Extractive Fact Decomposition for Interpretable Natural Language Inference in One Forward Pass

Nicholas Popovič, Michael Färber

Can atomic fact decomposition be distilled into encoder-only architectures to enable fast, scalable, and faithfully interpretable NLI without requiring LLMs at inference time?

OVERVIEW

Motivation

- Fact decomposition makes NLI interpretable and robust
- However, this requires a LLM at inference time

Our approach

- Reframe fact decomposition as an extractive task
- Build a synthetic data generation pipeline for additional annotation
- Distill fact decomposition into encoder-only models

Text

The Ottawa Sun is a daily tabloid newspaper in Ottawa, Ontario, Canada. It is published by Sun Media. It was first published in 1983 as the "Ottawa Sunday Herald", until it was acquired by (then) Toronto Sun Publishing Corporation in 1988. In April 2015, Sun Media papers were acquired by Postmedia.

Abstractive Fact Decomposition

1 The Ottawa Sun is a daily tabloid newspaper in Ottawa, Ontario, Canada
2 The Ottawa Sun was first published in 1983 as the "Ottawa Sunday Herald"
3 The Ottawa Sun was acquired by (then) Toronto Sun Publishing Corporation in 1988
4 Sun Media papers were acquired by Postmedia in April 2015
5 The Ottawa Sun is published by Sun Media
6 The Ottawa Sun is a tabloid newspaper

Extractive Fact Decomposition

The Ottawa Sun is a daily tabloid newspaper 6 in Ottawa, Ontario, Canada 1 . It is published by Sun Media 5 . It was first published in 1983 as the "Ottawa Sunday Herald" 2 , until it was acquired by (then) Toronto Sun Publishing Corporation in 1988 3 . In April 2015, Sun Media papers were acquired by Postmedia. 4

DATA COLLECTION

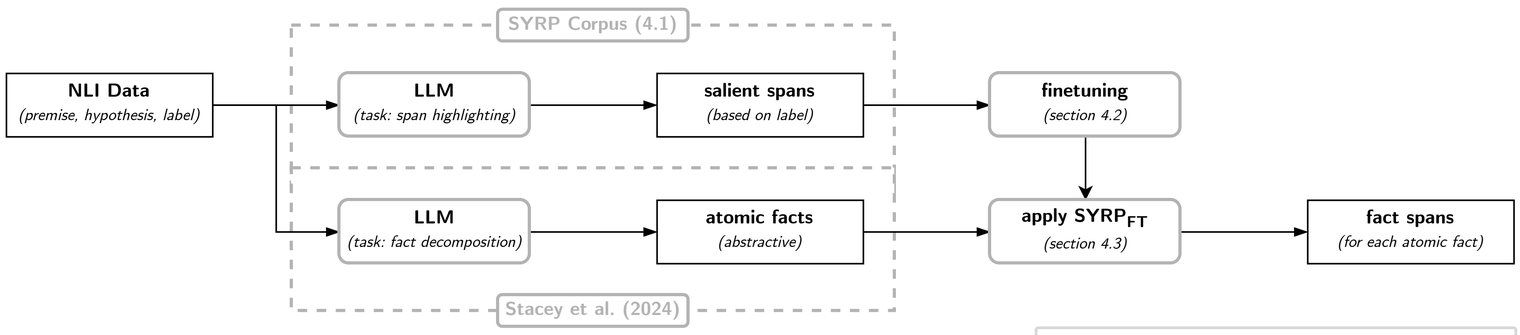
Premise

The Ottawa Sun is a daily tabloid newspaper in Ottawa, Ontario, Canada. It is published by Sun Media. It was first published in 1983 as the "Ottawa Sunday Herald", until it was acquired by (then) Toronto Sun Publishing Corporation in 1988. In April 2015, Sun Media papers were acquired by Postmedia.

Hypothesis Contradiction ✗

Toronto Sun Publishing acquired the Ottawa Sun in the late nineties.

The Ottawa Sun is a daily tabloid newspaper in Ottawa, Ontario, Canada. It is published by Sun Media. It was first published in 1983 as the "Ottawa Sunday Herald", untilit was acquired by (then) Toronto Sun Publishing Corporation in 1988. In April 2015, Sun Media papers were acquired by Postmedia.

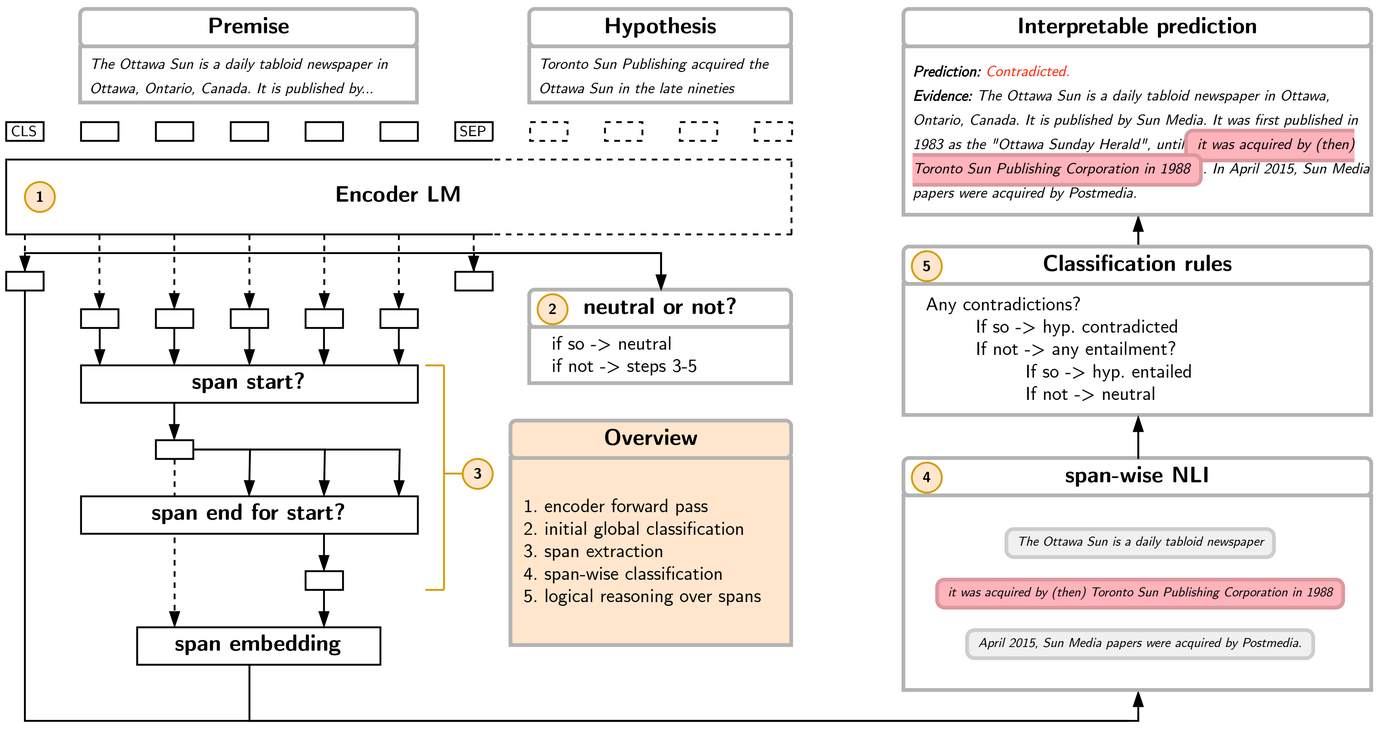


1 The Ottawa Sun is a daily tabloid newspaper in Ottawa, Ontario, Canada
2 The Ottawa Sun was first published in 1983 as the "Ottawa Sunday Herald"
3 The Ottawa Sun was acquired by (then) Toronto Sun Publishing Corporation in 1988
4 Sun Media papers were acquired by Postmedia in April 2015
5 The Ottawa Sun is published by Sun Media
6 The Ottawa Sun is a tabloid newspaper

Fact/salient spans

The Ottawa Sun is a daily tabloid newspaper FACT in Ottawa, Ontario, Canada FACT . It is published by Sun Media FACT . It was first published in 1983 as the "Ottawa Sunday Herald" FACT , until it was acquired by (then) Toronto Sun Publishing Corporation in 1988 SALIENT FACT . In April 2015, Sun Media papers were acquired by Postmedia. FACT

MODEL ARCHITECTURE



EVALUATION

Model	In-distribution R1	R2	R3	ANLI-all	Out-of-distribution ConTRoL	RTE	WNLI
not interpretable:							
DeBERTa$_{LARGE}$	78.3%	66.5%	61.7%	68.1%	56.0%	90.4%	68.9%
sentence atoms:							
SenLR	76.7%	64.8%	62.0%	67.5%	56.3%	86.3%	64.5%
JEDI$_{sent}$ (ours)	**77.4%**	**65.1%**	**62.3%**	**67.9%**	**57.6%**	**90.6%**	**67.8%**
span/fact atoms:							
SLR-NLI	74.7%	60.4%	58.3%	64.1%	**54.7%**	87.5%	65.8%
JEDI (ours)	**75.5%**	**63.1%**	**59.4%**	**65.6%**	54.3%	**87.7%**	**73.7%**
FGLR (+GPT-3.5-turbo)	76.2%	64.8%	63.1%	67.7%	52.7%	82.0%	77.0%
token atoms:							
SYRP$_{FT}$ (ours)	75.9%	63.3%	59.3%	65.8%	46.3%	88.8%	65.3%

Table 1: Test set scores for DeBERTa$_{LARGE}$. Results are averaged accuracies across 10 random seeds.

Model	interp.?	ANLI	HANS
JEDI	✓	65.6%	76.9%
DeBERTa$_{LRG}$	✗	68.1% ↑	79.1% ↑
JEDI$_{global\ only}$	✗	68.2% ↑	80.1% ↑
JEDI$_{no\ global}$	✓	59.7% ↓	73.3% ↓
JEDI$_{w/o\ ATLoss}$	✓	65.3% ↓	74.5% ↓
JEDI$_{sent}$	✓↓	67.9% ↑	83.1% ↑
SYRP$_{FT}$	✓↑	65.8% ↑	33.0% ↓

Table 3: Results of ablation study for models with DeBERTa$_{LARGE}$ as backbone. *interp.?* indicates whether interpretability is preserved despite the changes, with ✗ indicating no interpretability, and ✓(↑ / ↓) indicating interpretability at a higher or lower level of detail. Arrows ↑↓ indicate direction of changes over JEDI.