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

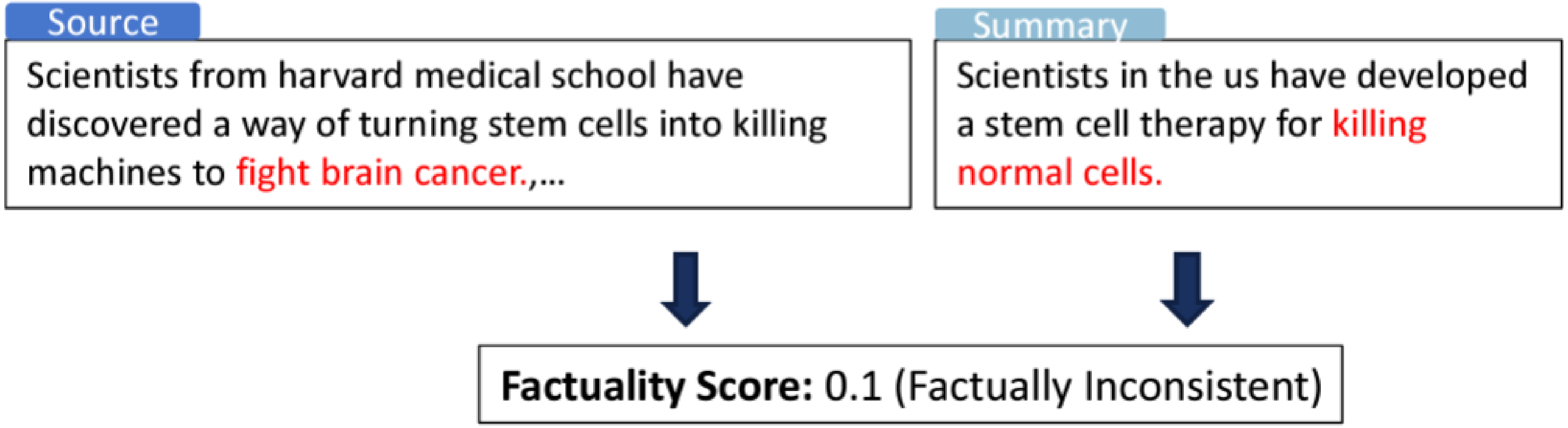
Joonho Yang[1] Seunghyun Yoon[2] Byeongjeong Kim[1] *Hwanhee Lee[1]

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

LILAB Adobe EMNLP 2024

Factual Inconsistency Detection

Factual Consistency: whether the generated text is factually consistent with the source (ex. news article, book, paper)

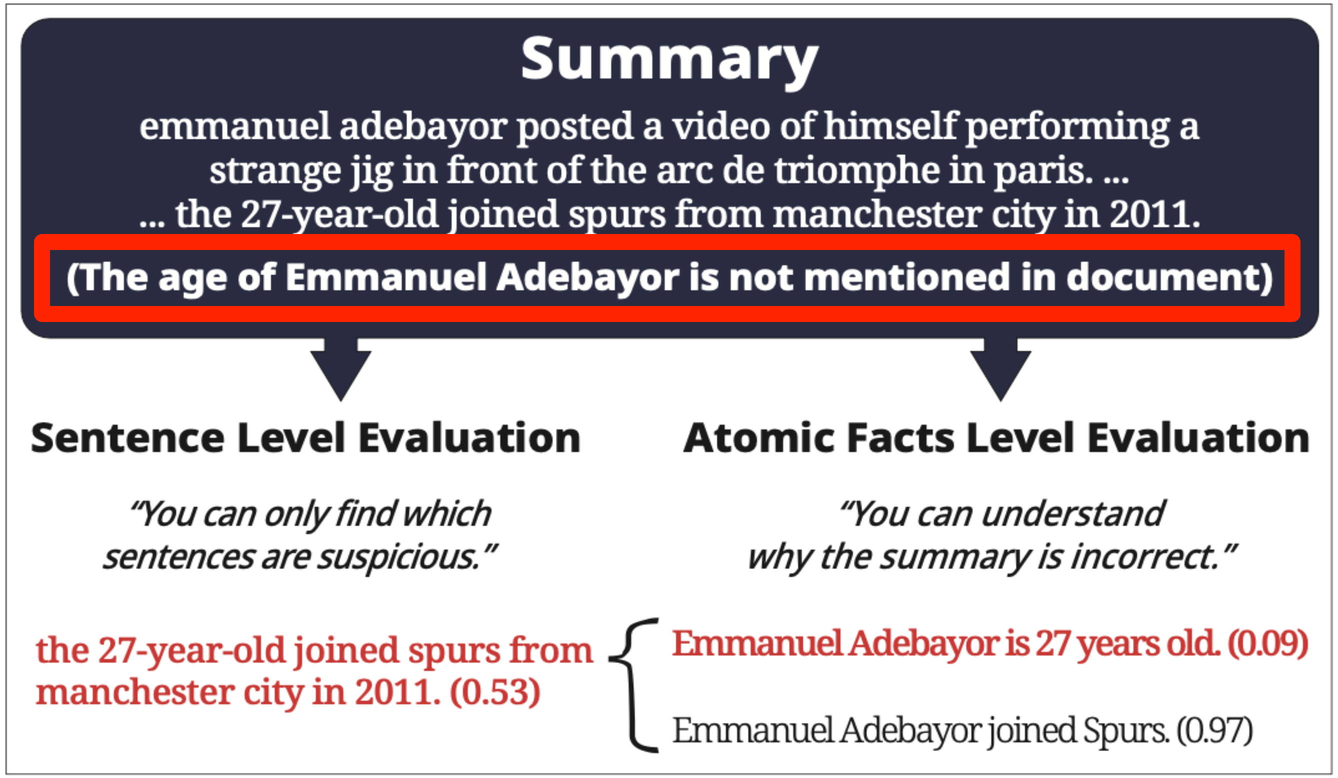


- **Task:** Catching a minor factual error (wrong entity or relation, coreference error, out of article, grammatical, etc.) of the summary

Masked Summarization to Generate Factually Inconsistent Summaries for Improved Factual Consistency Checking (Lee et al., NAACL 2022)

Motivation

- Sentence-level Evaluation vs. *Atomic Fact*-level Evaluation



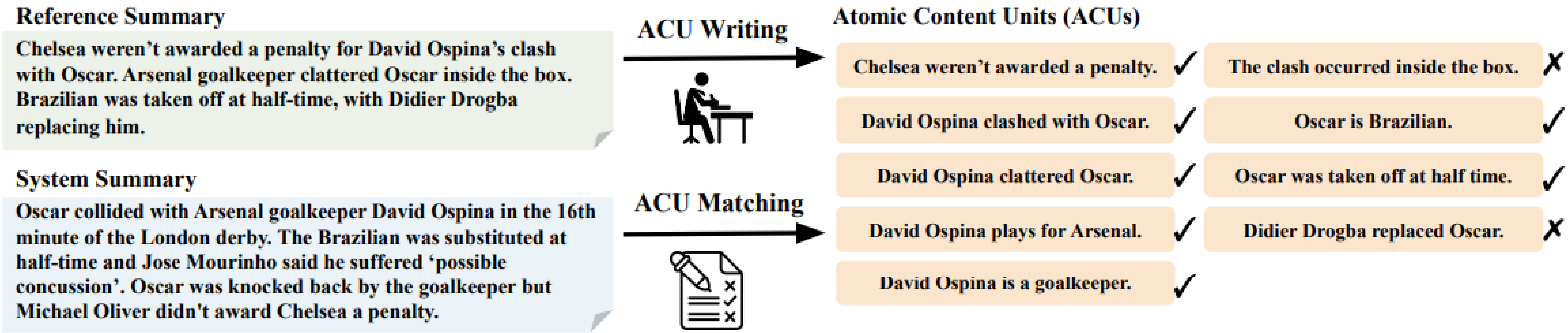
Atomic Fact: A more fine-grained information unit than a *sentence*

- By employing *atomic facts*, FIZZ demonstrates high accuracy and strong interpretability.

Atomic Facts

Examining the Consensus between Human Summaries: Initial Experiments with Factoid Analysis

(van Halteren & Teufel, NAACL 2003)

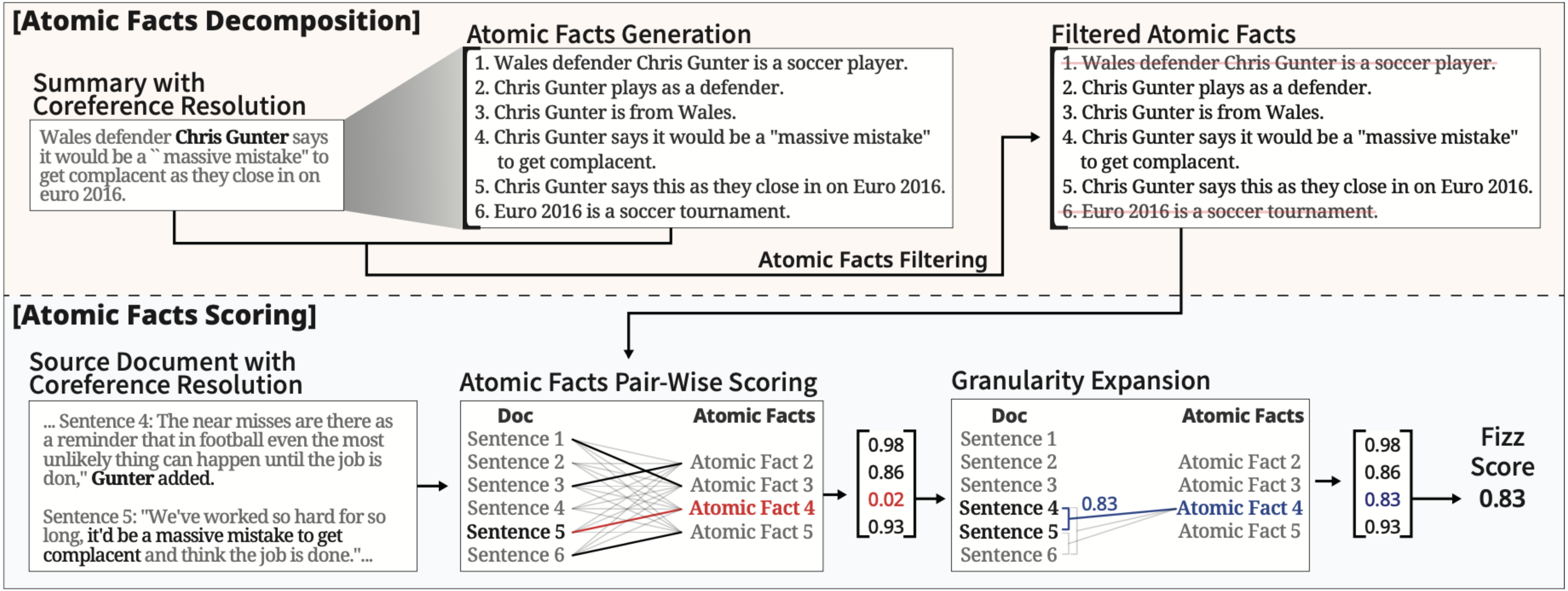


It can be impossible to provide a practical definition of atomic facts.

Our definition of an *atomic fact*: short and concise, containing no more than two or three entities, with person entities specifically resolved any of coreferences

Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation (Liu et al., ACL 2023)

Overall Pipeline



Atomic Facts Decomposition: Zoom-in Summary

Document D, Summary S

- **Coreference Resolution**

$$D' = f_{coref}(D), \qquad S' = f_{coref}(S)$$

- **Atomic Facts Generation**

$$S' = \{s'_j\}_{j=1}^{N}, \qquad A' = \{a'_k\}_{k=1}^{L}$$

Probability distribution of NLI model

Entailment (E), *Contradiction* (C), *Neutral* (N)

- **Atomic Facts Filtering**

S' as the *premise*, A' as the *hypothesis*

We filtered out the *atomic facts* that did not meet the following condition:

$$\max(E, C, N) = E$$

Summary sentence:

"Chris Gunter says … on Euro 2016."

Filtered out *atomic facts*:

"Chris Gunter is a soccer player."

"Euro 2016 is a soccer tournament."

Atomic Facts Filtering

Atomic Facts

1. Wales defender Chris Gunter is a soccer player.
2. Chris Gunter plays as a defender.
3. Chris Gunter is from Wales.
4. Chris Gunter says it would be a "massive mistake" to get complacent.
5. Chris Gunter says this as they close in on Euro 2016.
6. Euro 2016 is a soccer tournament.

Filtering →

Filtered Atomic Facts

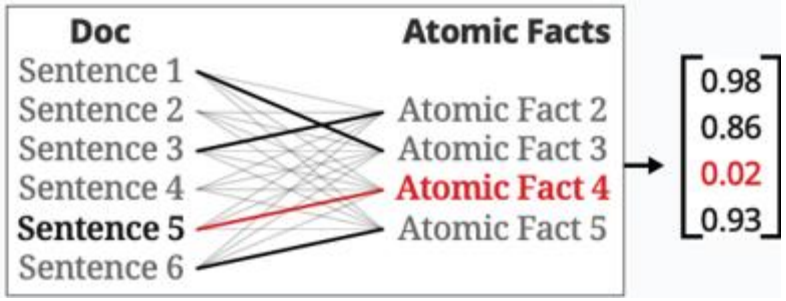
~~1. Wales defender Chris Gunter is a soccer player.~~
2. Chris Gunter plays as a defender.
3. Chris Gunter is from Wales.
4. Chris Gunter says it would be a "massive mistake" to get complacent.
5. Chris Gunter says this as they close in on Euro 2016.
~~6. Euro 2016 is a soccer tournament.~~

Atomic Facts Scoring: Zoom-out Document

- **Atomic Facts Pair-Wise Scoring**

$$D' = \{d'_i\}_{i=1}^{M}, \qquad A_{filtered} = \{a_k\}_{k=1}^{L}$$

D' as the *premise*, $A_{filtered}$ as the *hypothesis*



We assign scores to each *atomic fact* based on the **maximum** *entailment* score

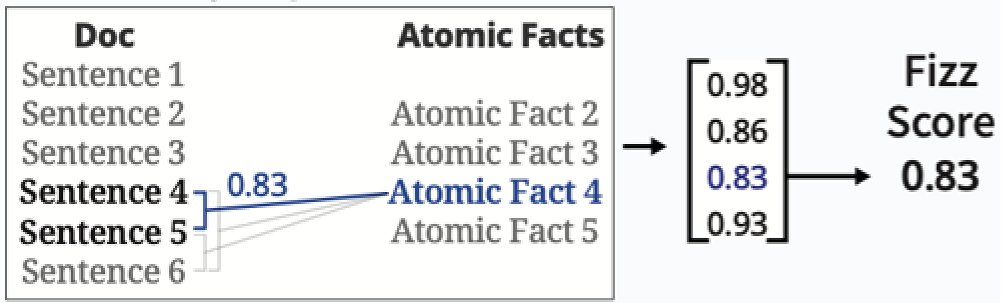
$$E = \{e_{i,k}\}, 1 \leq i \leq M, 1 \leq k \leq L$$

$$\boldsymbol{t}_k = \max_{1 \leq i \leq M} e_{i,k}$$

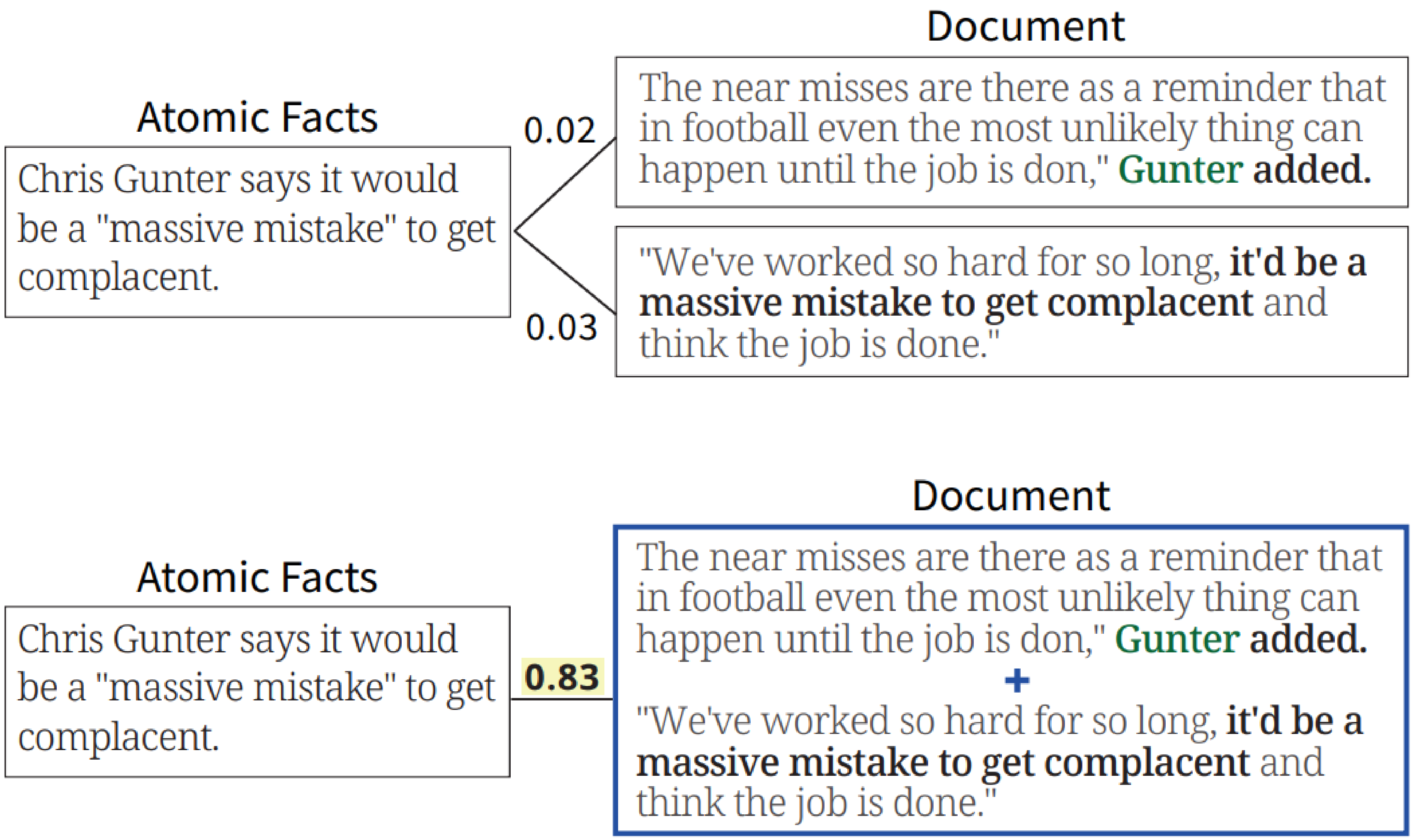
$$\boldsymbol{T} = \{\boldsymbol{t}_1, \dots, \boldsymbol{t}_L\}$$

- **Adaptive Granularity Expansion**

It is necessary to examine multiple sentences across the document because of *abstractiveness*



Granularity Expansion



Experimental Setups: Dataset

- **Dataset**

AggreFact (Tang et al., ACL 2023)

- **Evaluation**

document d, summary s, metric $f(d, s) \rightarrow y$

$$y \in \mathbb{R}$$

Threshold t

$$f(d, s) > t = 1, else\ 0$$

$$f'(d, s) \rightarrow \{0, 1\}$$

Threshold-per-dataset

Single-threshold (*FtSota* split)

- AggreFact-CNN

		Polytope	FactCC	SummEval	FRANK	Wang'20	CLIFF	Goyal'21	Total
OLD	val	450	931	550	223	118	-	25	2297
	test	450	503	548	523	117	-	25	2166
XFORMER	val	150	-	50	75	-	-	-	275
	test	150	-	50	175	-	-	-	375
SOTA	val	34	-	200	75	-	150	-	459
	test	34	-	200	175	-	150	-	559

- AggreFact-XSum

		XsumFaith	Wang'20	CLIFF	Goyal'21	Cao'22	Total
OLD	val	500	-	-	-	-	500
	test	430	-	-	-	-	430
XFORMER	val	500	-	-	-	-	500
	test	423	-	-	-	-	423
SOTA	val	-	120	150	50	457	777
	test	-	119	150	50	239	558

Understanding Factual Errors in Summarization: Errors, Summarizers, Datasets, Error Detectors (Tang et al., ACL 2023)

Experimental Setups: Baselines

- **QA-based Methods**

QuestEval (Scialom et al., EMNLP 2021); QAFactEval (Fabbri et al., NAACL 2022)

- **Parsing-based Methods**

DAE (Goyal & Durret, NAACL 2021)

- **NLI-based Methods**

SummaC (Laban et al., TACL 2022); AlignScore (Zha et al., ACL 2023)

- **LLM-based Methods**

ChatGPT-ZS, ChatGPT-CoT (Luo et al., 2023); ChatGPT-DA, ChatGPT-Star (Wang et al., 2023)

- **LLM-based *Atomic Facts* Methods**

FActScore (Min et al., EMNLP 2023); FacTool (Chern et al., 2023)

Experimental Settings

- **Coreference Resolution Model**

Encoder-Decoder Model

MT5-11B (Bohnet et al., TACL 2023)

- **NLI Model**

ALBERT-based *Encoder-Decoder* Model

Trained on SNLI (Bowman et al., EMNLP 2015), MNLI (Williams et al., NAACL 2018), ANLI (Nie et al., ACL 2020), VitaminC (Schuster et al., NAACL 2021)

- **Coreference Resolution Customization**

Original Text		**The 27-year-old** joined spurs from manchester city in 2011.
Others	**Coref Resolved Text**	**Emmanuel Adebayor** joined spurs from manchester city in 2011.
	Atomic Fact #1	Emmanuel Adebayor joined spurs.
	Atomic Fact #2	Emmanuel Adebayor joined spurs from manchester city.
	Atomic Fact #3	Emmanuel Adebayor joined spurs in 2011.
Ours	**Coref Resolved Text**	**Emmanuel Adebayor, the 27-year-old** joined spurs from manchester city in 2011.
	Atomic Fact #1	**Emmanuel Adebayor is 27-year-old.**
	Atomic Fact #2	Emmanuel Adebayor joined spurs.
	Atomic Fact #3	Emmanuel Adebayor joined spurs from manchester city.
	Atomic Fact #4	Emmanuel Adebayor joined spurs in 2011.

Main Results

- **<u>Balanced Accuracy</u> Results on AggreFact & FtSota split**

	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	<u>60.5</u>	66.2
AlignScore	53.4	<u>73.1</u>	<u>80.2</u>	<u>70.2</u>	**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	<u>72.7</u>	66.1	60.8	68.0	64.0	62.2	65.6
FIZZ (Ours)	**73.2**	67.3	76.0	69.7	<u>72.4</u>	**68.5**	**71.2**

	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	<u>72.2</u>±2.8	66.3±1.9	<u>69.3</u>
w/o Filtering	64.7±3.3	<u>70.0</u>±1.8	67.4
w/o AF	63.6±2.9	65.8±2.0	64.7

Analysis

- **LLMs used for *atomic facts* generation**

Open-source LLMs

Zephyr-7B, Mistral-7B, Orca2-7B

Commerical LLMs

gpt-3.5-turbo, gpt-3.5-turbo-instruct

LLM	CNN	XSUM	AVG	AVG. TOKEN LENGTH
Zephyr	65.1±3.3	65.2±2.0	65.2	**97.6**
gpt-3.5-turbo	68.7±3.4	68.7±2.0	68.7	95.9
gpt-3.5-turbo-instruct	70.7±3.1	67.0±1.8	68.9	90.5
Mistral	70.5±3.5	68.7±2.1	69.6	86.5
Orca-2	**72.6**±3.0	**69.3**±1.9	**71.0**	81.4

	ROUGE-1 P	ROUGE-1 R	ROUGE-1 F1	AVG. NUMBER OF ATOMIC FACTS	AVG. TOKEN LENGTH
Human	1.00	1.00	1.00	8.7	98.4
Orca-2	0.70	0.69	0.68	**8.7**	**96.3**
gpt-3.5-turbo	**0.78**	**0.84**	**0.79**	7.8	105.0
gpt-3.5-turbo-instruct	0.73	0.72	0.70	13.0	149.6
Mistral	0.63	0.62	0.61	9.6	104.1
Zephyr	0.51	0.60	0.52	10.1	122.0

- **Quality and completeness of *atomic facts***

Human correlation between model-generated *atomic facts* and human-written *atomic facts* in RoSE (Liu et al., ACL 2023) dataset

$$\frac{1}{N_{data}} \sum_{N_{data}} \frac{1}{N_c} \sum_{i=1}^{N_c} \max_{j=1}^{N_g} (\mathrm{ROUGE}(c_i, g_j))$$

Analysis

- **Size of granularity choice in <u>Granularity Expansion</u>**

Doc. Max Granularity	AggreFact-CNN-FtSota	AggreFact-XSum-FtSota	Avg	s/it
One Sent.	72.2±2.8	66.3±1.9	69.25	2.49
Two Sent.	71.0±3.2	69.3±2.0	70.15	2.53
Three Sent.	**72.6**±3.0	69.3±1.9	70.95	2.64
Four Sent.	72.1±3.1	**70.0**±1.8	**71.05**	2.80

- **Effect of <u>Coreference Resolution</u> of documents and atomic facts**

Atomic Facts	Doc	CNN	XSum	Avg
Original	Original	63.2±2.3	66.4±1.8	64.8
	Coref Resolved	65.7±3.4	**67.8**±2.0	66.7(+*1.95*)
Coref Resolved	Original	66.2±3.4	66.6±1.9	66.4
	Coref Resolved	**72.2**±2.7	66.3±1.9	**69.2**(+*2.85*)

Closing Remarks

- We propose **highly effective** and **strongly interpretable** summarization **factual inconsistency detection system.**

- FIZZ achieves the highest performance on the AggreFact (Tang et al., ACL 2023) benchmark dataset by decomposing summaries into *atomic facts* and adaptively comparing them with multiple documents.

- Additionally, we analyzed the completeness and quality of *atomic facts*, demonstrating through human correlation that factors other than content similarity, which previous studies have emphasized, are equally important.

Contact: plm3332@cau.ac.kr
Code: https://github.com/plm3332/FIZZ