DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, Smaranda Muresan

Motivation and Background

- Why do we need fact-checking?

  Misinformation/disinformation on the rise (Ireton, 2018)

  Journalists must examine “check-worthy” claims for truth (Hassan et al, 2017)

- Related Work

  Fact-checking websites (Vlachos and Reidel, 2014; Alhindi et al., 2018)

  News articles (Pomerleau and Rao, 2017)

  Community forums (Mihaylova et al., 2018)

Challenges

1) Professional fact-checking organizations need to synthesize evidence from **multiple sources** (Graves, 2018)

2) Numerical comparatives alone are indicative of truthful statements (Rashkin et al., 2017) but **temporal reasoning** is challenging (Mirza and Tonnelli, 2016)

3) Fact-checking is sensitive to **ambiguous entities** (Thorne and Vlachos, 2018) or **lexical features** (Nakashole and Mitchell, 2014)

Towards Addressing these Challenges

Data

Development of adversarial claims

Methods

Development of fact-checking system with targeted improvements

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Data

Murda Beatz's real name is Marshall Mathers.

[Murda Beatz] 

Shane Lee Lindstrom, known professionally as Murda Beatz ...

REFUTES

Data

FEVER 1.0 (Thorne et al. 2018)

FEVER 2.0 (Thorne et al. 2019)

- **adversarial** "attacks" submitted by participants
- evaluated against FEVER systems

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Data: Adversarial Claims

1) Multiple propositions
    a) CONJUNCTION

Janet Leigh was from New York. Janet Leigh was an author.

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Data: Adversarial Claims

1) Multiple propositions
    a) CONJUNCTION

Janet Leigh was from New York **and** ~~Janet Leigh~~ was an author.

Data: Adversarial Claims

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING

[The Nice Guys] 

The Nice Guys is a 2016 action comedy film.

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Data: Adversarial Claims

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING

[The Nice Guys] 

The Nice Guys is a 2016 action comedy film **directed by a**

Danish screenwriter known for the 1987 action film Lethal Weapon.

[Shane Black] 

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Data: Adversarial Claims

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS

Duff McKagan is an American citizen.

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Data: Adversarial Claims

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS

Duff McKagan is an American citizen **born in Seattle.**

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Data: Adversarial Claims

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS
2) Temporal reasoning
   a) DATE MANIPULATION

   Artpop was Gaga's second consecutive number one record in the United States in 2009.

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Data: Adversarial Claims

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS
2) Temporal reasoning
   a) DATE MANIPULATION

Artpop was Gaga's second consecutive number one record in the United States ~~in 2009~~ **before 2010**.

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Data: Adversarial Claims

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS
2) Temporal reasoning
   a) DATE MANIPULATION
   b) MULTI-HOP TEMPORAL REASONING

[William Henry Harrison] **[Indiana Territory]**

The first governor of the Indiana Territory lived long enough to see it become a state.

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Data: Adversarial Claims

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS
2) Temporal reasoning
   a) DATE MANIPULATION
   b) MULTI-HOP TEMPORAL REASONING

[William Henry Harrison] **[Indiana Territory]**

The first governor of the Indiana Territory lived long enough to see it become a state. (death 1841)

(admittance 1816)

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Data: Adversarial Claims

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS
2) Temporal reasoning
   a) DATE MANIPULATION
   b) MULTI-HOP TEMPORAL REASONING
3) Ambiguity and lexical variation
   a) ENTITY DISAMBIGUATION

Kate Hudson is a left wing political activist.

[Kate Hudson] 

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Data: Adversarial Claims

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS
2) Temporal reasoning
   a) DATE MANIPULATION
   b) MULTI-HOP TEMPORAL REASONING
3) Ambiguity and lexical variation
   a) ENTITY DISAMBIGUATION

Kate Hudson is a left wing political activist.

[Kate Hudson] [Kate Hudson (activist)] 

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Data: Adversarial Claims

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS
2) Temporal reasoning
   a) DATE MANIPULATION
   b) MULTI-HOP TEMPORAL REASONING
3) Ambiguity and lexical variation
   a) ENTITY DISAMBIGUATION
   b) LEXICAL SUBSTITUTION

   The Last Song began filming in 2009.

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

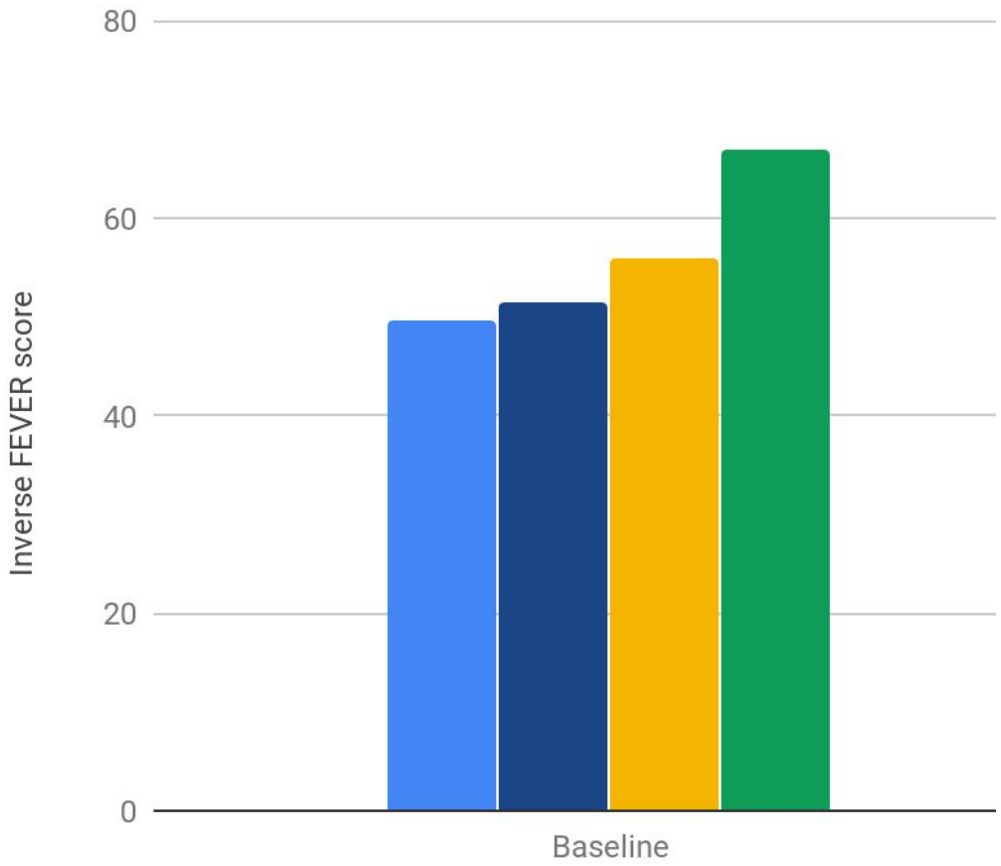
Data: Adversarial Claims

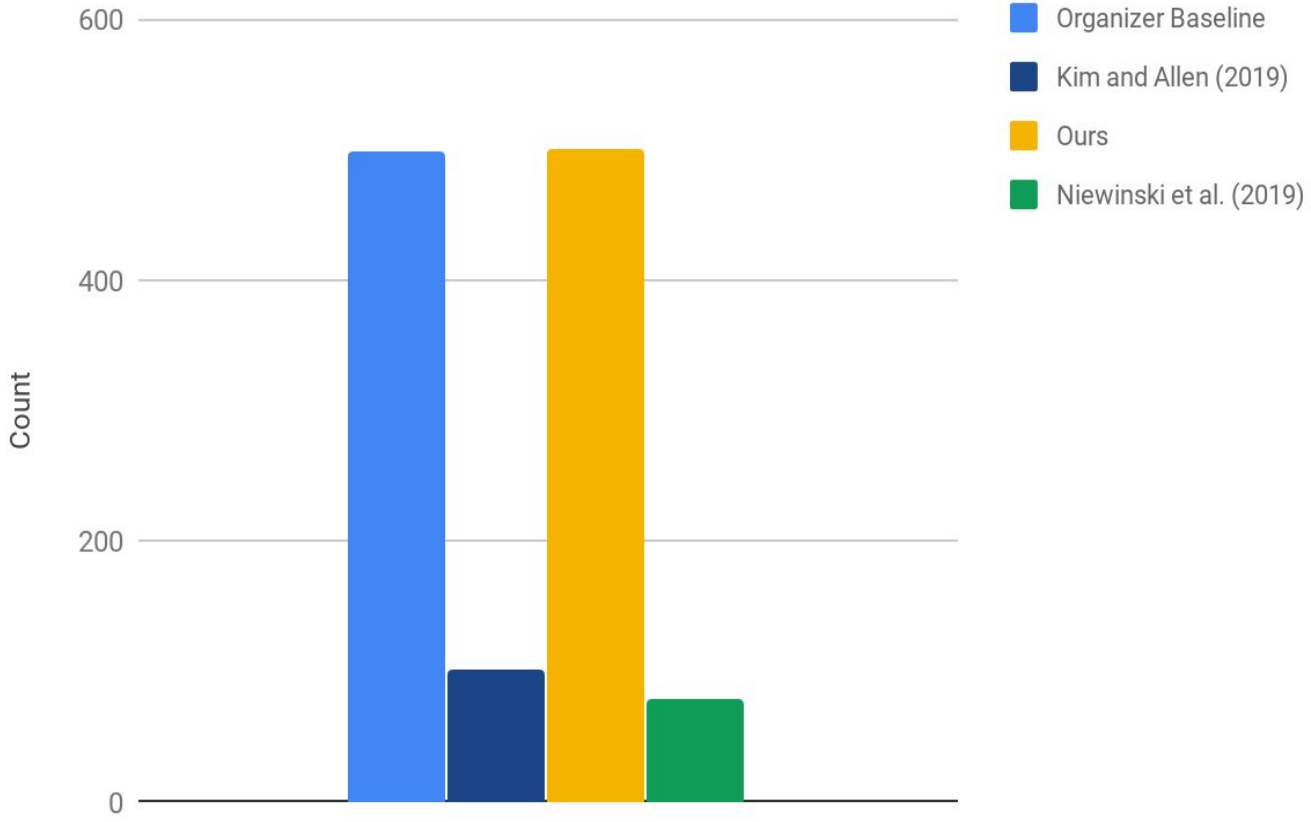
1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS
2) Temporal reasoning
   a) DATE MANIPULATION
   b) MULTI-HOP TEMPORAL REASONING
3) Ambiguity and lexical variation
   a) ENTITY DISAMBIGUATION
   b) LEXICAL SUBSTITUTION

The Last Song began ~~filming~~ **shooting** in 2009.

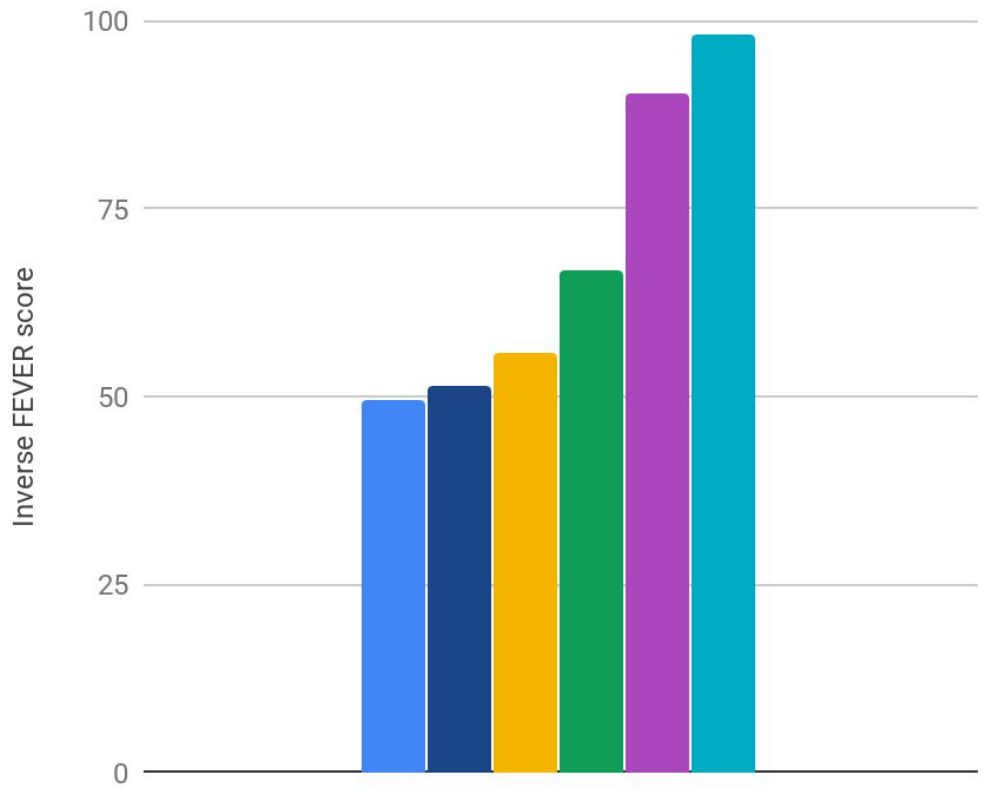
DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

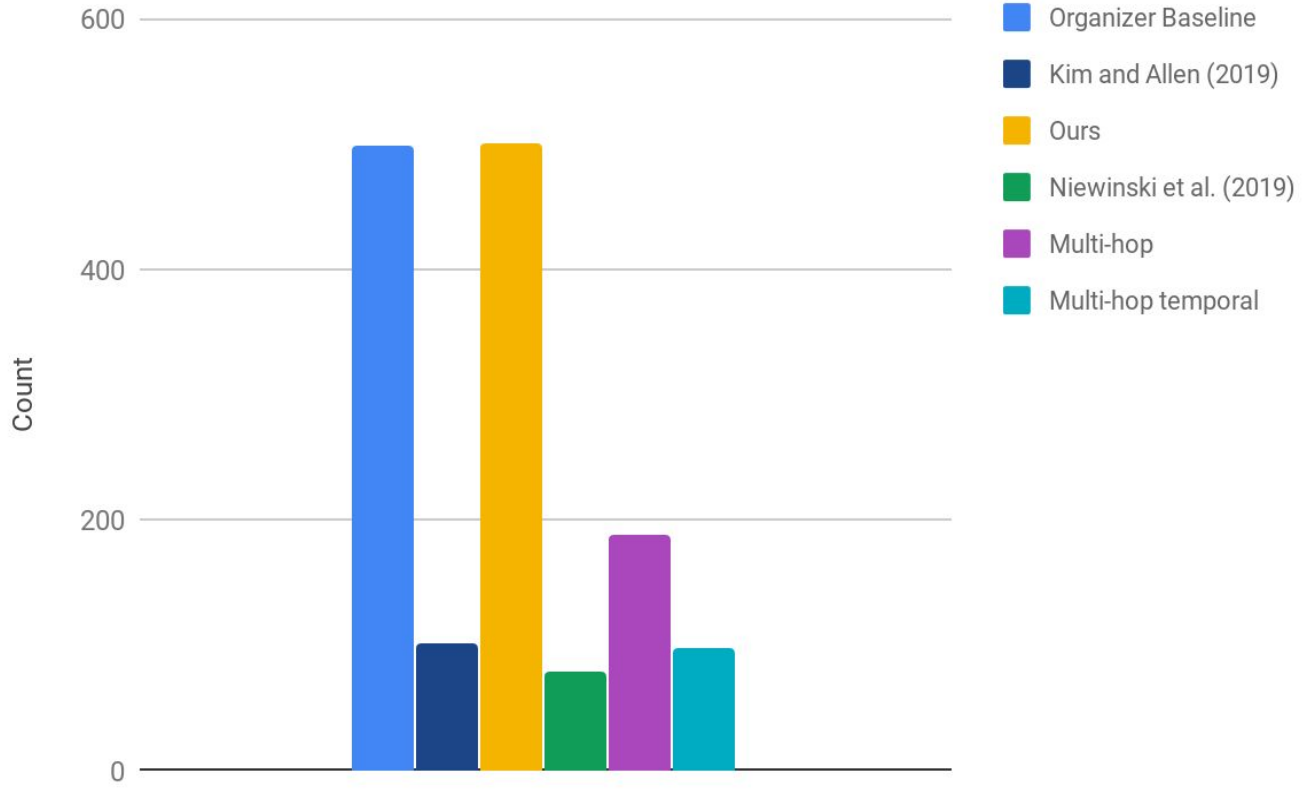
Evaluation: Data



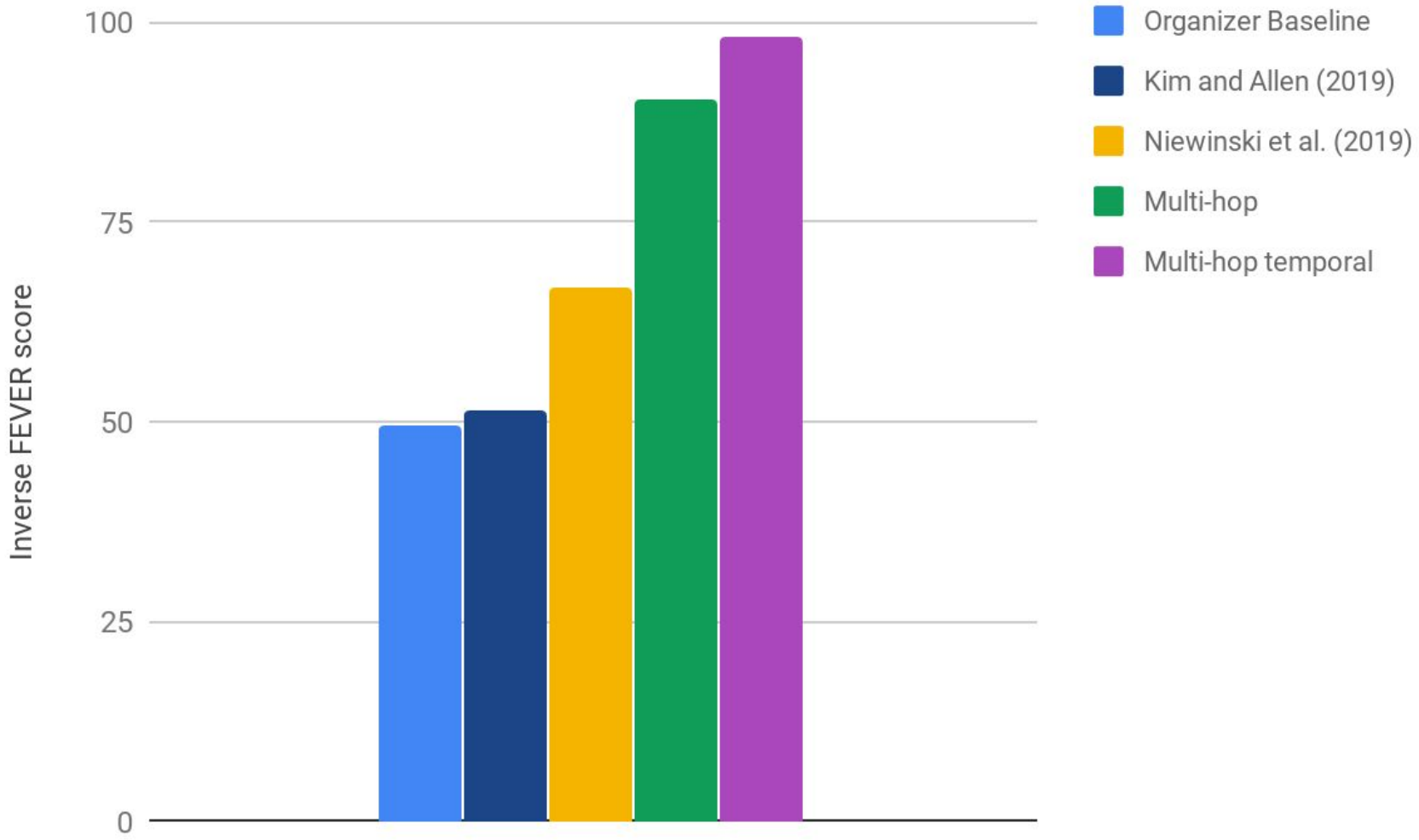


Evaluation: Data

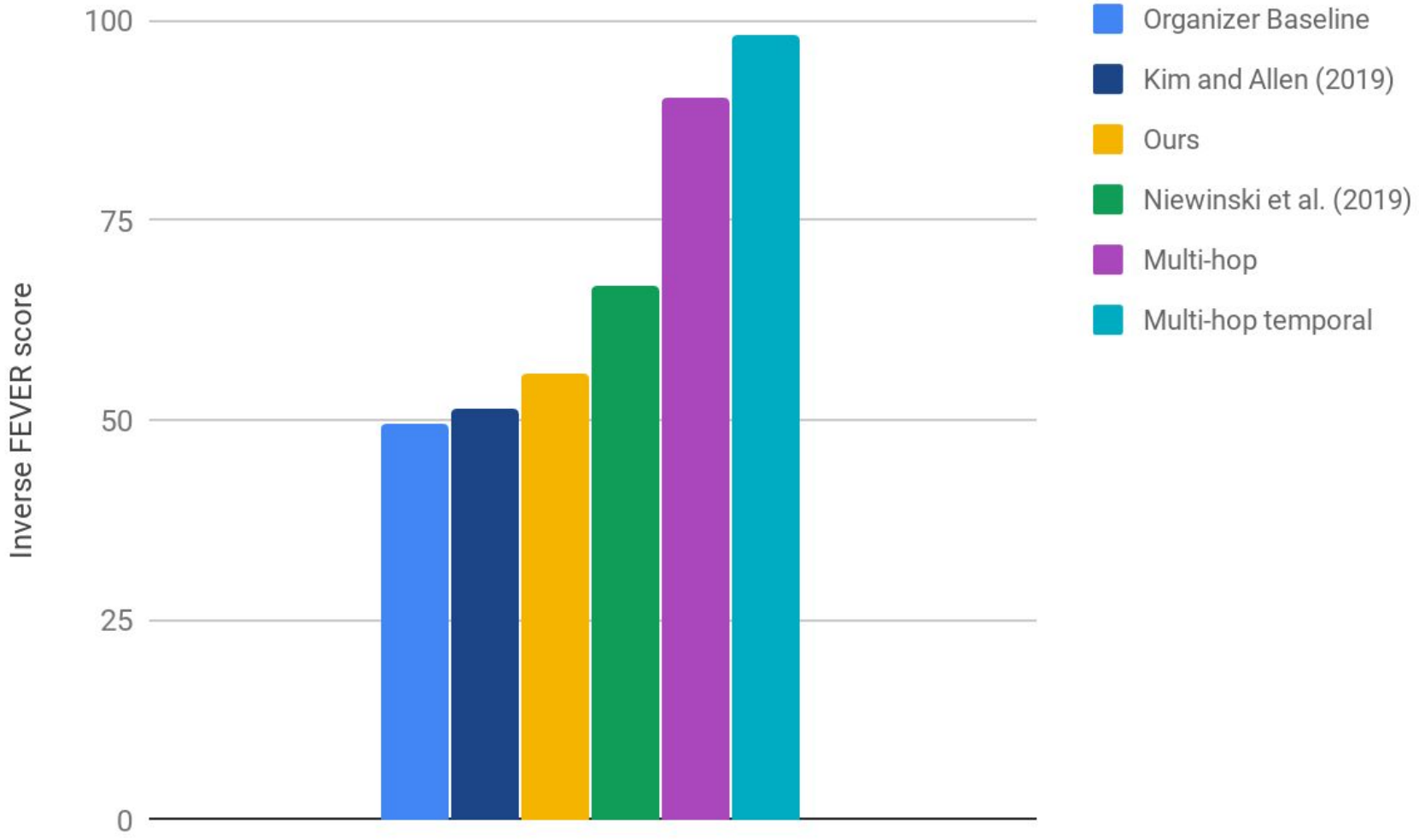




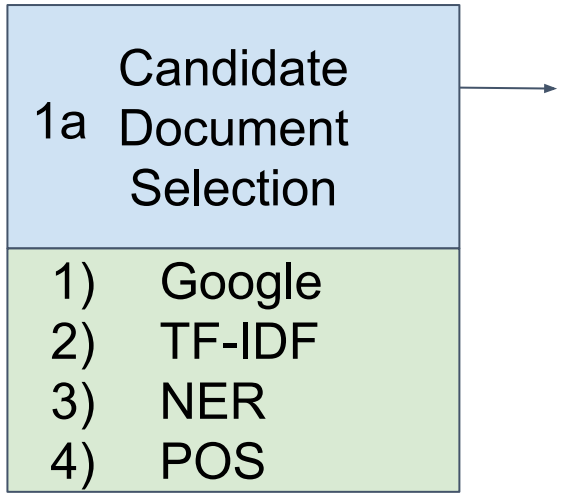
Evaluation: Data



Evaluation: Data

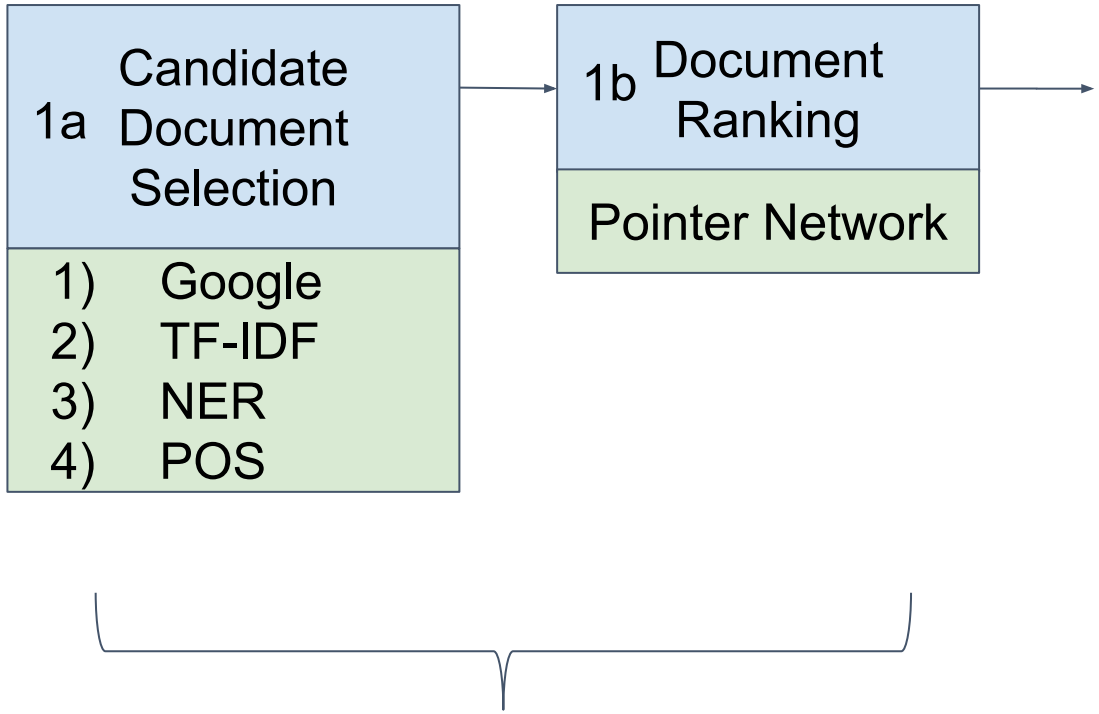


Methods: Fact-Checking System



DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

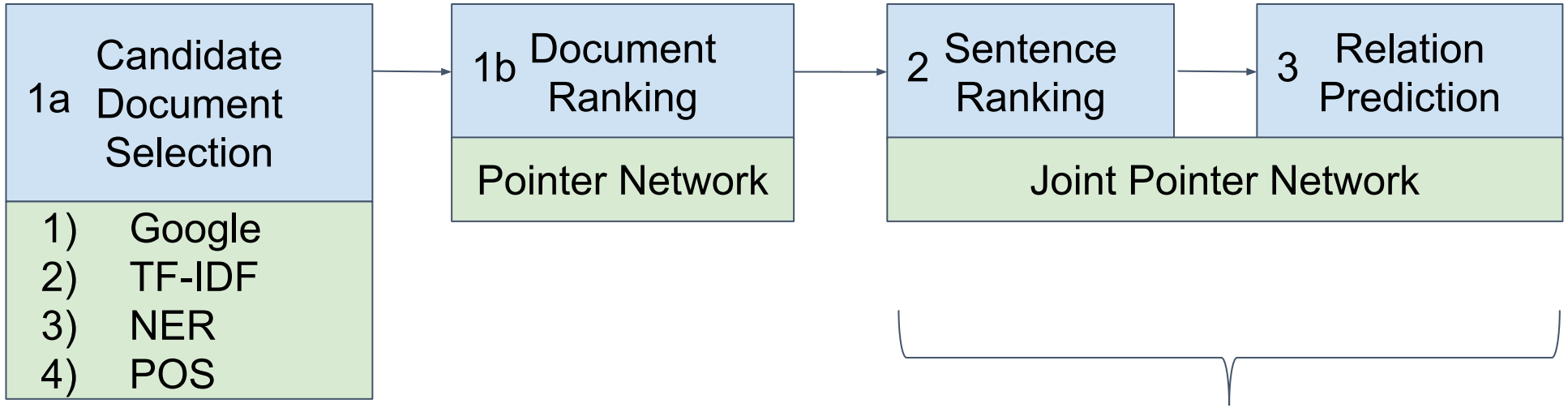
Methods: Fact-Checking System



Overgenerate and re-rank to handle ambiguity

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Methods: Fact-Checking System



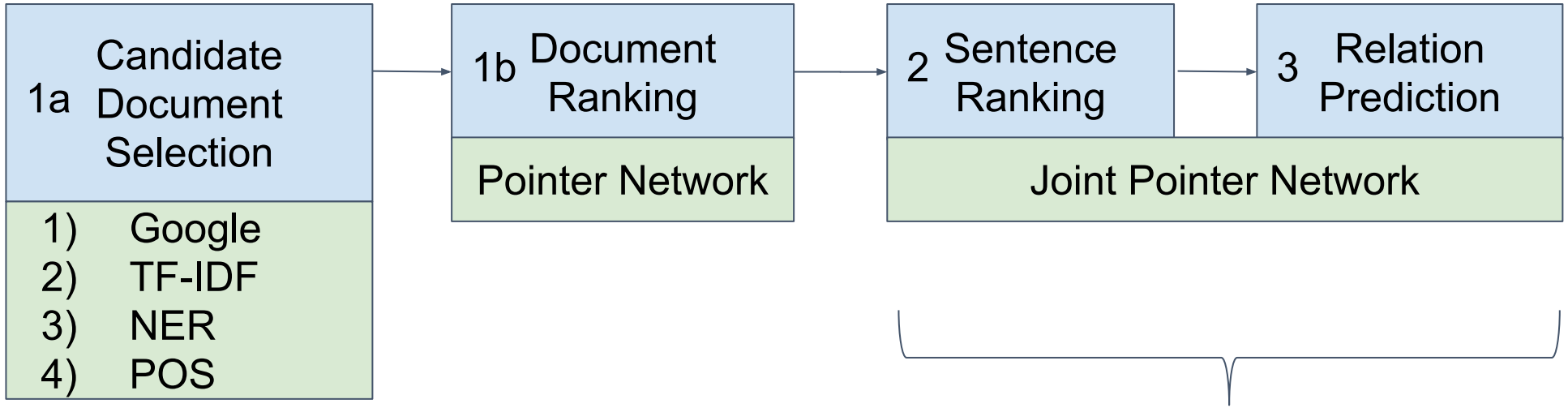
Overgenerate and re-rank to handle ambiguity

Sequence prediction to handle multiple propositions

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Methods: Fact-Checking System

Post-processing to handle temporal relations

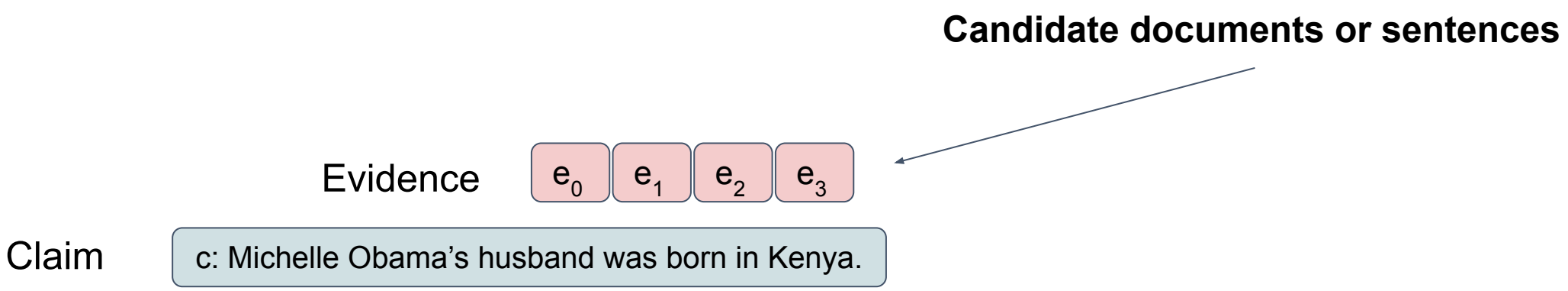


Sequence prediction to handle multiple propositions

Overgenerate and re-rank to handle ambiguity

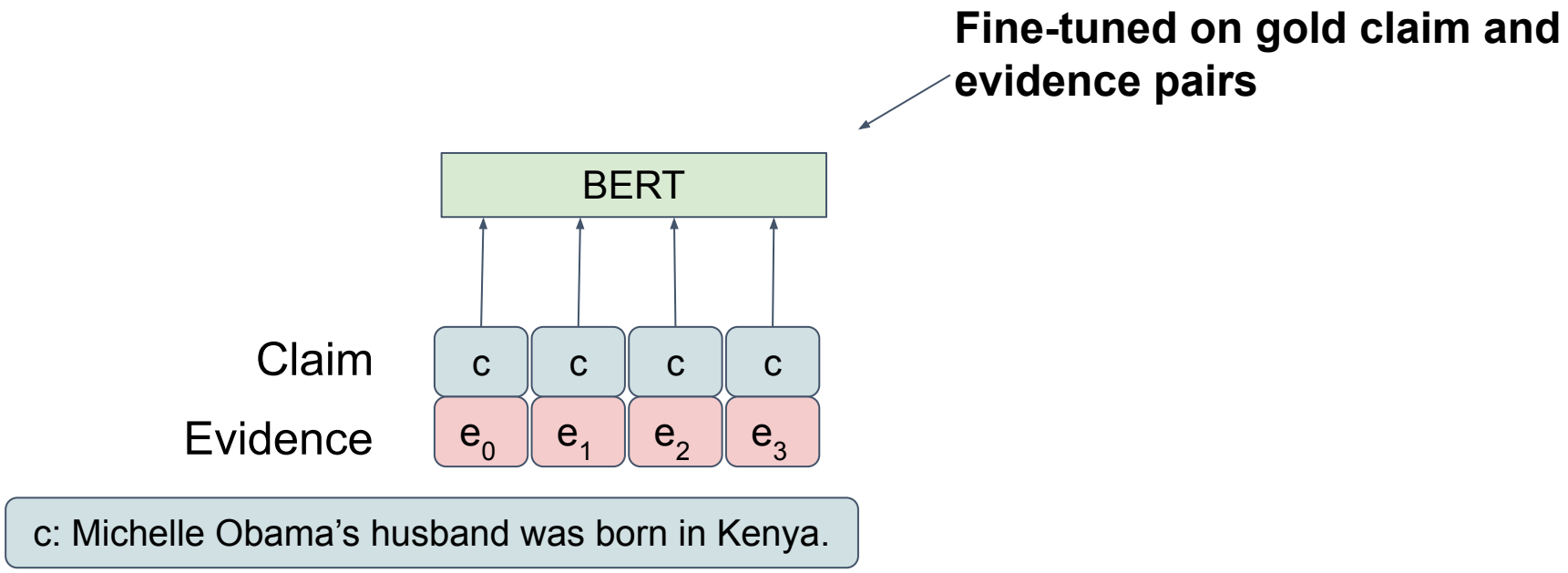
DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Methods: Pointer Network



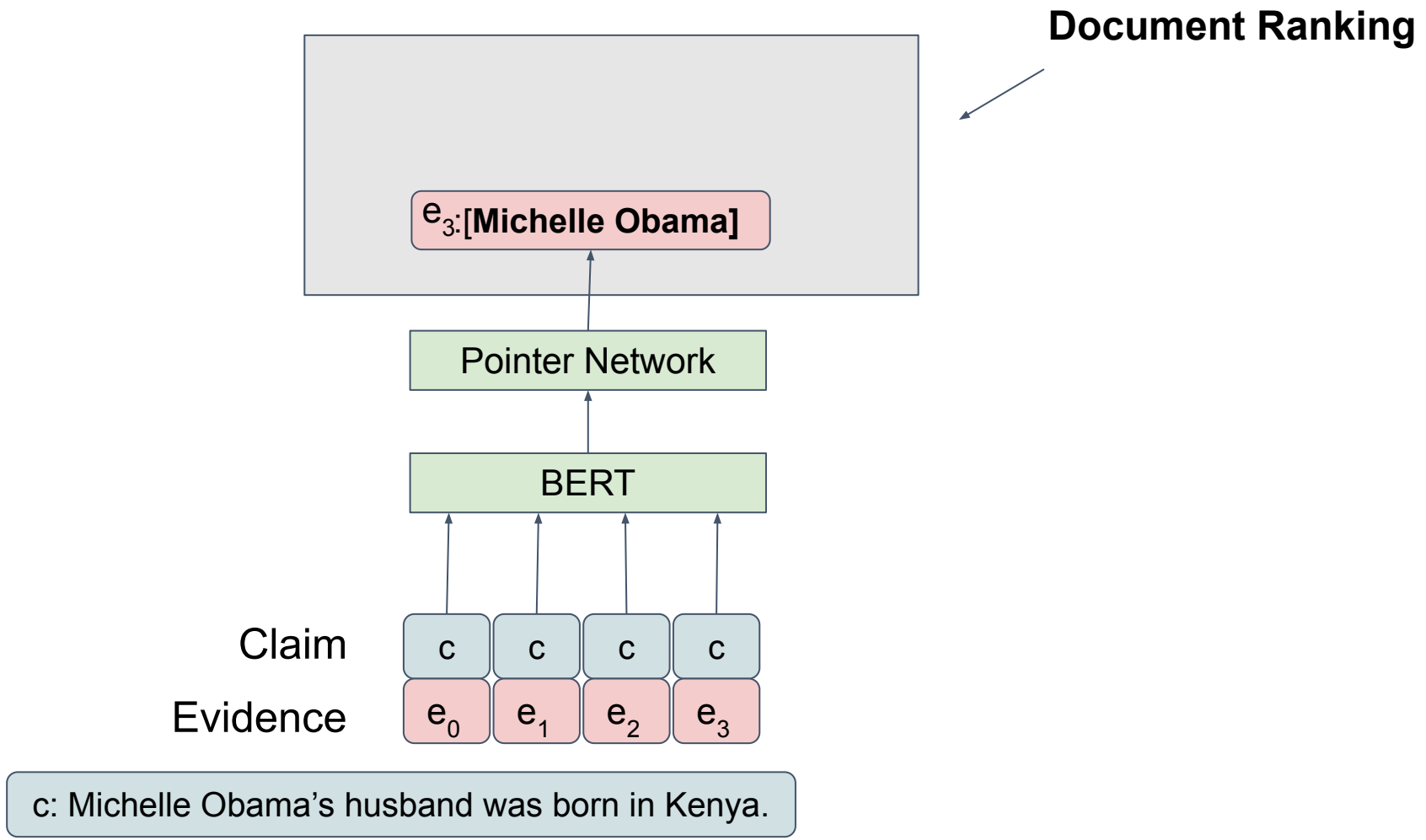
DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Methods: Pointer Network



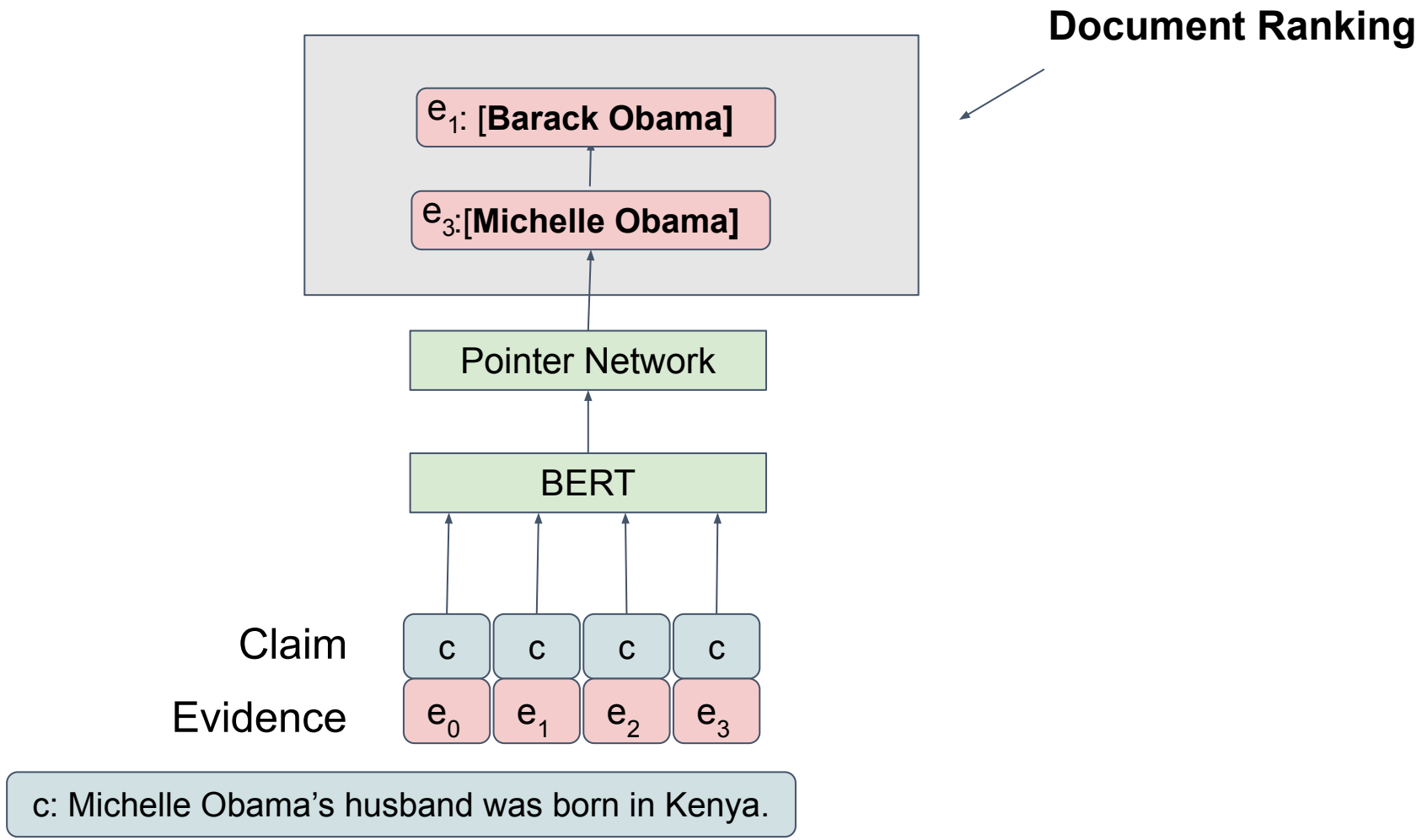
DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Methods: Pointer Network



DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

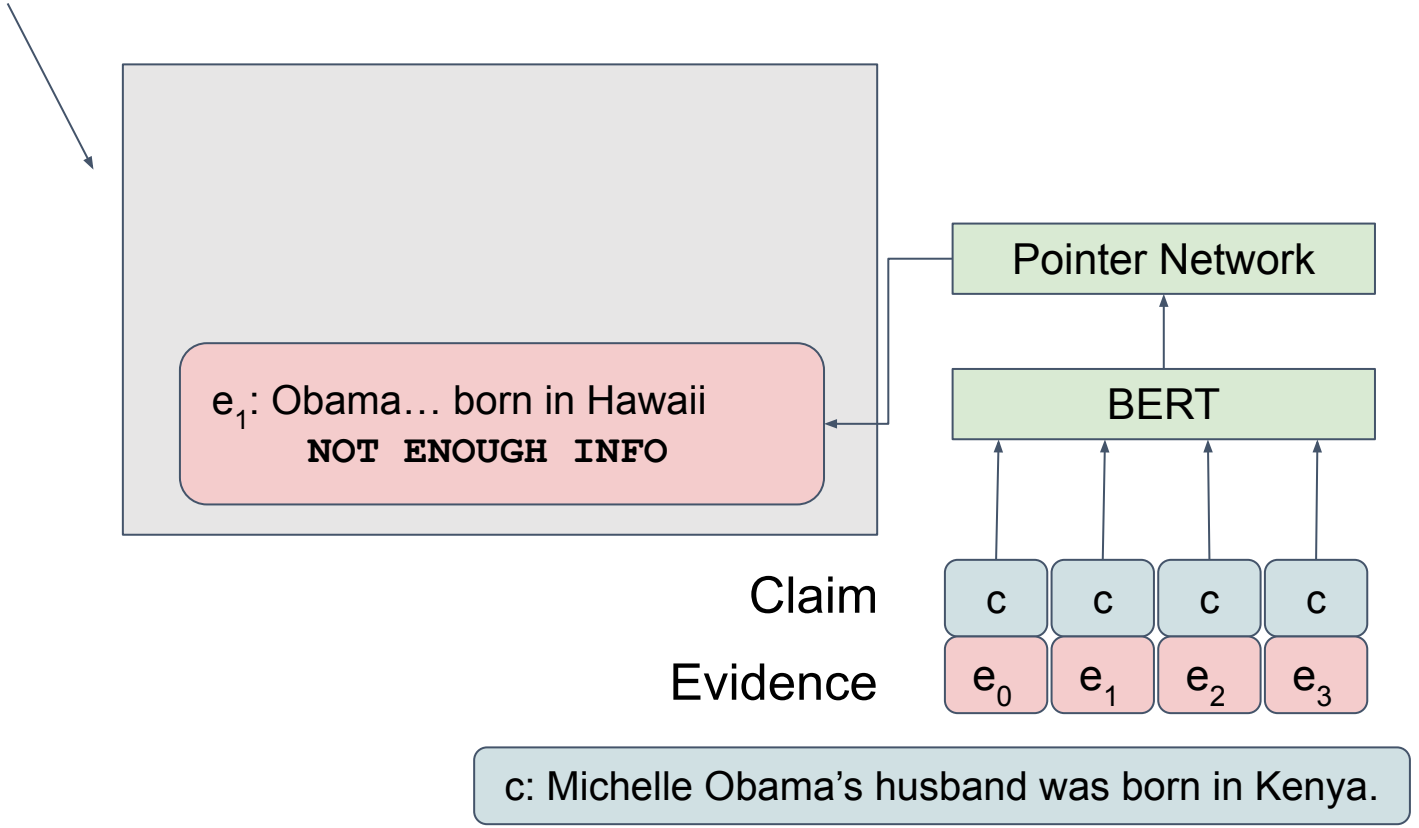
Methods: Pointer Network



DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Methods: Pointer Network

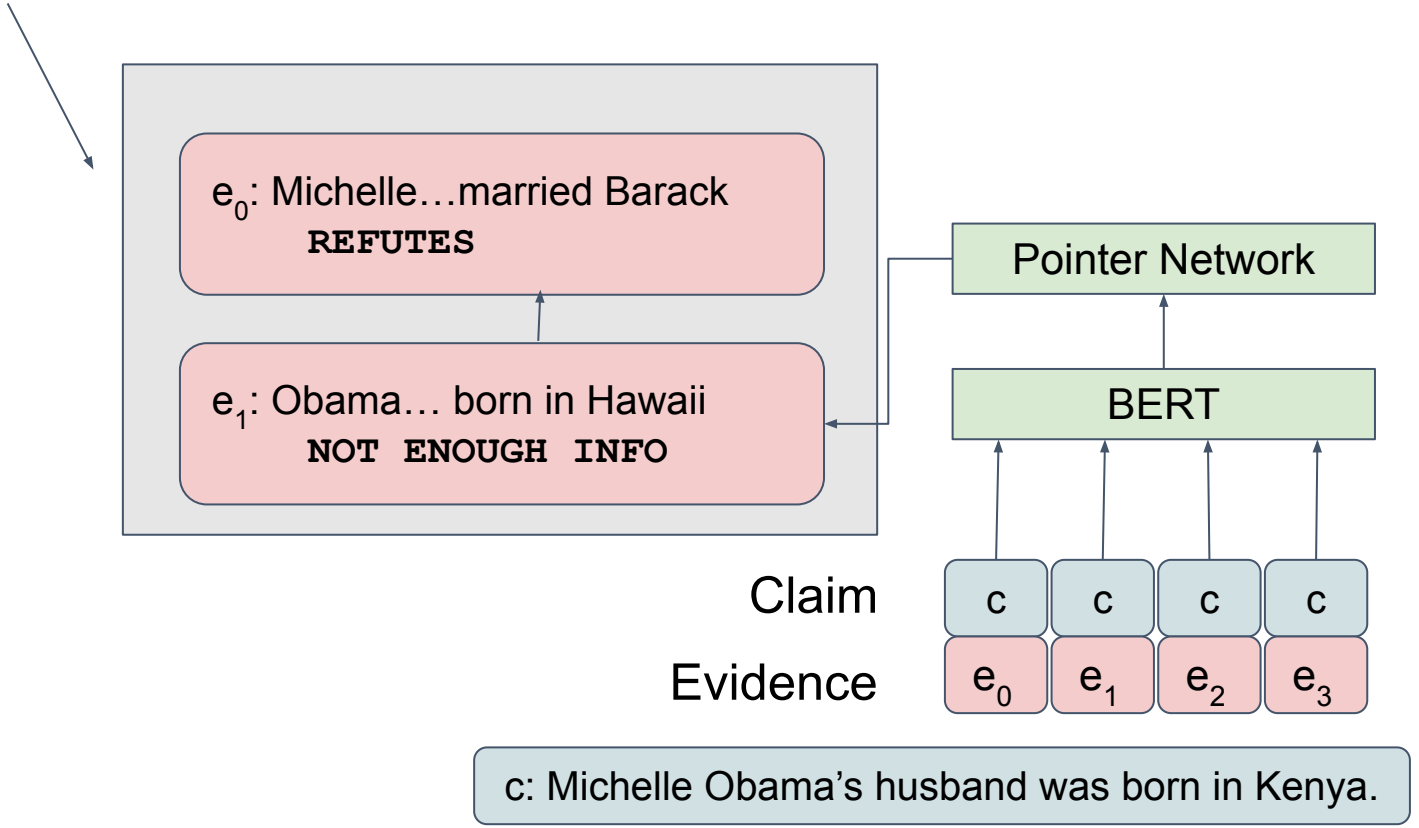
Joint Sentence Selection and Relation Prediction



DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

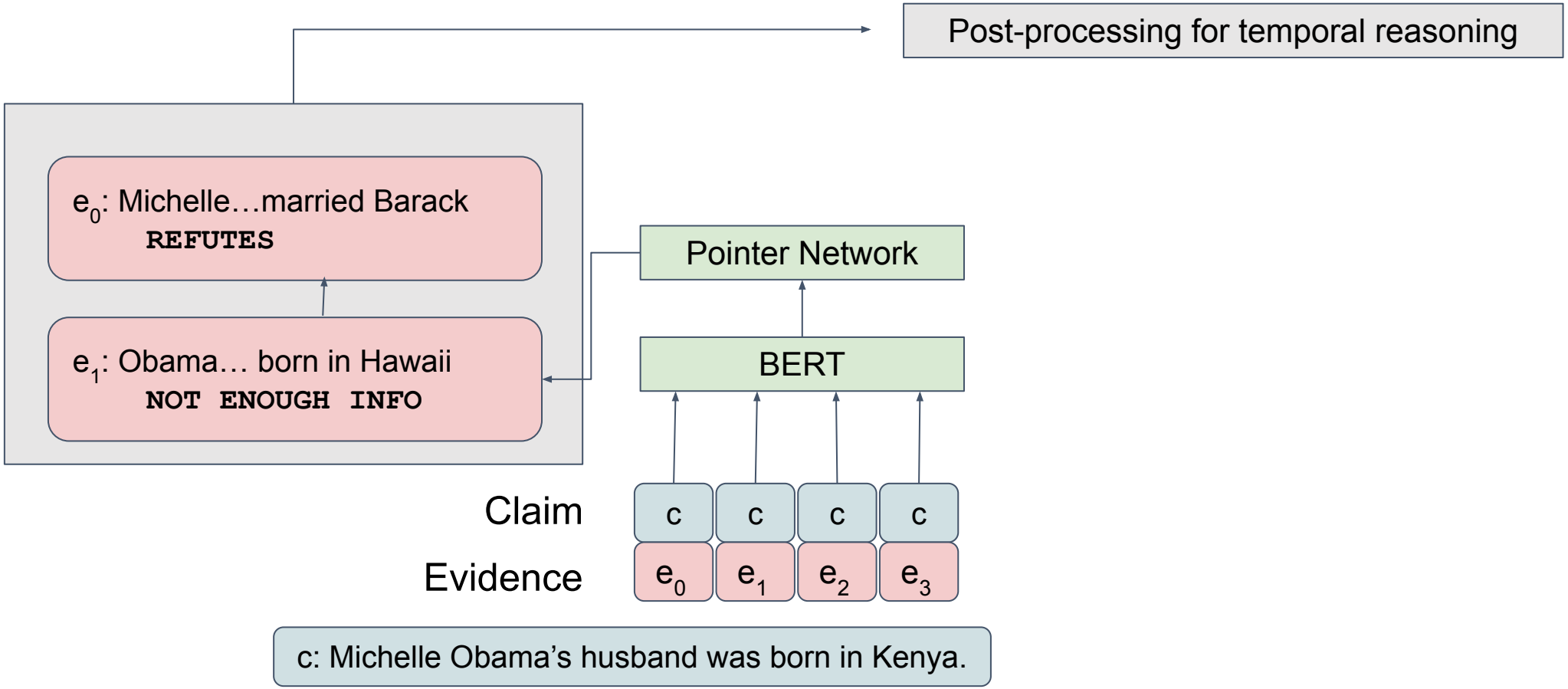
Methods: Pointer Network

Joint Sentence Selection and Relation Prediction



DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Methods: Pointer Network



DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.

Methods: Temporal Post-processing

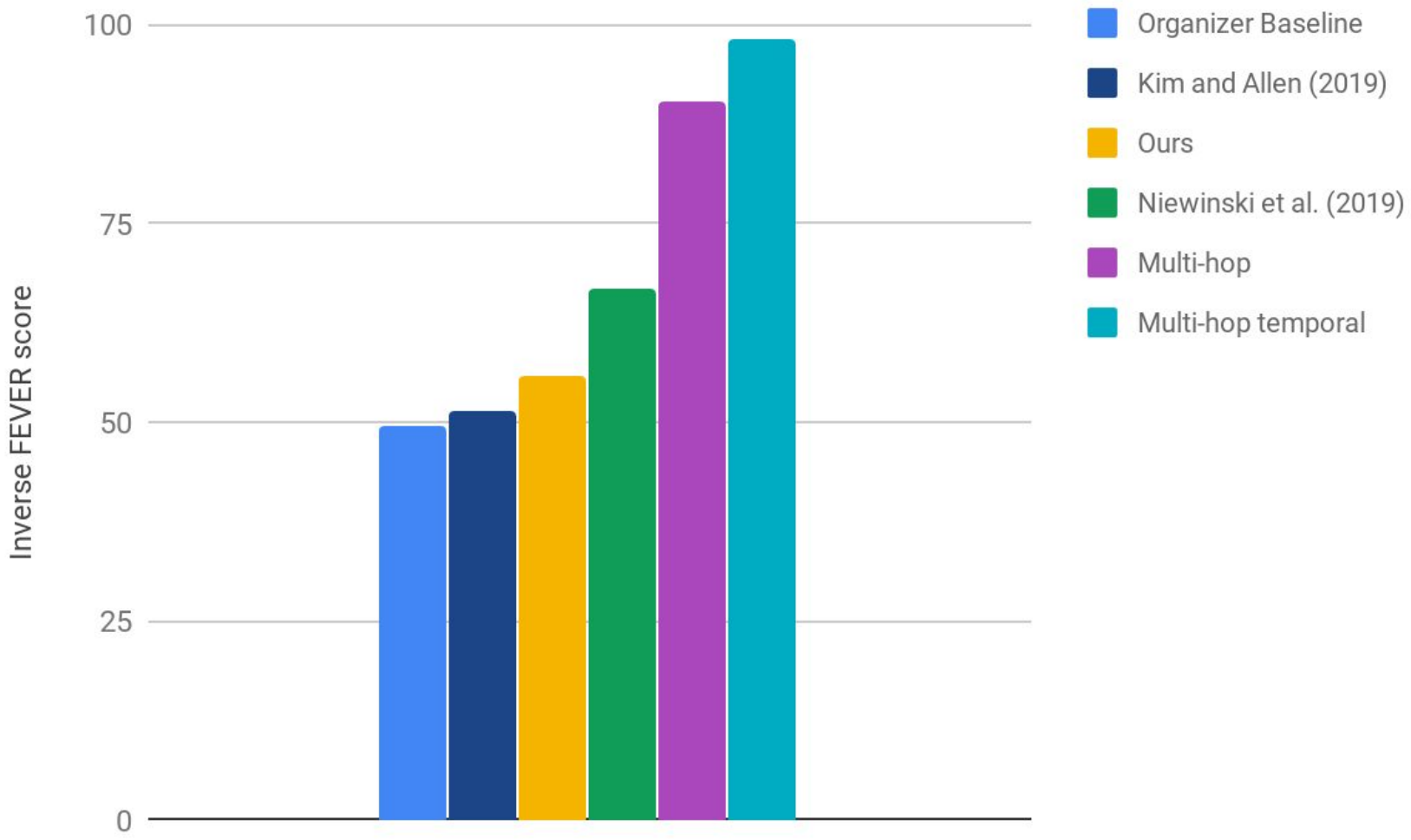
1. Extract with Open IE and normalize:

   The Latvian Soviet Socialist Republic was a republic of the Soviet Union **3 years after 2009**.

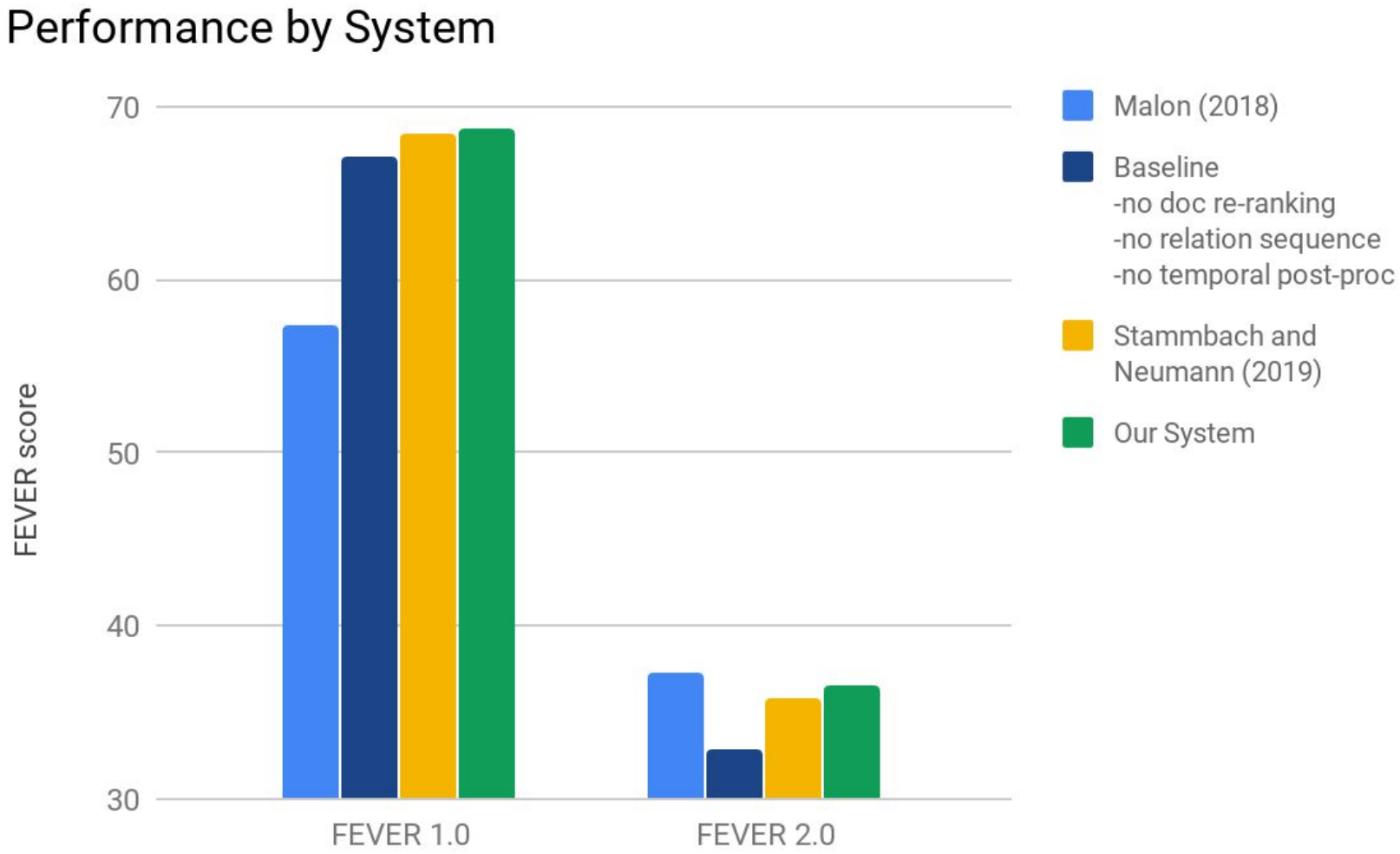
2. Compare only dates in retrieved evidence:

   The Soviet Union … existed **from 1922 to 1991**.

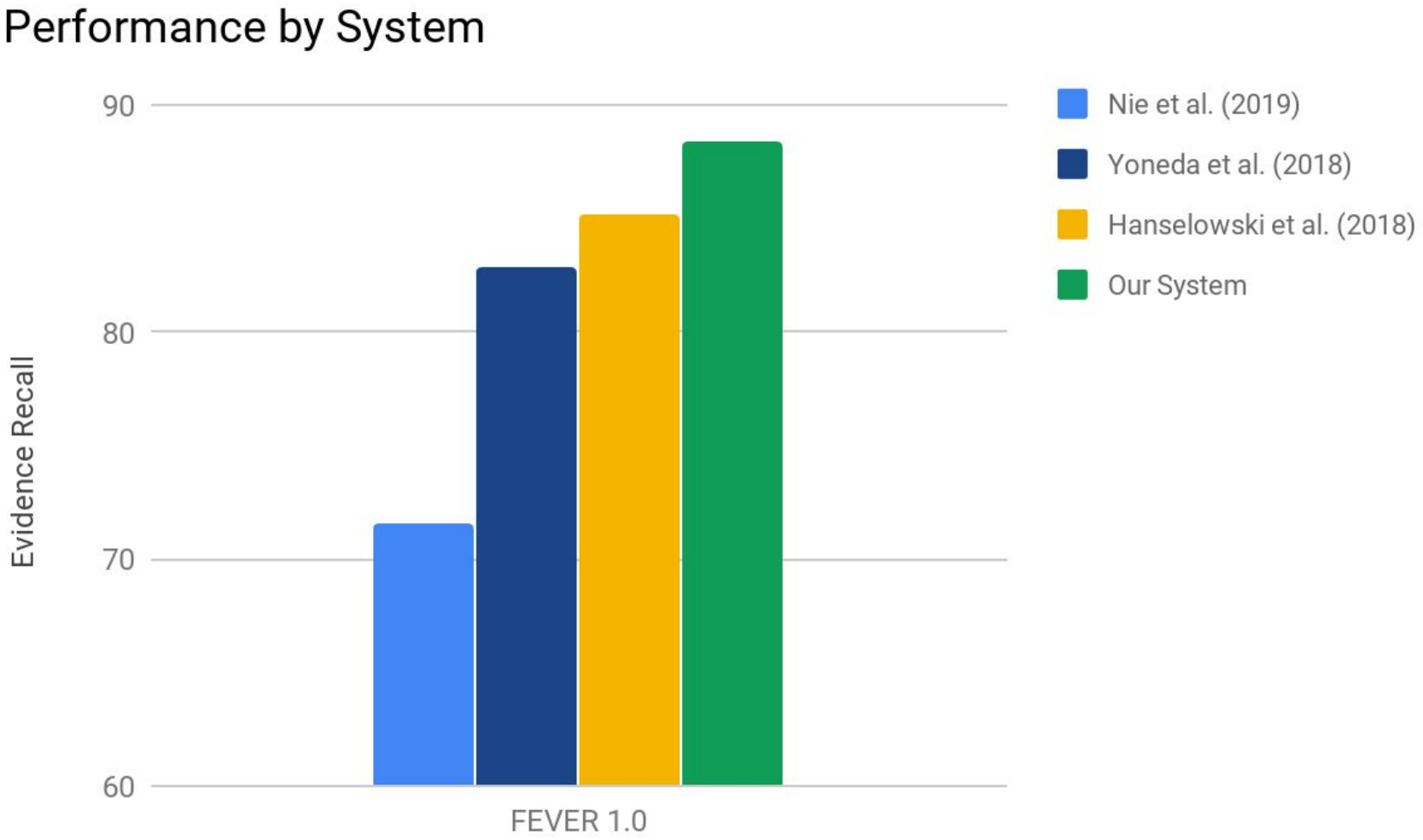
Evaluation: Data



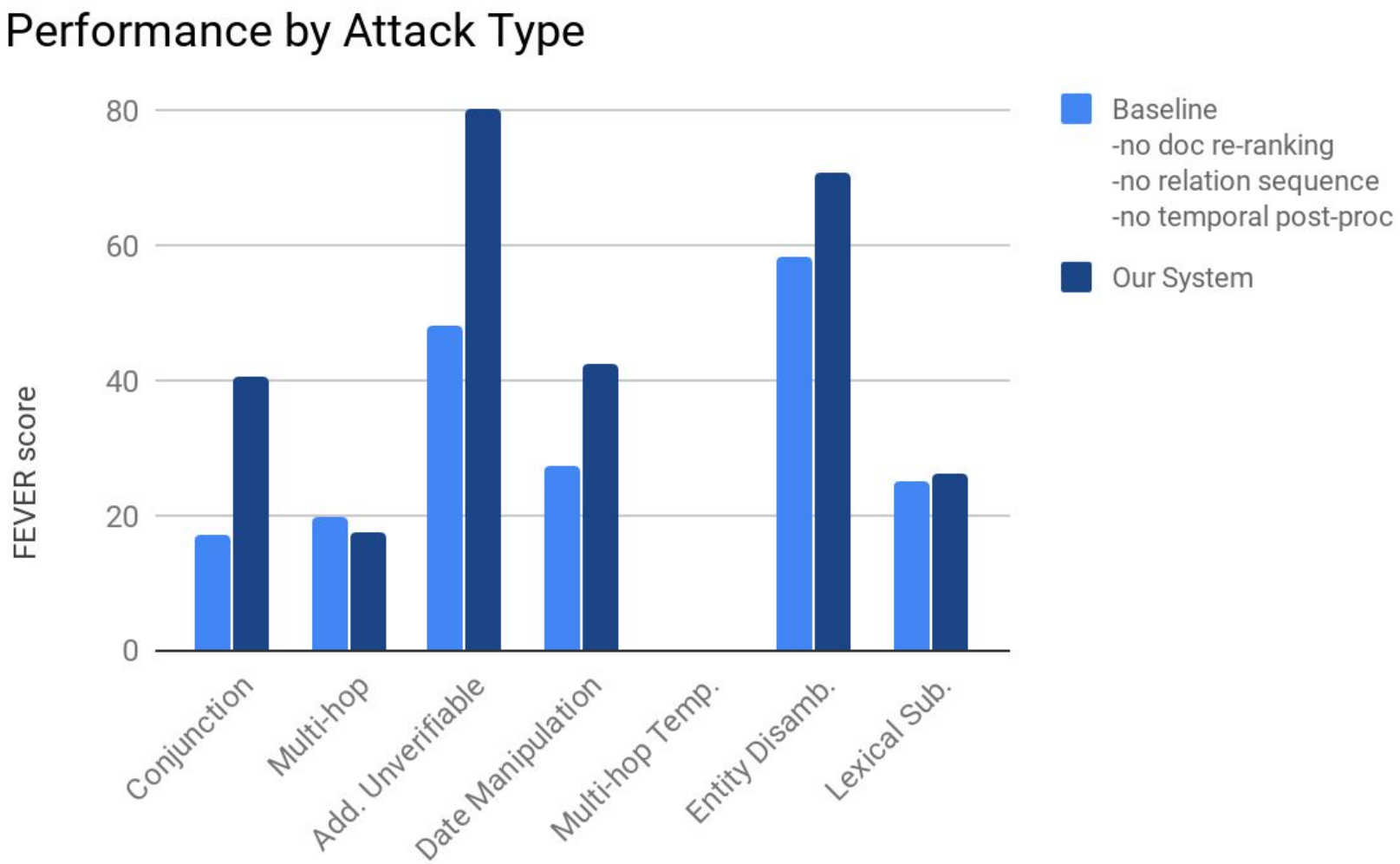
Evaluation: Methods



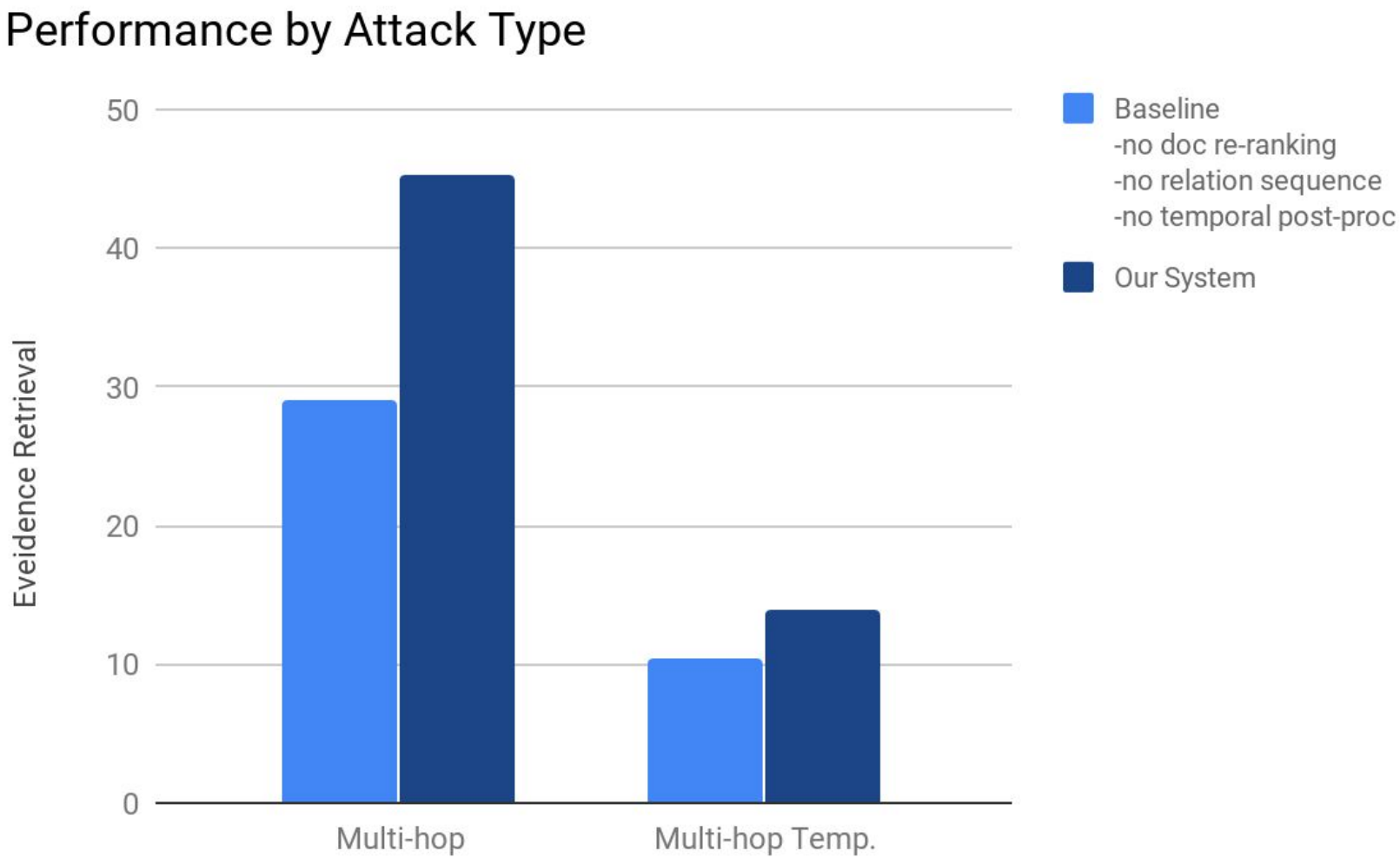
Evaluation: Methods



Evaluation: Attack Types



Evaluation: Attack Types



Evaluation: Qualitative Analysis

Baseline:

Honeymoon is a major-label record by Elizabeth Woolridge Grant.

[Honeymoon] **[Lana del Rey]**

SUPPORTS

Evaluation: Qualitative Analysis

Our System:

Honeymoon is a major-label record by Elizabeth Woolridge Grant.

✓ ✓

[Honeymoon (Lana del Rey album)] [Lana del Rey]

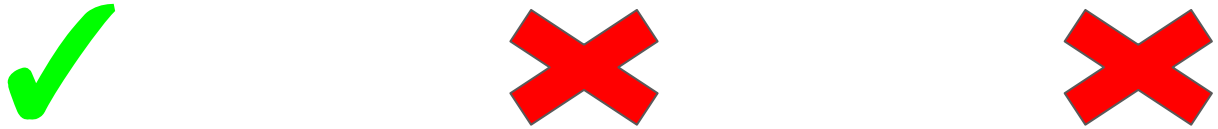
✓

SUPPORTS

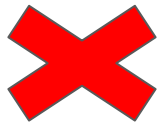
Evaluation: Qualitative Analysis

Baseline:

The MVP of the 1976 Canada Cup tournament was born before the tournament was first held.



[Bobby Orr] [World Cup of Hockey] [Canada Cup]



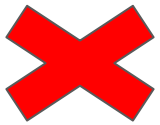
REFUTES

Evaluation: Qualitative Analysis

Our System:

The MVP of the 1976 Canada Cup tournament was born before the tournament was first held.


[Bobby Orr] [1976 Canada Cup]



NOT ENOUGH INFO

Questions?

Come join our discussion slot(s):

Wednesday July 8, 2020 14B NLP Applications-11 18:00 UTC+0

Wednesday July 8, 2020 15B NLP Applications-12 21:00 UTC+0

Evaluation: Ablation

System	FEVER 1.0 Dev	FEVER 2.0 Dev
Builders	69.2	36.2
Ptr+rel.seq.loss	73.17	39.86
+doc. rank	-	41.91
+dateProc **(Fixers)**	-	43.36

Evaluation: Data

Team	#	Potency
Baseline	498	49.68
Kim and Allen (2019)	102	51.54
Ours	501	55.79
Niewinski et al. (2019)	79	66.83
Multi-hop	188	**90.34**
Multi-hop Temporal	55	**98**

Evaluation: Methods

Team	FEVER 1.0	FEVER 2.0
Athene	61.58	25.35
UNC	64.21	30.47
Our Baseline - RL	67.08	32.92
Dominiks	68.46	35.82
UCL MR	62.52	35.83
Our System	**68.8**	36.61
Papelo	57.36	**37.31**