FIZZ: Factual Inconsistency Detection by Zoom-in Summary and Zoom-out Document

EMNLP 2024

Joonho Yang[1], Seunghyun Yoon[2], Byeongjeong Kim[1], Hwanhee Lee[1]*

[1]Department of Artificial Intelligence, Chung-Ang University, [2]Adobe Research USA

Adobe

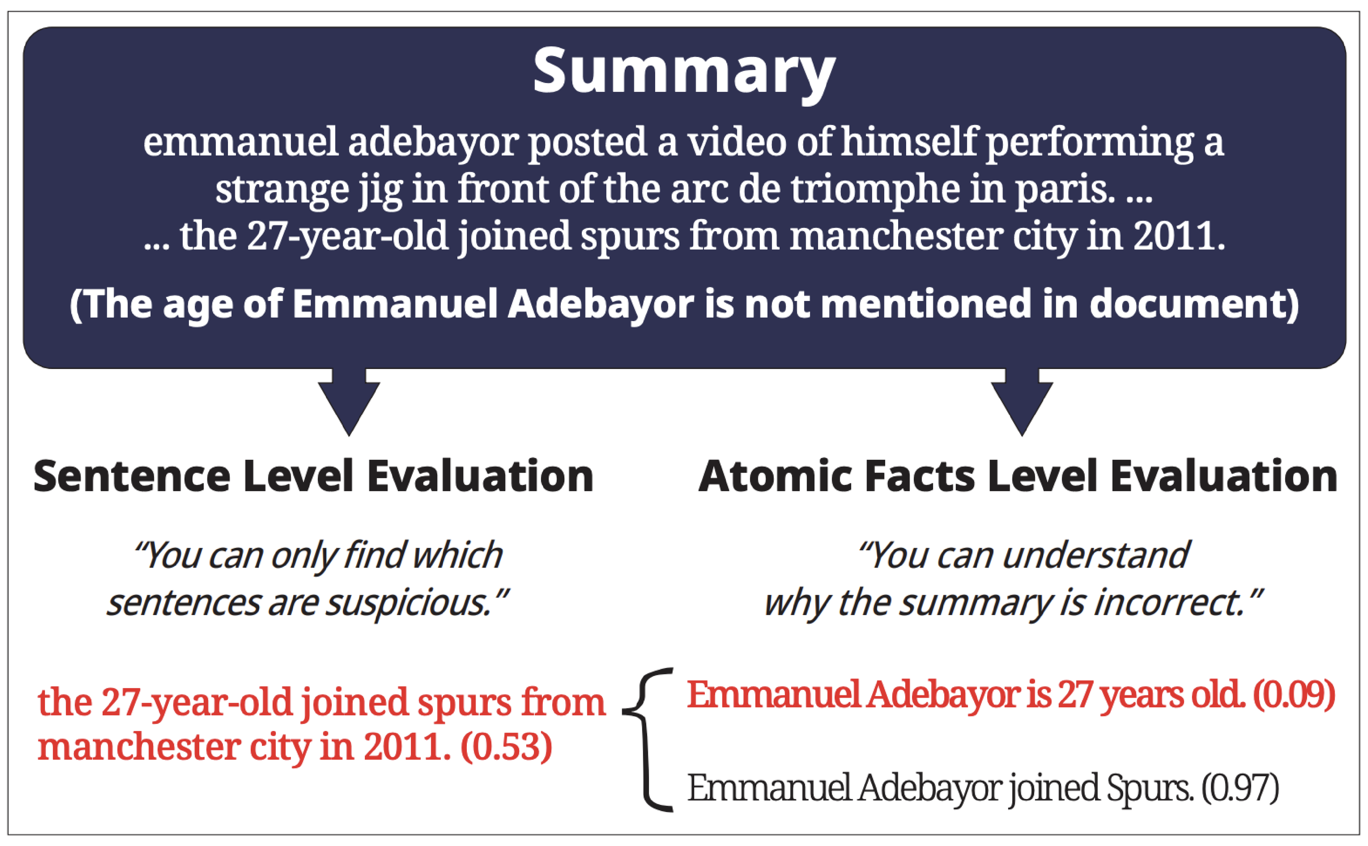
Factual Inconsistency in Summary

- Abstractive summarization systems based on pre-trained language models aim to generate human-like summaries.
- However, these models may generate **factually inconsistent summaries** that **do not align with the source document**, often due to hallucinations.

Motivation

- Current **evaluation systems** for verifying the **factual consistency** of summaries have significant limitations in both **accuracy** and **interpretability**.

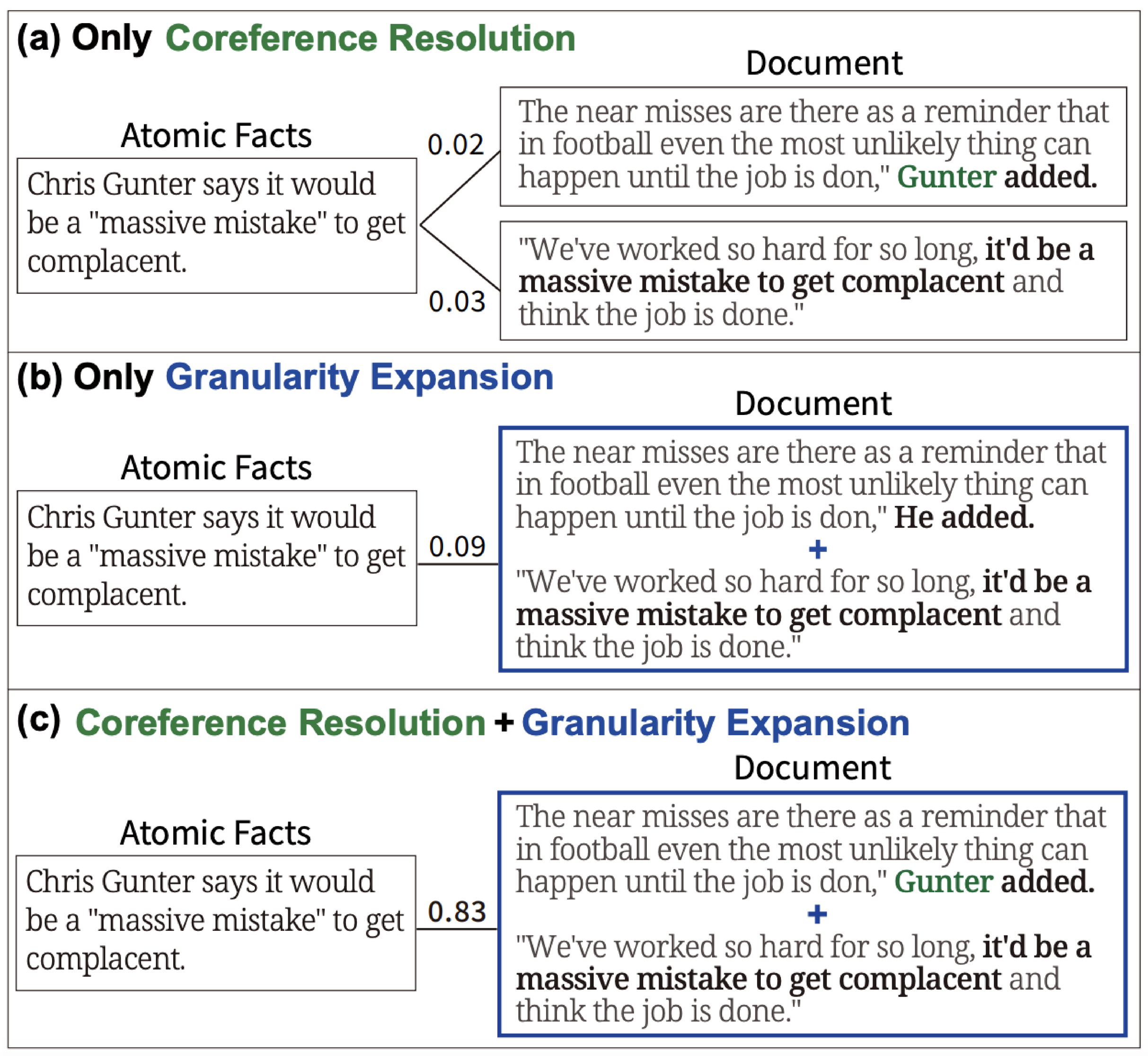
Atomic Fact Level Evaluation



(A) Atomic facts level evaluation for the summary

- We propose a novel summarization inconsistency detection system, which significantly enhances both **accuracy** and **interpretability** by **breaking down summaries into atomic facts** and **expanding the granularity** of document analysis.

Coreference Resolution & Granularity



(C) The effect of granularity expansion and coreference resolution

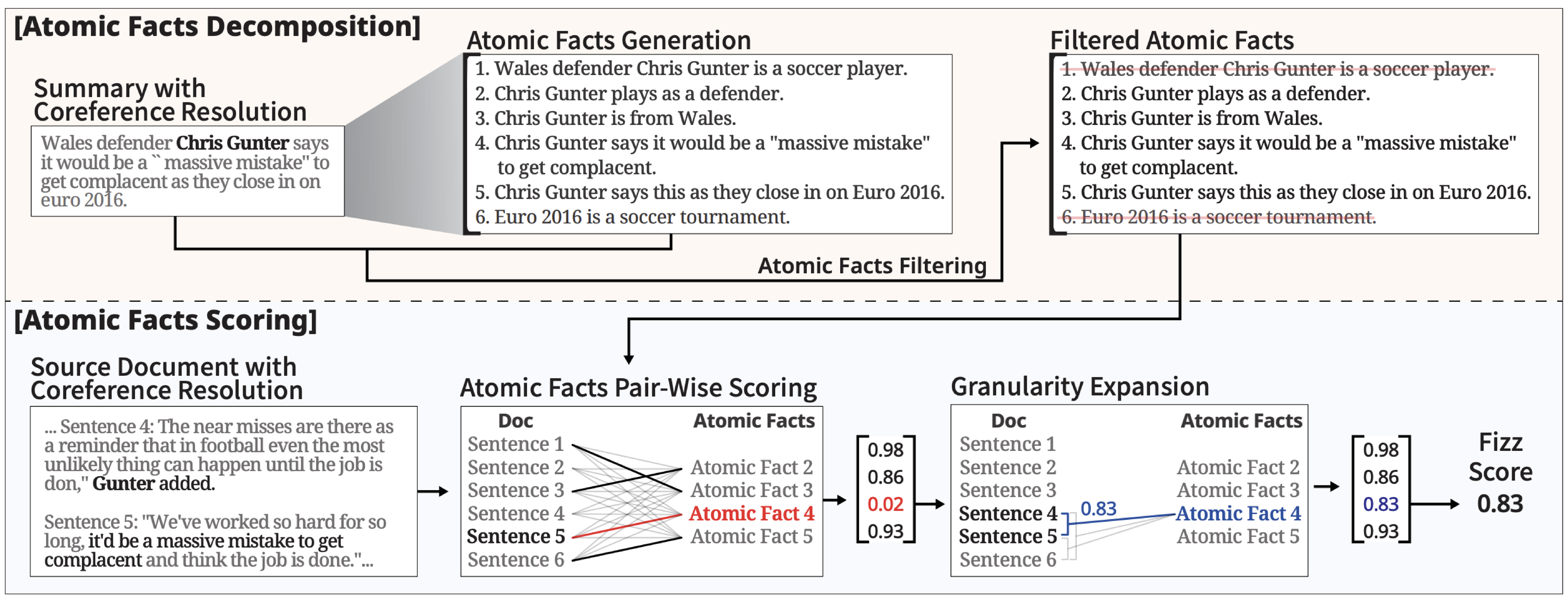
Experimental Results

	AggreFact-Cnn-FtSota	AggreFact-XSum-FtSota	Avg
DAE	65.4±4.4	**70.2**±2.3	67.8
QuestEval	70.2±3.2	59.5±2.7	64.9
SummaC-ZS	64.0±3.8	56.4±1.2	60.2
SummaC-Conv	61.0±3.9	65.0±2.2	63.0
QAFactEval	67.8±4.1	63.9±2.4	65.9
AlignScore	62.5±3.3	69.6±1.7	66.1
ChatGPT-ZS	56.3±2.9	62.7±1.7	59.5
ChatGPT-COT	52.5±3.3	55.9±2.1	54.2
ChatGPT-DA	53.7±3.5	54.9±1.9	54.3
ChatGPT-Star	56.3±3.1	57.8±0.2	57.1
FactScore	60.8±3.2	68.0±2.0	64.4
FacTool	49.3±3.5	59.0±2.0	54.2
FIZZ (Ours)	**72.6**±3.0	69.3±1.9	**71.0**
w/o GE	72.2±2.8	66.3±1.9	69.3
w/o Filtering	64.7±3.3	70.0±1.8	67.4
w/o AF	63.6±2.9	65.8±2.0	64.7

	AggreFact-Cnn FtSota	AggreFact-Cnn ExF	AggreFact-Cnn Old	AggreFact-XSum FtSota	AggreFact-XSum ExF	AggreFact-XSum Old	Avg
Baseline	50.0	50.0	50.0	50.0	50.0	50.0	50.0
DAE*	59.4	67.9	69.7	**73.1**	-	-	67.5
QuestEval	63.7	64.3	65.2	61.6	60.1	59.7	62.4
SummaC-ZS	63.3	**76.5**	76.3	56.1	51.4	53.3	62.8
SummaC-Cv	70.3	69.8	78.9	67.0	64.6	67.5	69.7
QAFactEval	61.6	69.1	**80.3**	65.9	59.6	60.5	66.2
AlignScore	53.4	73.1	80.2	70.2	**80.1**	63.7	70.1
ChatGPT-ZS	66.2	64.5	74.3	62.6	69.2	60.1	66.2
ChatGPT-CoT	49.7	60.4	66.7	56.0	60.9	50.1	57.3
ChatGPT-DA	48.0	63.6	71.0	53.6	65.6	61.5	60.6
ChatGPT-Star	55.8	65.8	71.2	57.7	70.6	53.8	62.5
FactScore	69.9	71.6	73.9	68.0	63.5	66.8	69.0
FacTool	72.7	66.1	60.8	68.0	64.0	62.2	65.6
FIZZ (Ours)	**73.2**	67.3	76.0	69.7	72.4	**68.5**	**71.2**

- We present the performance of various methods on the **AggreFact** benchmark dataset. (Tang et al., 2023)
- Experimental results show that our proposed system FIZZ achieves state-of-the-art performance.
- FIZZ exhibits **high interpretability** by **utilizing atomic facts.**

Overall Pipeline



(B) Overall pipeline of *FIZZ*

1) The pipeline begins by applying coreference resolution to both the summary and the document.

2) The summary is decomposed into **atomic facts** using an LLM.

3) The **atomic facts** are filtered and scored against the document.

4) The scores are refined through granularity expansion of the document.

5) The **minimum** score is defined as the final score.