BEMEAE: Moving Beyond Exact Span Match for Event Argument Extraction

Enfa Fane, Md Nayem Uddin, Oghenevovwe Ikumariegbe, Daniyal Kashif, Eduardo Blanco, Steven Corman

Computational Language Understanding Lab (UArizona)

The Center on Narrative, Disinformation and Strategic Influence (ASU)

THE UNIVERSITY OF ARIZONA

Arizona State University

Office of Naval Research Science & Technology

Event Argument Extraction

Last year, Mohammed Rehman was convicted of plotting a bomb attack.

Last year, Mohammed Rehman was **convicted** of plotting a bomb attack.

event trigger

argument → defendant

Last year, **Mohammed Rehman** was convicted of plotting a bomb attack.

argument event trigger

Last year, Mohammed Rehman was convicted of plotting a bomb attack.

Evaluation via Exact Span Match

Gold/ Reference arg: (3,4) ≠ Prediction/Candidate: (5,6)

Let's perform error analysis on the model predictions.

Consider this example.

There are no craters, and the vehicles are largely undamaged which wouldn't be the case if there had been an airstrike, said Konashenkov. Later, the UN also walked back its claim that the convoy was hit by military planes.

*There are no craters, and the vehicles are largely undamaged which wouldn't be the case if there had been an **airstrike**, said Konashenkov. Later, the UN also walked back its claim that the convoy was hit by military planes.*

Here the event is an airstrike. The task? Identify the (alleged) target

*There are no craters, and **the vehicles** are largely undamaged which wouldn't be the case if there had been an airstrike, said Konashenkov. Later, the UN also walked back its claim that the convoy was hit by military planes.*

Here the event is an airstrike. The task? Identify the (alleged) target

According to the annotation, the target is 'the vehicles'.

*There are no craters, and **the vehicles** are largely undamaged which wouldn't be the case if there had been an airstrike, said Konashenkov. Later, the UN also walked back its claim that the **convoy** was hit by military planes.*

Here the event is an airstrike. The task? Identify the (alleged) target

According to the annotation, the target is 'the vehicles'. But the model predicted 'convoy'

*There are no craters, and **the vehicles** are largely undamaged which wouldn't be the case if there had been an airstrike, said Konashenkov. Later, the UN also walked back its claim that the **convoy** was hit by military planes.*

Here the event is an airstrike. The task? Identify the (alleged) target

The reference (gold annotation) is 'the vehicles', but the model predicted 'convoy'

But.. isn't that the same entity as 'the vehicles'?

*There are no craters, and **the vehicles** are largely undamaged which wouldn't be the case if there had been an airstrike, said Konashenkov. Later, the UN also walked back its claim that the **convoy** was hit by military planes.*

Here the event is an airstrike. The task? Identify the (alleged) target

The reference (gold annotation) is 'the vehicles', but the model predicted 'convoy'

But.. isn't that the same entity as 'the vehicles'? … isn't the prediction actually right?

As we continued our analysis we realized,

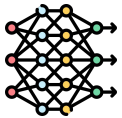
it wasn't just once!

Good predictions were consistently considered wrong.

What if our models are better than we think

but the metric is failing to capture it?

Datasets: RAMS and GENEVA



9 Models via TextEE Framework

PAIE (Ma et al., 2022)
TagPrime-CR (Hsu et al., 2023a)
TagPrime-C (Hsu et al., 2023a)
X-Gear (Huang et al., 2022)
AMPERE (Hsu et al., 2023b)
DEGREE (Hsu et al., 2022)
BART-Gen (Li et al., 2021)
CRF Tagging (Huang et al., 2024)
EEQA (Du and Cardie, 2020)

When does Exact Span Match fail?

When does Exact Span Match fail?

- **Identical Mentions** marked incorrect due to position in text

reference

Clinton is recovering well with antibiotics and rest [...]

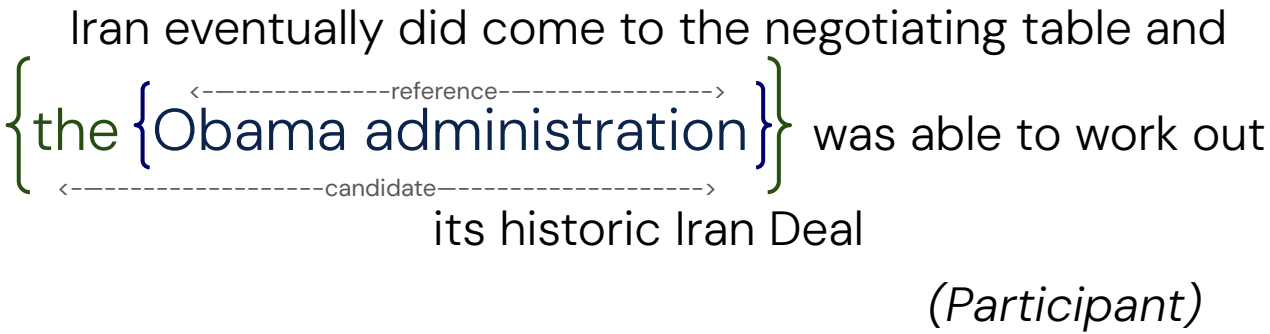
candidate

Clinton's aides say she 'll return to the campaign [...]

(Victim)

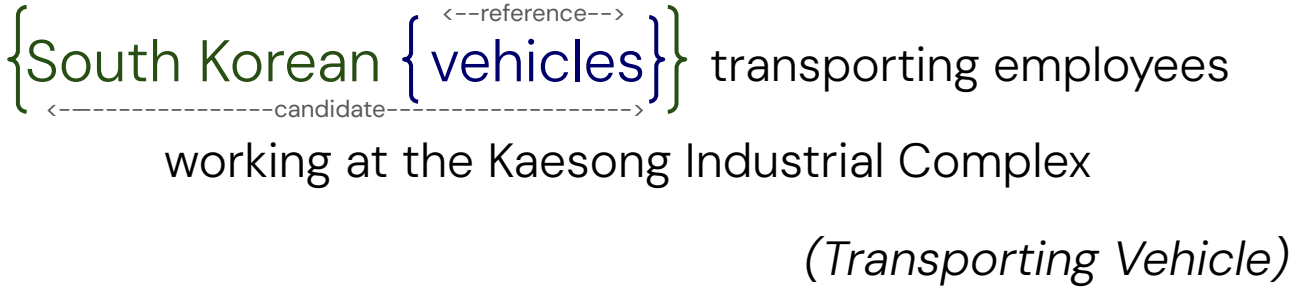
When does Exact Span Match fail?

- **Harmless Tokens** (e.g., determiners, punctuation) unfairly penalized



When does Exact Span Match fail?

- **Relevant additional information** is considered incorrect



When does Exact Span Match fail?

- **Alternative Mentions** like coreference and metonymy not credited

Assange has characterized the investigation as part of a broader conspiracy on the part of the 

{ { U.S. } government } to incarcerate him
<candidate>

(Jailer)

- **Aggregated Lists** treated as mismatches

[...] accused Hezbollah of recruitment to carry out terrorist attacks ,

and of smuggling weapons and explosives

(Smuggled Artifact)

Reference : (1) weapons and explosives

Candidate: (1) weapons (2) explosives

When does Exact Span Match fail?

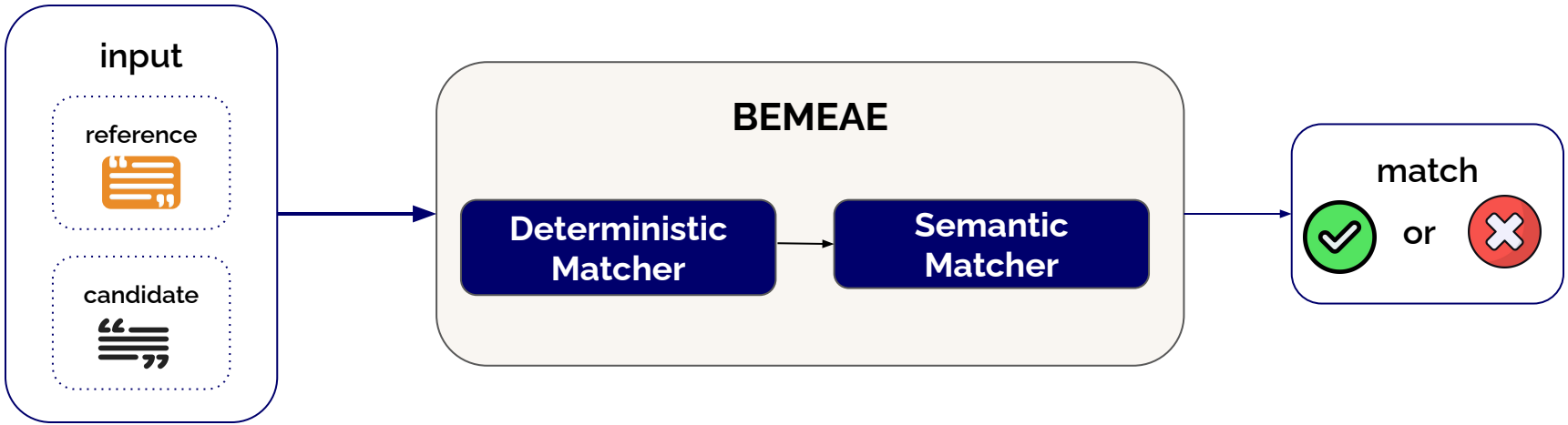
- **Identical Mentions**
- **Harmless Tokens**
- **Relevant additional information**
- **Alternative Mentions**
- **Aggregated Lists**

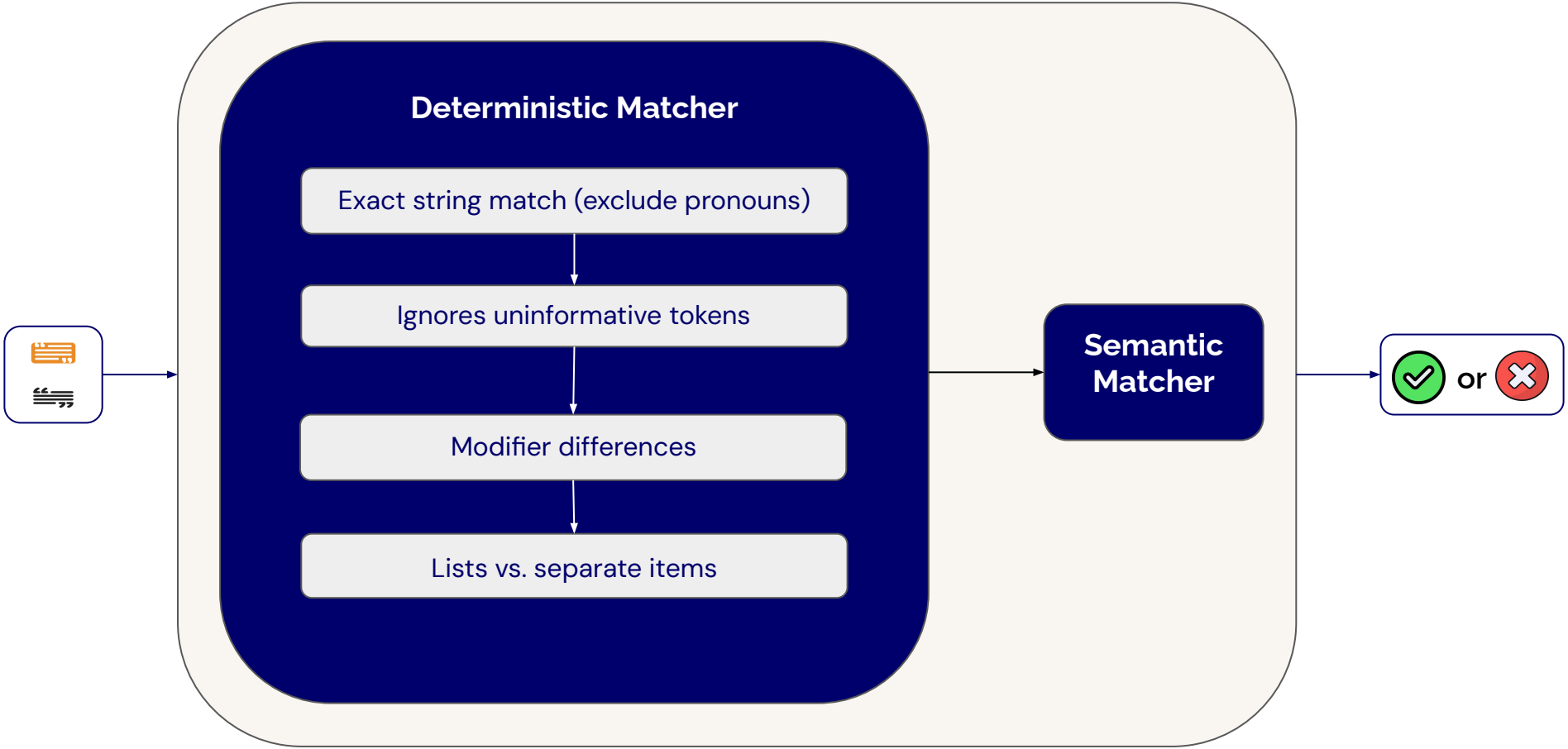
We need a metric that better reflects human judgments

Introducing

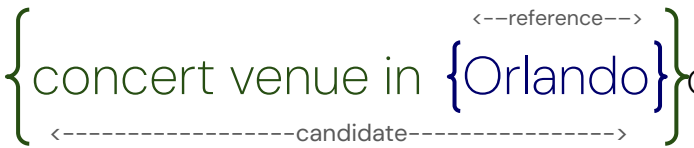
BEMEAE

Beyond Exact Span Match for Event Argument Extraction


input
reference
candidate
BEMEAE
Deterministic Matcher
Semantic Matcher
match
or


Deterministic Matcher
Exact string match (exclude pronouns)
Ignores uninformative tokens
Modifier differences
Lists vs. separate items
Semantic Matcher
or

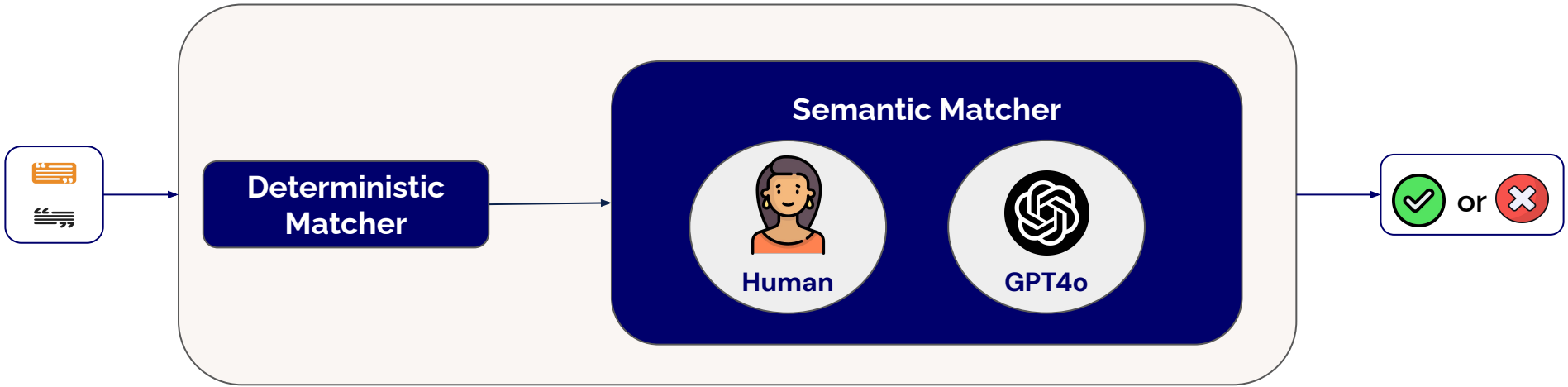
Christina Grimmie, 22, [..] was signing autographs at a



on 10 June when an

assailant shot her

(Place of Attack)

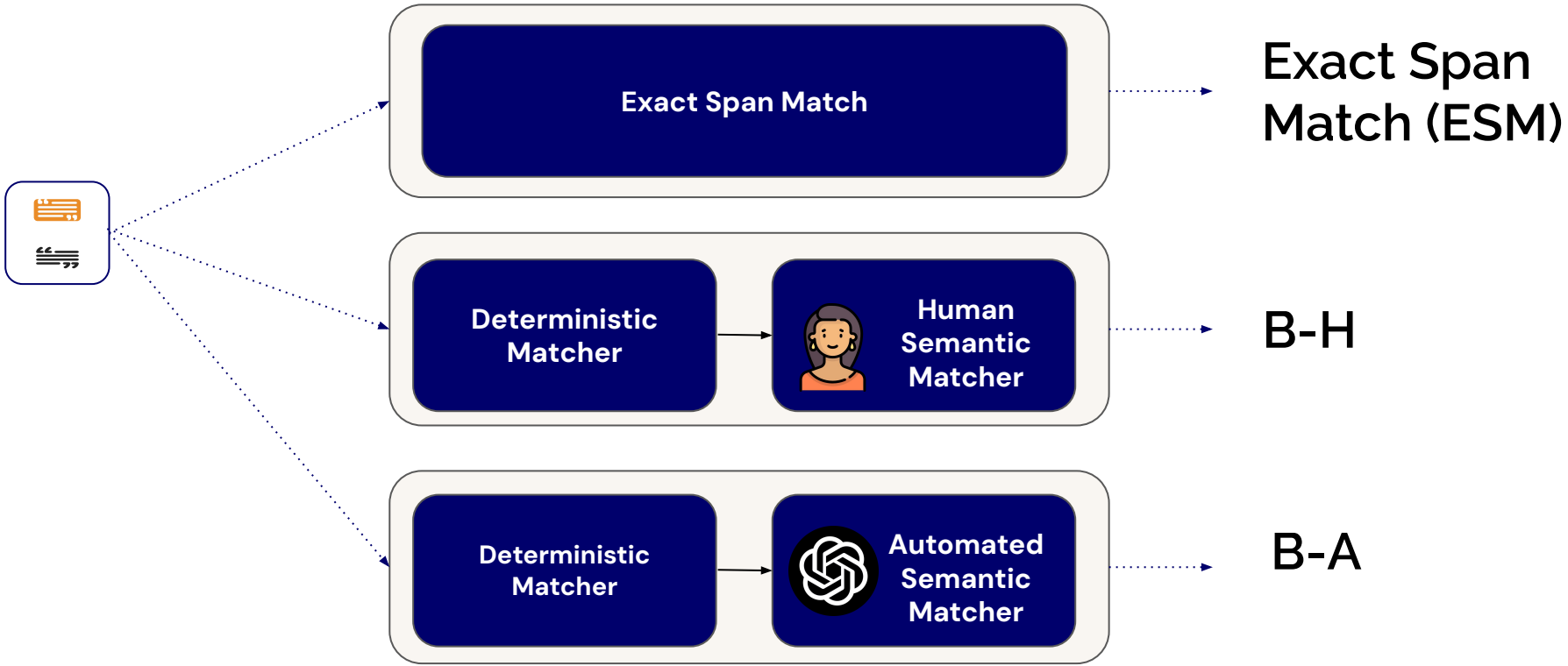


500 most predicted candidate arguments, deemed incorrect by Exact Span Match and not matched by deterministic components, were manually annotated.

High Human agreement

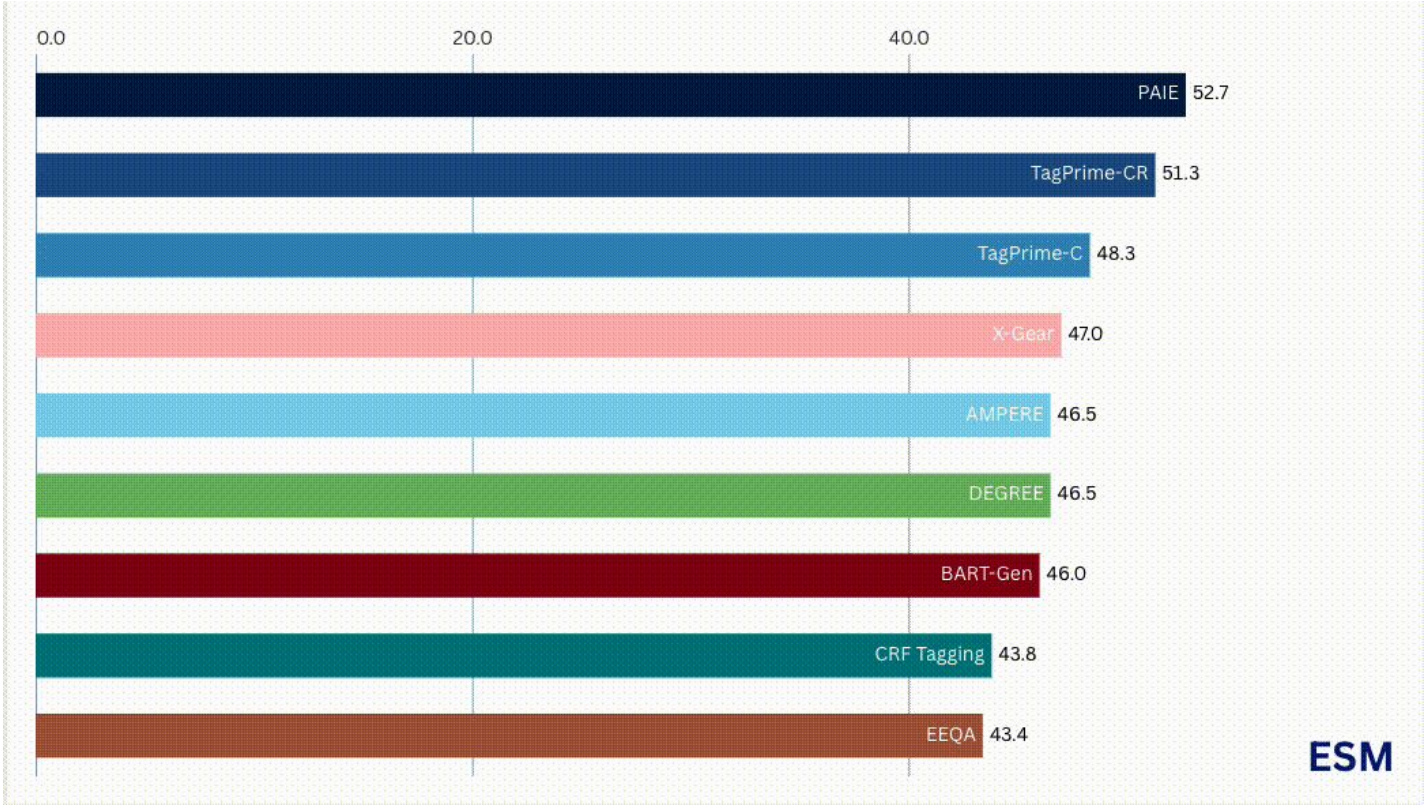
GPT-4o outperformed all other semantic matchers tested, moderate agreement with humans

Evaluation Setup

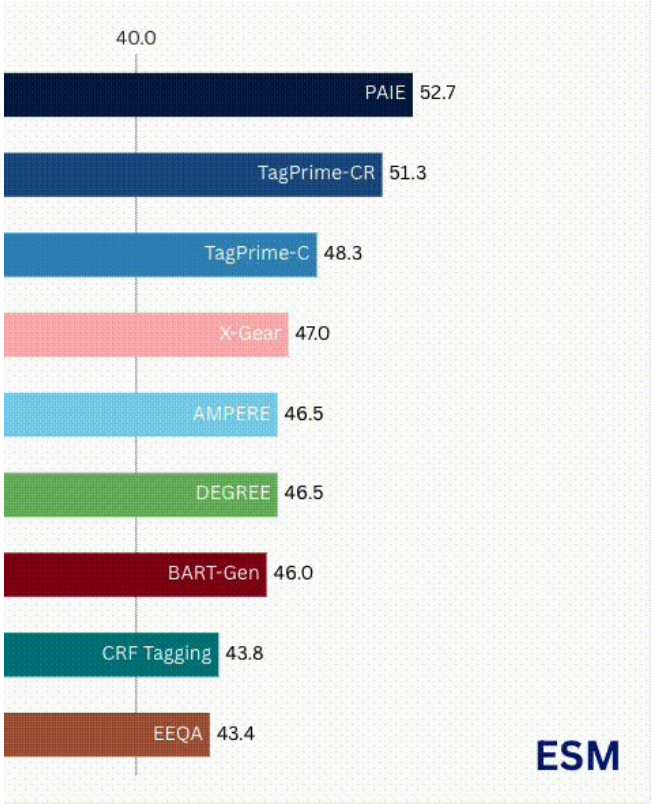


What is the model performance under these metrics?

BEMEAE Rewrites the Leaderboard



Higher F1 overall



Note: Graph truncated to emphasize differences

Higher F1 overall, but gains are uneven

System	Δ F1
PAIE	10.2
TagPrime-CR	8.7
TagPrime-C	10.1
X-Gear	10.6
AMPERE	12.2
DEGREE	12.9
BART-Gen	14.9
CRF Tagging	9.7
EEQA	6.1

Uneven gains lead to rank changes

System	Δ rank
PAIE	
TagPrime-CR	-1
TagPrime-C	-3
X-Gear	-3
AMPERE	
DEGREE	2
BART-Gen	5
CRF Tagging	
EEQA	

Uneven gains lead to rank changes

System	Δ rank
PAIE	
TagPrime-CR	-1
TagPrime-C	-3
X-Gear	-3
AMPERE	
DEGREE	2
BART-Gen	5
CRF Tagging	
EEQA	

- Exact Span Match ranks do not correlate well with BEMEAE-Human ranks (Kendall's $\tau = 0.44$)

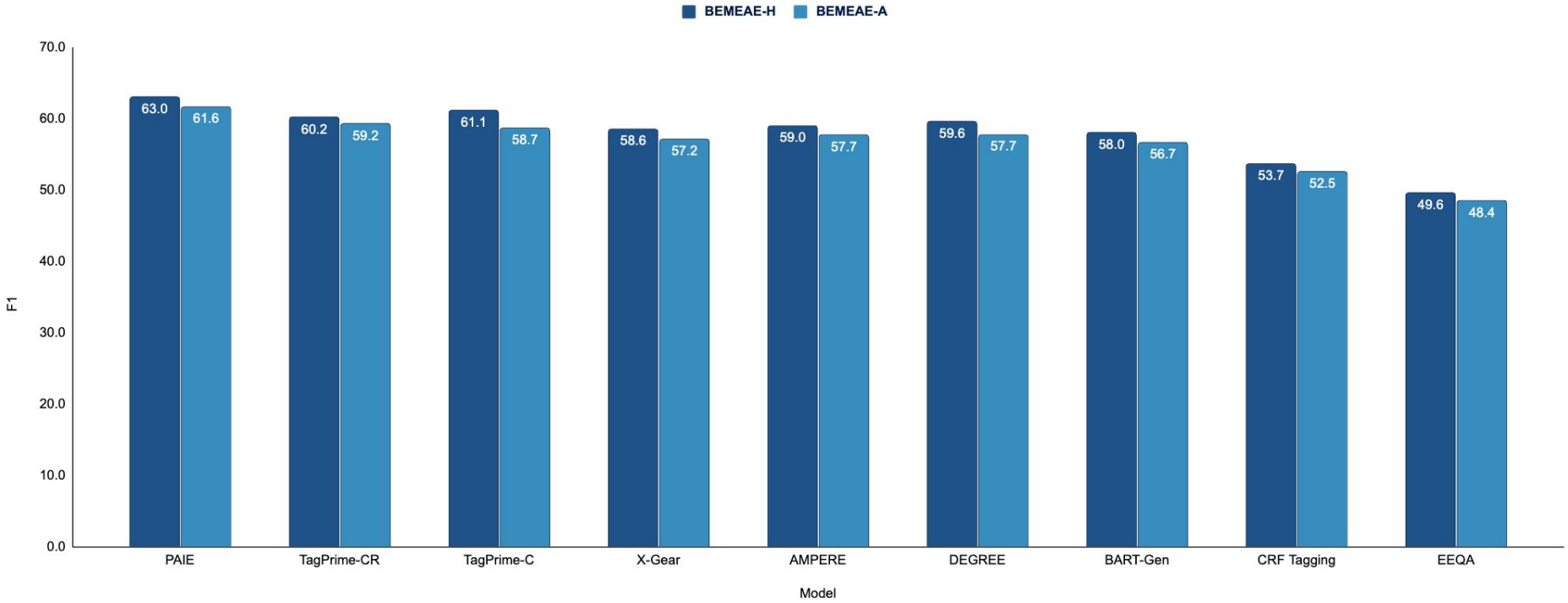
- Exact Span Match ranks do not correlate well with BEMEAE-Human ranks (Kendall's $\tau = 0.44$)

BEMEAE > Exact Span Match

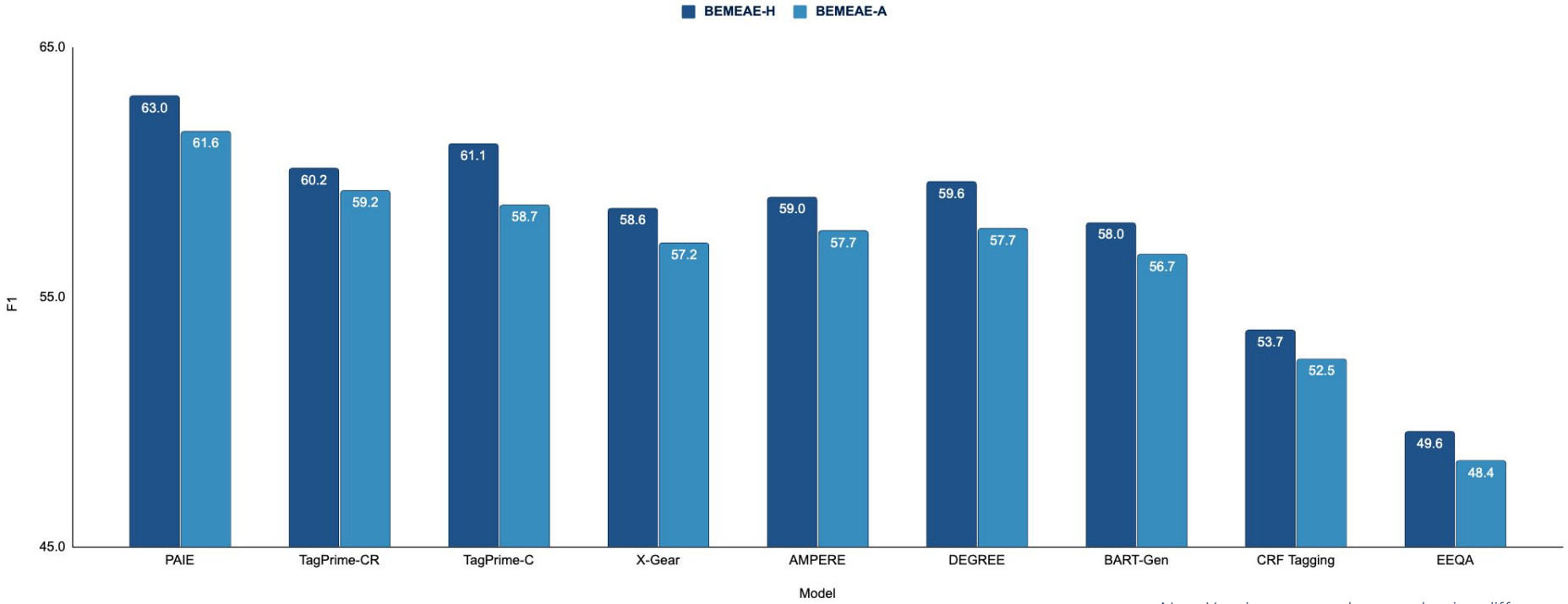
But human as semantic matchers is not scalable.

How does BEMEAE-A compare?

BEMEAE-A follows BEMEAE-H closely

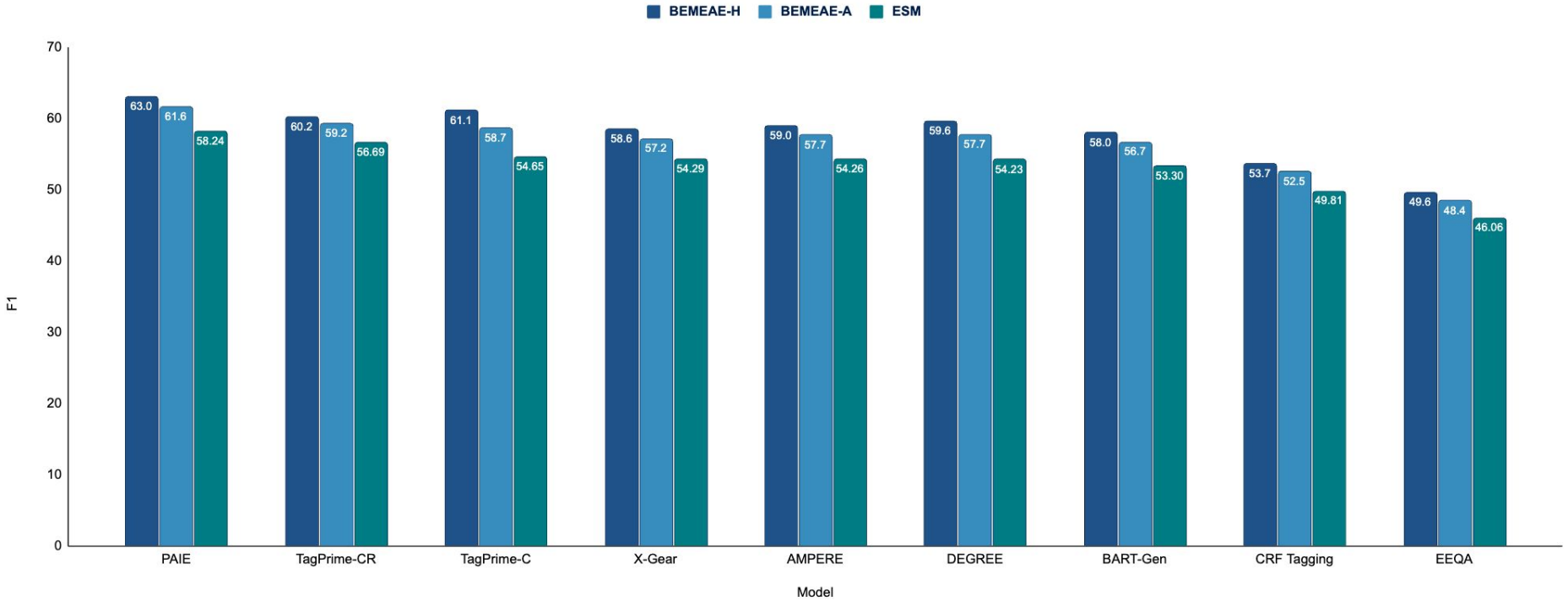


BEMEAE-A follows BEMEAE-H closely

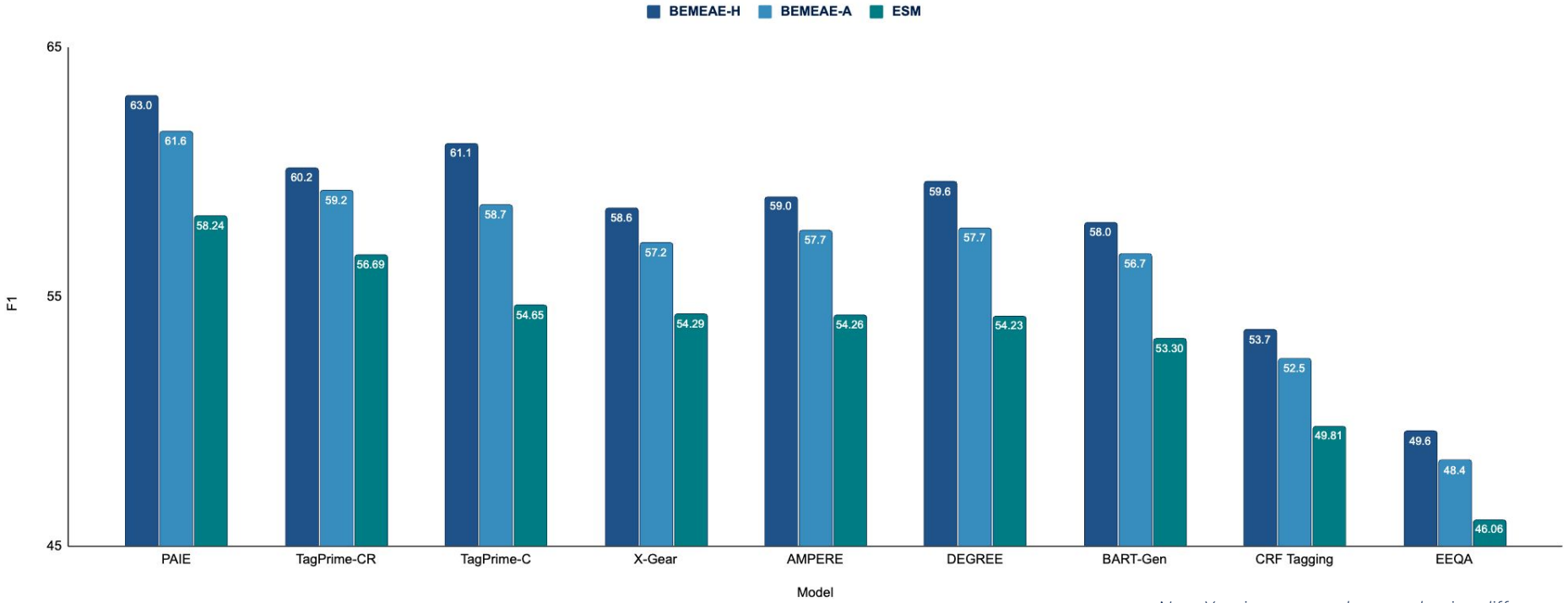


Note:Y-axis truncated to emphasize differences

BEMEAE-A follows BEMEAE-H closely, Exact Span Match less so



BEMEAE-A follows BEMEAE-H closely, Exact Span Match less so



Note:Y-axis truncated to emphasize differences

BEMEAE-A closer to BEMEAE-H

1. **Moderate agreement with human annotations** (Cohen's $\kappa = 0.43$)
2. **Strong correlation with BEMEAE-H model rankings** (Kendall's $\tau = 0.94$)
3. **F1 difference by 2-3 points**

For human-aligned evaluation,

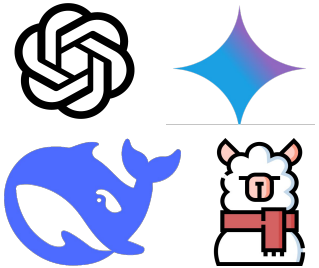
BEMEAE-H > BEMEAE-A > Exact Span Match

Is BEMEAE the perfect metric?

- No, But it is a significant improvement over Exact Span Match.
- BEMEAE partly largely depends on the Semantic Matcher used.
- Better semantic matchers enable truer, more human-aligned evaluation.

Evaluation in the era of LLMs

- LLM-based model predictions may exhibit greater surface-level variability
- Potensial underestimation of performance when using Exact Span Match
- BEMEAE offers a more robust and human-aligned evaluation.



BEMEAE: Moving Beyond Exact Span Match for Event Argument Extraction

Enfa Fane, Md Nayem Uddin, Oghenevovwe Ikumariegbe, Daniyal Kashif, Eduardo Blanco, Steven Corman

BEMEAE: Moving Beyond Exact Span Match for Event Argument Extraction

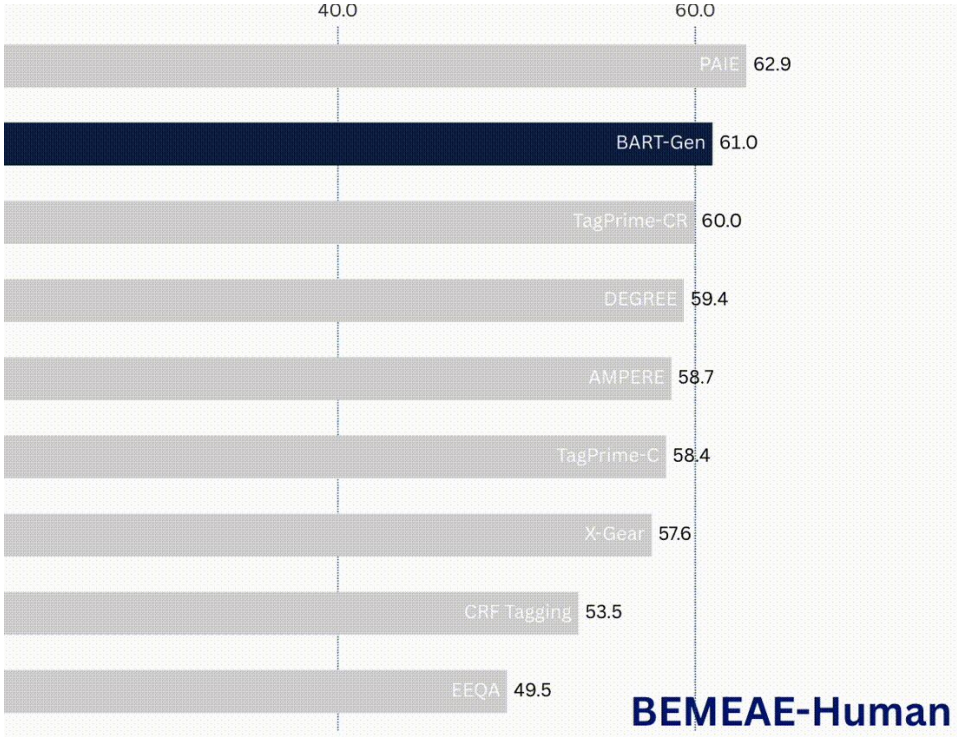
Enfa Fane, Md Nayem Uddin, Oghenevovwe Ikumariegbe, Daniyal Kashif, Eduardo Blanco, Steven Corman

- This research was supported by the U.S. Office of Naval Research (Grant No. N00014-22-1-2596.).
- We thank creators of the TEXTEE framework, which enabled fair and consistent model comparisons.
- Special thanks to our research team at Arizona State University for their valuable contributions.
- Icons in this presentation are from Flaticon.com

Is Exact Span Match truly reflecting your model's performance,
or underestimating it

…like it did for BART-Gen?

We invite you to find out



Note: Graph truncated to emphasize differences

Evaluate your EAE model
with BEMEAE today!

github.com/beingenfa/bemeae

enfafane@gmail.com

Scan QR Code

BEMEAE: Moving Beyond Exact Span Match for Event Argument Extraction

Enfa Fane, Md Nayem Uddin, Oghenevovwe Ikumariegbe, Daniyal Kashif,
Eduardo Blanco, Steven Corman

THE UNIVERSITY OF ARIZONA

Arizona State University

Computational Language Understanding

Office of Naval Research Science & Technology