI	69.5	69.8
	MNLI + VitC	67.9	71.2
BERT-Large	SNLI	66.6	62.4
	MNLI	70.9	73.0
	MNLI + VitC	**72.1**	**74.4**

Highlights:

1. Better NLI leads to better inconsistency detection.

2. SummaCConv only leads to improvements over ZS for better NLI.

NLI Category Selection

Category			SummaCConv Acc.	
E	N	C	MNLI + VitC	MNLI
✔			**74.4**	72.6
	✔		71.2	66.4
		✔	72.5	72.6
✔	✔		73.1	72.6
✔		✔	74.0	**73.0**
	✔	✔	69.2	72.6
✔	✔	✔	69.7	**73.0**

Highlights:

1. **E**ntailment-only models are very strong.
2. **N**eutral is not useful.
3. **C**ontradiction helps the MNLI model a little.

Interpretation:

Inconsistency detection is not about detecting contradictions. It is more about finding support (entailment) for summary claims.

Discussion

1. **Choice of granularity.** See paper for experiment. TL/DR: finer granularity is better.

2. **Focus on News Summarization.** Future annotations should focus on new domains (dialogue, medical, etc.) and languages for annotation.

3. **Towards Consistent Summarization.** Inconsistency detection is the first step. Can we make the next generation of summarizers consistent?

Thank you

TL;DR: New SummaC Benchmark to tackle factual consistency. Two new NLI-based models: SummaCZS, SummaCConv are strong performers.

Code/Data: github.com/tingofurro/summac

Questions? Get in touch, in person at ACL or online: plaban@salesforce.com