Paper Review

AlignScore: Evaluating Factual Consistency with a Unified Alignment Function

Alejandro Jorba, Camille Dubois, Killian Varescon

Introduction

Objective :

Introduce AlignScore, a new metric for evaluating factual consistency in text generation tasks.

Motivation :

Existing metrics are task-specific (NLI, QA) and limited in generalizability.

Solution :

AlignScore, a holistic metric trained on diverse datasets from multiple tasks.

Table of contents

I. Paper Presentation

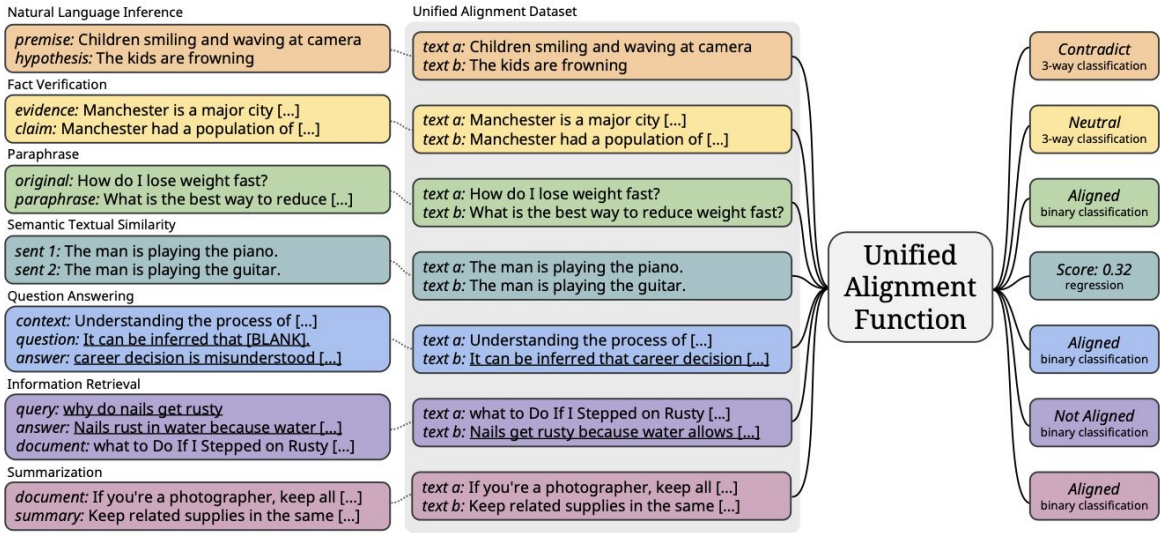
Concept: Given texts *a* and *b*, *b* is aligned with *a* if all information of *b* is present in *a* and does not contradict *a*.

Mapping: $f(a, b) \longmapsto y$

Challenges :

- Different input/output formats across tasks
- Unifying tasks into a uniform alignment training corpus

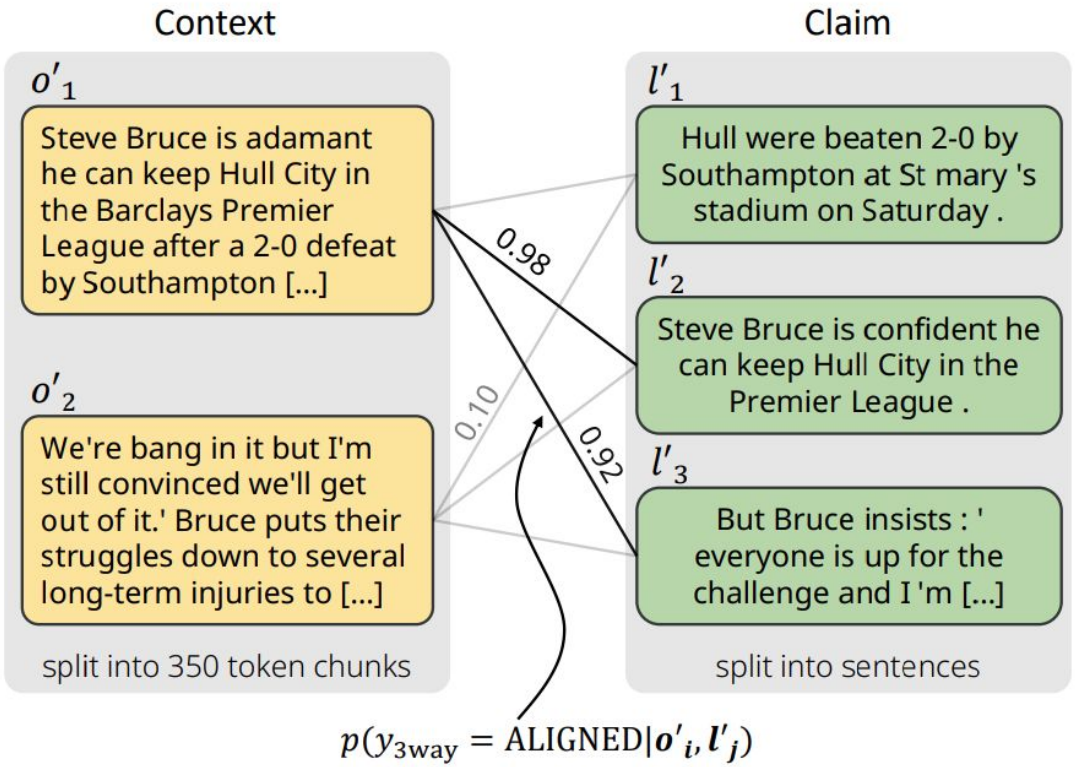
I. Paper presentation - Unifying formats



Captured from Yuheng Zha & al., 2023.

How to deal with long contexts ? Can a claim be “halfly true” ?

I. Paper presentation - ALIGN SCORE



$$\textsc{AlignScore}(\boldsymbol{o}, \boldsymbol{l}) = \underset{j}{\text{mean}} \underset{i}{\max} \text{alignment}(\boldsymbol{o'_i}, \boldsymbol{l'_j})$$

I. Paper presentation - Results

- Several benchmarks, mainly TRUE and SummaC,

- At-least-similar performance as other baseline metrics (FEQA, MNLI, …),

- Meaning of the comparison ? Is it reliable ?

II. Implementation

- We ran the SummaC benchmark, with 3 baseline metrics.

→ Huge time spent on debugging both the Align-Score & the SummaC codes.

- Great coherence in the results, but on one dataset (up to ~ 40% difference).

https://github.com/alejorba/AlignScore-summac-benchmark

II. Implementation - Results

Type	Metric	CGS	XSF	PolyTope	SummEval	FRANK
Similarity Matching	BERTScore	63.1	49.0	61.3	70.1	84.8
NLI	MNLI	44.9	46.6	51.1	45.5	59.4
Misc	BLANC	54.1	53.5	70.4	60.5	83.4
AlignScore	ALIGNSCORE-large	86.4	75.7	53.3	81.0	91.4

Table 1: Results obtained on our implementation of the SummaC benchmark

Type	Metric	CGS	XSF	PolyTope	SummEval	FRANK
Similarity Matching	BERTScore	63.1	49.0	85.3	79.6	84.9
NLI	MNLI	44.9	46.6	45.0	43.5	59.3
Misc	BLANC	54.1	53.5	74.7	68.6	83.4
AlignScore	ALIGNSCORE-large	86.4	75.8	92.4	91.7	91.4

Table 2: SummaC benchmark results presented on the paper [extract]

III. Analysis

- Original & interesting idea to assess factual consistency on a large variety of tasks.
- Well motivated problem.
- How to unify datasets of different tasks ? Key aspect of the paper.
- Complex metrics interpretability and comparison.

III. Analysis - Results

- Comparison with other baseline metrics that are task-specific = different context

→ Factual consistency in a Q&A task implies that the answer is aligned with the question AND the context.

III. Analysis - Practical application

Why not using ALIGN-SCORE to evaluate factual consistency on one LLM performing different NLP task ?

→ “Does ChatGPT hallucinate more when paraphrasing or answering questions ?”

Thanks !

Now, let’s try out our question answering factual consistency regarding the paper…

Questions ?