Cure My FEVER: Building, Breaking, and Fixing Models for Fact-Checking

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

Columbia University

Automated Fact-checking and Related Tasks

Source Trustworthiness

Fact-checking

Automated Fact-Checking

Datasets and Problem Formulation

Dataset	Source	Size	Input	Output	Evidence Type
Truth of Varying Shades Rashkin et al. (2017)	Politifact + news websites	74K	Claim sentences	6 truth levels	No evidence
LIAR Wang (2017)	Politifact	12.8K	Claim Sentences	6 truth levels	metadata
Emergent Ferreira and Vlachos (2016)	Snopes.com Twitter	300 claims 2,595 articles	Pair (claim, article headline)	for, against, observes	News Articles
FNC-1 Pomerleau and Rao (2017)	Emergent	50K	Pair (headline, article body)	agree, disagree, discuss, unrelated	News Articles
FEVER Thorne et al. (2018)	Synthetic	185K	Claim sentences	Support, Refute, Not Enough Info	Sentences from Wikipedia

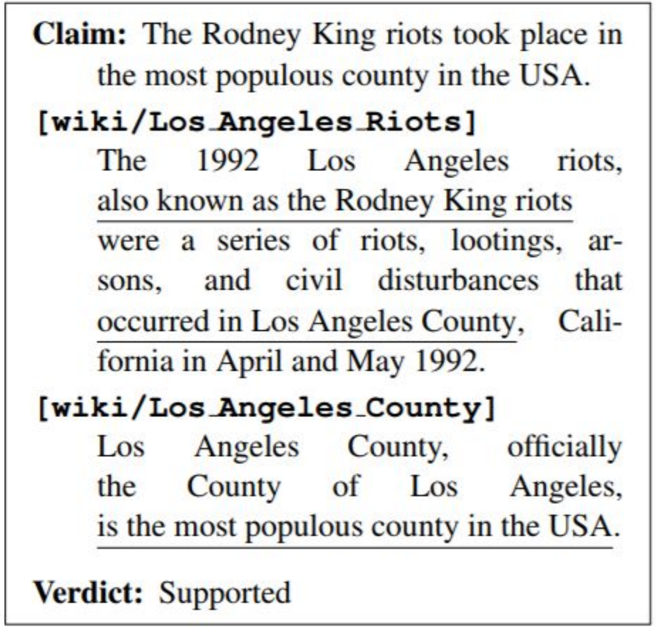
Automated Fact-Checking

Datasets and Problem Formulation

Dataset	Source	Size	Input	Output	Evidence Type
Truth of Varying Shades Rashkin et al. (2017)	Politifact + news websites	74K	Claim sentences	6 truth levels	No evidence
LIAR Wang (2017)	Politifact	12.8K	Claim Sentences	6 truth levels	metadata
Emergent Ferreira and Vlachos (2016)	Snopes.com Twitter	300 claims 2,595 articles	Pair (claim, article headline)	for, against, observes	News Articles
FNC-1 Pomerleau and Rao (2017)	Emergent	50K	Pair (headline, article body)	agree, disagree, discuss, unrelated	News Articles
FEVER Thorne et al. (2018)	Synthetic	185K	Claim sentences	Support, Refute, Not Enough Info	Sentences from Wikipedia

Overview

- FEVER: Fact Extraction and VERification of 185,445 claims
- Dataset
  - Claim Generation
  - Claim Labeling
- System
  - Document Retrieval
  - Sentence Selection
  - Textual Entailment



Claim: The Rodney King riots took place in the most populous county in the USA.

`[wiki/Los_Angeles_Riots]`

The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

`[wiki/Los_Angeles_County]`

Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported

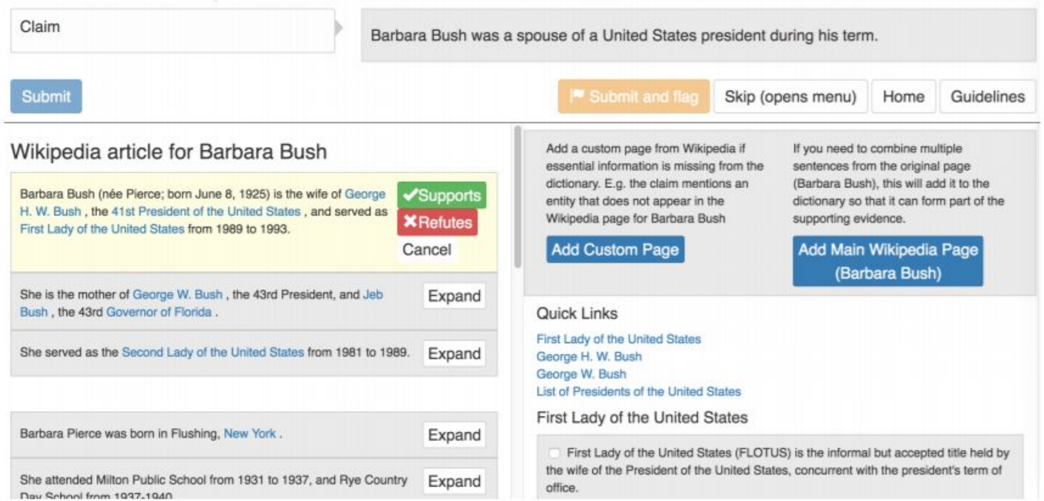
Claim Generation

- Sample sentences from the introductory section of 50,000 popular pages (5,000 of Wikipedia’s most accessed pages and their linked pages)

- **Task:** given a sample sentence, generate a set of claims containing a single piece of information focusing on the entity that its original Wikipedia page was about.
  - Entities: a dictionary of terms with wikipedia pages.
  - Create mutations of the claims.
  - Average claim length is 9.4 tokens

Claim Labeling

- In 31.75% of the claims more than one sentence was considered appropriate evidence
- Claims require composition of evidence from multiple sentences in 16.82% of cases.
- In 12.15% of the claims, this evidence was taken from multiple pages.
- IAA in evidence retrieval 95.42% precision and 72.36% recall.



Barbara Bush was a spouse of a United States president during his term.

Wikipedia article for Barbara Bush

Barbara Bush (née Pierce; born June 8, 1925) is the wife of George H. W. Bush , the 41st President of the United States , and served as First Lady of the United States from 1989 to 1993.

She is the mother of George W. Bush , the 43rd President, and Jeb Bush , the 43rd Governor of Florida .

She served as the Second Lady of the United States from 1981 to 1989.

Barbara Pierce was born in Flushing, New York .

She attended Milton Public School from 1931 to 1937, and Rye Country Day School from 1937-1940

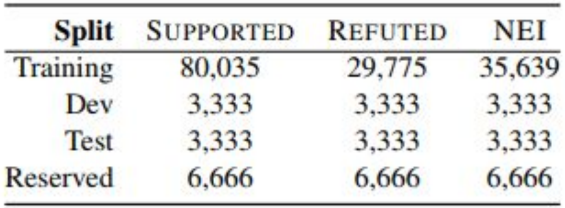
Add a custom page from Wikipedia if essential information is missing from the dictionary. E.g. the claim mentions an entity that does not appear in the Wikipedia page for Barbara Bush

If you need to combine multiple sentences from the original page (Barbara Bush), this will add it to the dictionary so that it can form part of the supporting evidence.

First Lady of the United States
George H. W. Bush
George W. Bush
List of Presidents of the United States

First Lady of the United States

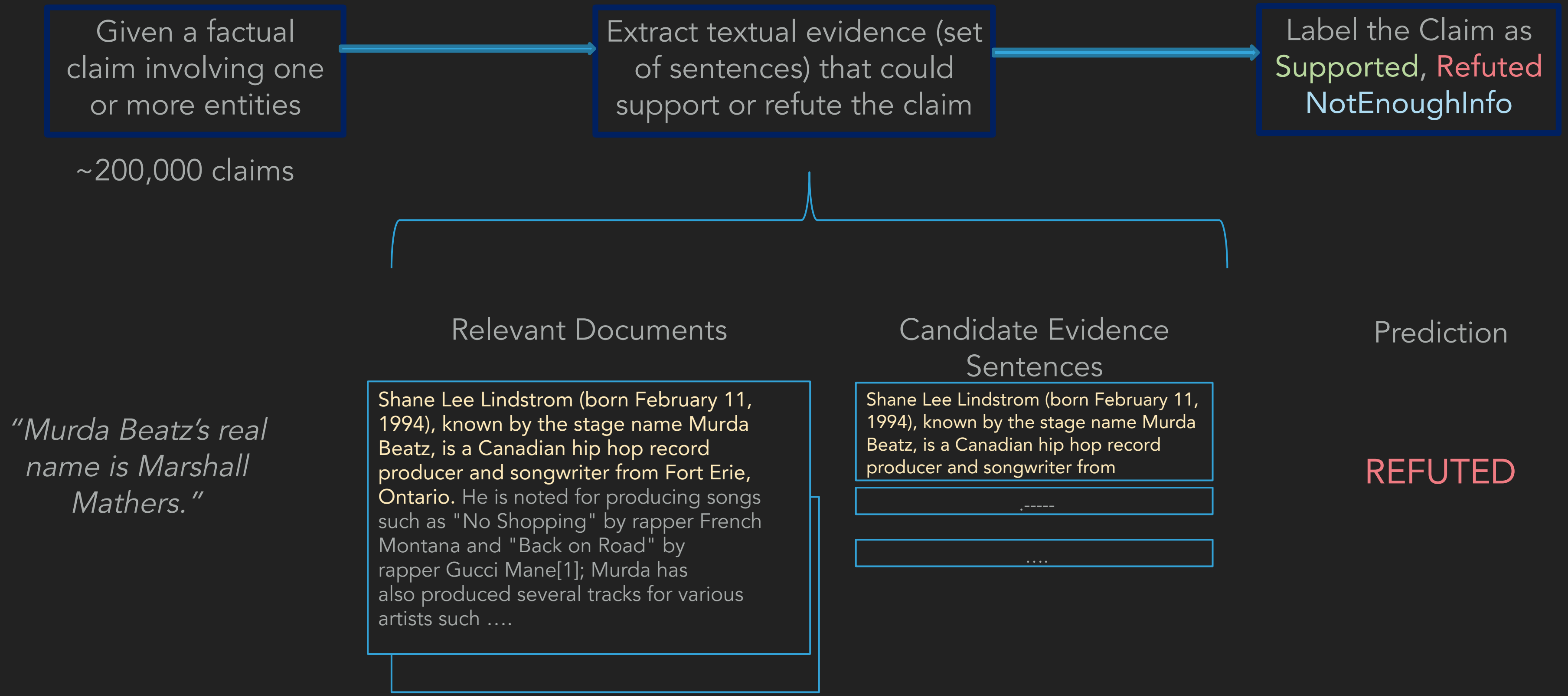
□ First Lady of the United States (FLOTUS) is the informal but accepted title held by the wife of the President of the United States, concurrent with the president's term of office.



Split	SUPPORTED	REFUTED	NEI
Training	80,035	29,775	35,639
Dev	3,333	3,333	3,333
Test	3,333	3,333	3,333
Reserved	6,666	6,666	6,666

Table 1: Dataset split sizes for SUPPORTED, REFUTED and NOTENOUGHINFO (NEI) classes

FACT EXTRACTION AND VERIFICATION (FEVER)



Given a factual claim involving one or more entities

~200,000 claims

Extract textual evidence (set of sentences) that could support or refute the claim

Label the Claim as Supported, Refuted NotEnoughInfo

"Murda Beatz's real name is Marshall Mathers."

Relevant Documents

Shane Lee Lindstrom (born February 11, 1994), known by the stage name Murda Beatz, is a Canadian hip hop record producer and songwriter from Fort Erie, Ontario. He is noted for producing songs such as "No Shopping" by rapper French Montana and "Back on Road" by rapper Gucci Mane[1]; Murda has also produced several tracks for various artists such ….

Candidate Evidence Sentences

Shane Lee Lindstrom (born February 11, 1994), known by the stage name Murda Beatz, is a Canadian hip hop record producer and songwriter from

.-----

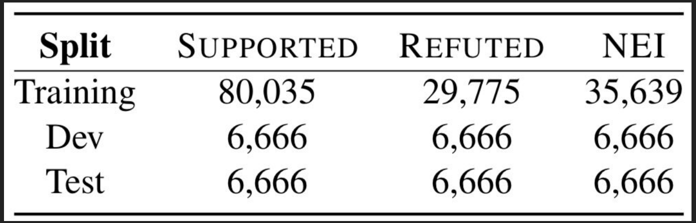
….

Prediction

REFUTED

DATA AND METRICS

- 185,445 Claims



Split	SUPPORTED	REFUTED	NEI
Training	80,035	29,775	35,639
Dev	6,666	6,666	6,666
Test	6,666	6,666	6,666

- Metric:
  - FEVER score = label accuracy conditioned on providing at least one complete set of evidence

FACT EXTRACTION AND VERIFICATION (FEVER) VERSION 1.0

(Chakrabarty, Alhindi, Muresan, 2018)

Relevant Documents

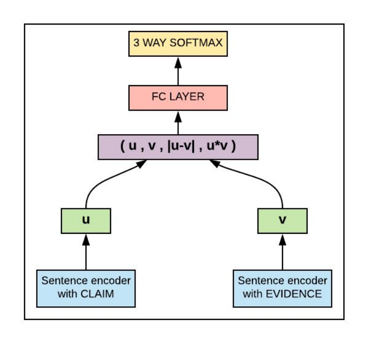
- Google API: retrieve top documents for the claim
- Wikipedia API: Retrieve top documents for each named entity in the claim
- Query Wikipedia Search API with the subject of the claim

Candidate Evidence Sentences

- Use contextualized word embeddings (ELMO) to represent the claim and candidate evidence sentences.
- Compute cosine similarity and retrieve the top 5 most relevant sentences from the relevant documents

Textual Entailment Task

- Model each Claim – Candidate Evidence pair separately
- Do on top 3 candidates

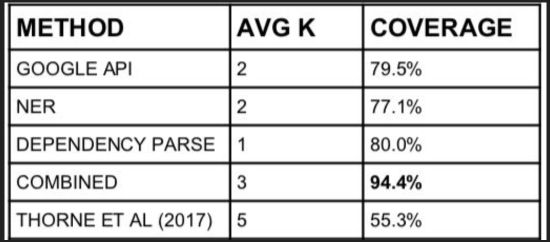


Ranked 6th on the task last year on FEVER score

(Chakrabarty, Alhindi, Muresan, 2018)

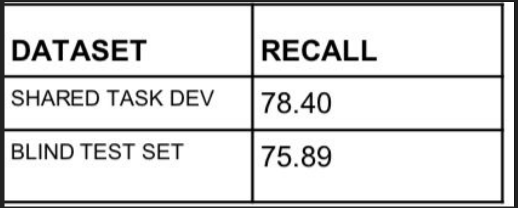
RESULTS FOR ALL STAGES

- Doc retrieval



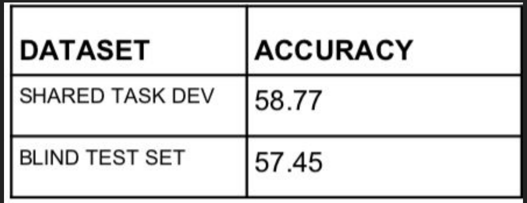
METHOD	AVG K	COVERAGE
GOOGLE API	2	79.5%
NER	2	77.1%
DEPENDENCY PARSE	1	80.0%
COMBINED	3	**94.4%**
THORNE ET AL (2017)	5	55.3%

- Evidence Recall



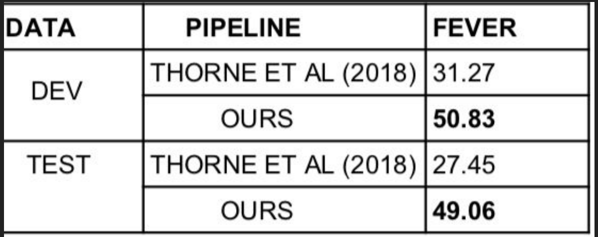
DATASET	RECALL
SHARED TASK DEV	78.40
BLIND TEST SET	75.89

- Entailment Accuracy



DATASET	ACCURACY
SHARED TASK DEV	58.77
BLIND TEST SET	57.45

- FEVER score



DATA	PIPELINE	FEVER
DEV	THORNE ET AL (2018)	31.27
	OURS	**50.83**
TEST	THORNE ET AL (2018)	27.45
	OURS	**49.06**

(Chakrabart, Alhindi, Muresan, 2018)

ERROR ANALYSIS

- System wrongly penalized for not matching gold evidence

Claim: Aristotle spent time in Athens

System Prediction (correct): Supported

System Evidence (not in gold): *At seventeen or eighteen years of age, he joined Plato's Academy in Athens and remained there until the age of thirty-seven*

System Evidence (not in gold): *Shortly after Plato died , Aristotle left Athens and at the request of Philip II of Macedon ,tutored Alexander the Great beginning in 343 BC*

(Chakrabart, Alhindi, Muresan, 2018)

ERROR ANALYSIS

- Need better semantics (to distinguish NotEnoughInfo from Supported)

Claim: *Happiness in Slavery is a gospel song by Nine Inch Nails*

System Prediction: Supported

Gold Label: NotEngoughInfo

System Evidence: *Happiness in Slavery,is a song by American industrial rock band Nine Inch Nails from their debut extended play (EP), Broken(1992)*

Fact Extraction and VERification (FEVER) Version 2

Breakers

Development of adversarial claims

Builders & Fixers

Development of initial system and targeted improvements

Breakers

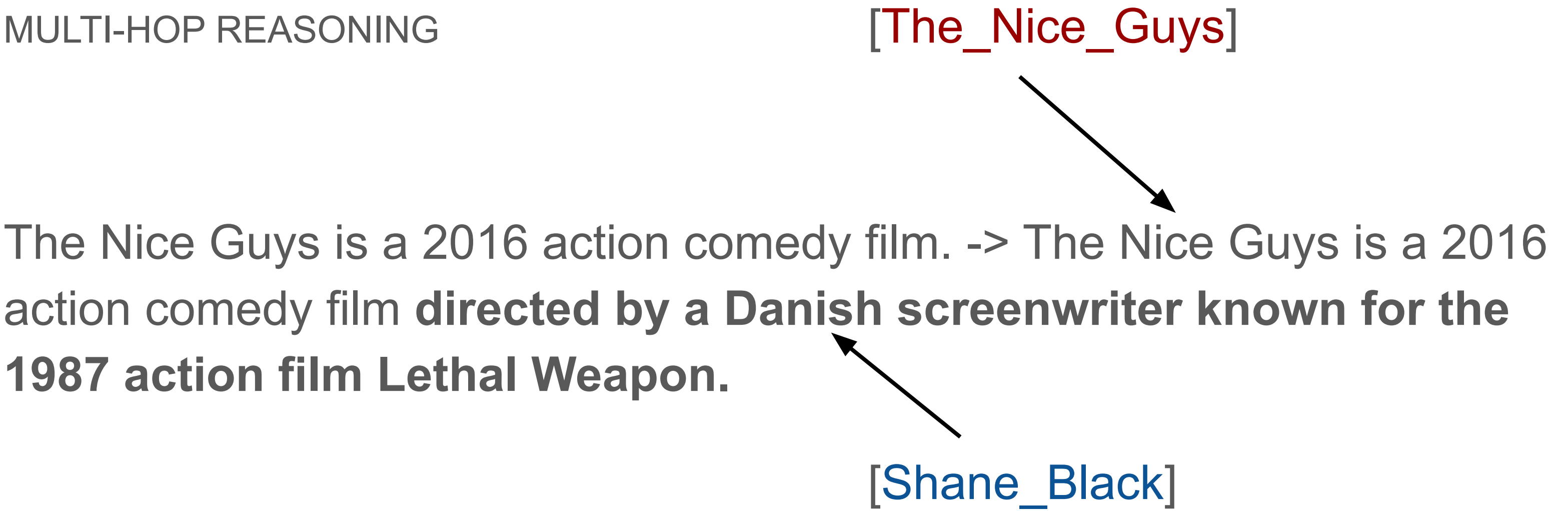
1) Multiple propositions : Claims that require multi-hop document or sentence retrieval
   a) CONJUNCTION

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

      -> Janet Leigh was from New York and was an author.

Breakers

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



[The_Nice_Guys]

The Nice Guys is a 2016 action comedy film. -> The Nice Guys is a 2016 action comedy film **directed by a Danish screenwriter known for the 1987 action film Lethal Weapon.**

[Shane_Black]

Breakers

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

Duff McKagan is an American citizen

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

Breakers

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

        in 2001 -> in the first decade of the 21st century

        in 2009→ 3 years before 2012

Breakers

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

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

Admittance of Indiana Territory (1816) BEFORE William Henry Harrison (death 1841)

Breakers

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

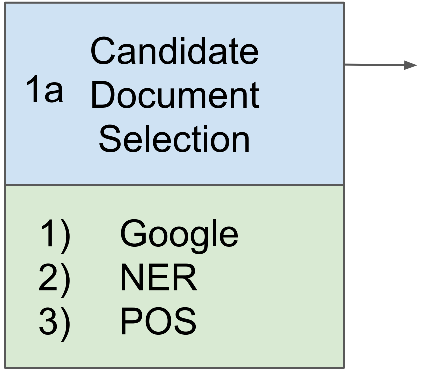
       Patrick Stewart -> Patrick Maxwell Stewart

Breakers

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

       filming -> shooting

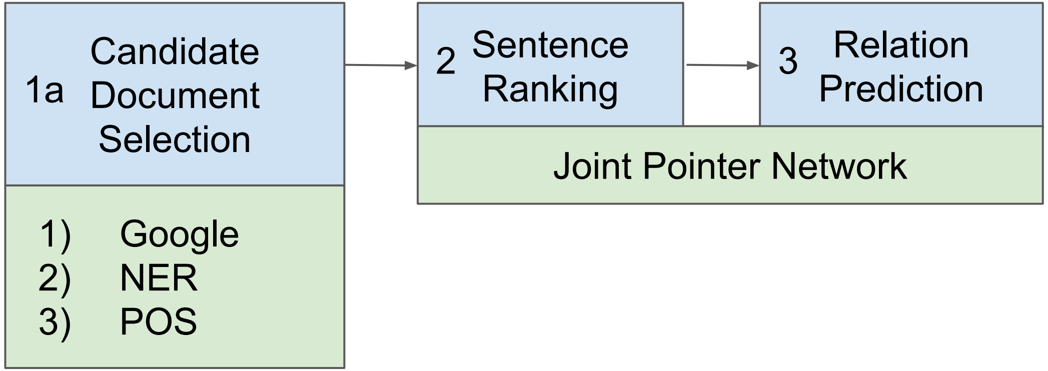
Builders



1a Candidate Document Selection

1) Google
2) NER
3) POS

Builders



1a Candidate Document Selection

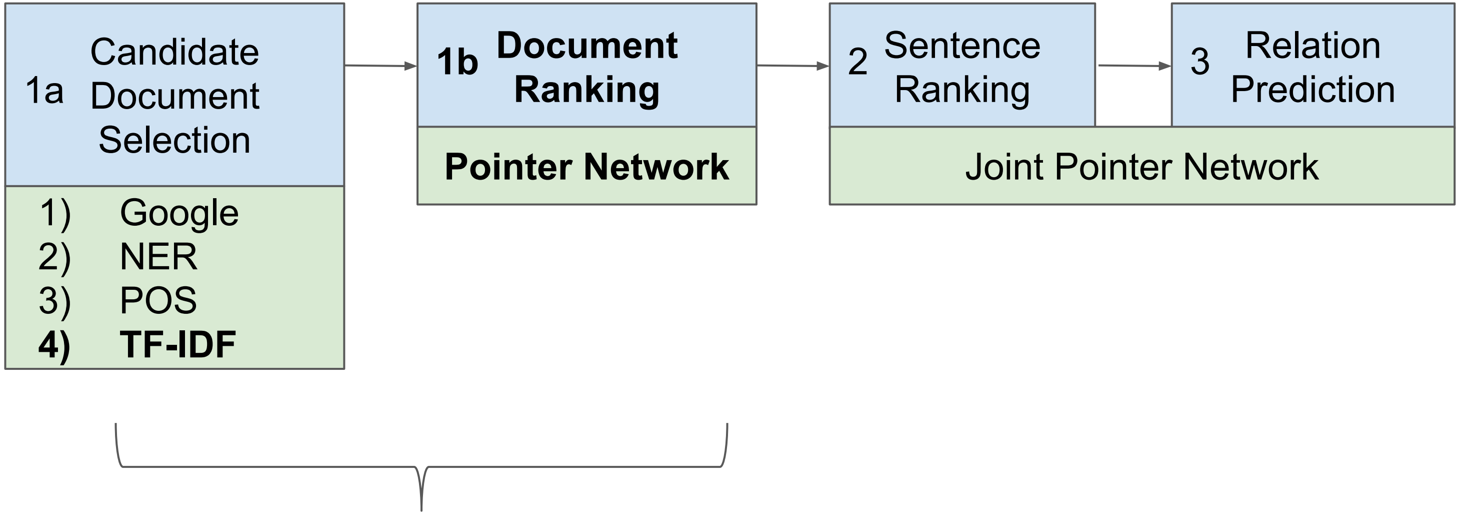
1) Google
2) NER
3) POS

2 Sentence Ranking

3 Relation Prediction

Joint Pointer Network

Fixers



1a Candidate Document Selection

1) Google
2) NER
3) POS
4) **TF-IDF**

1b Document Ranking

Pointer Network

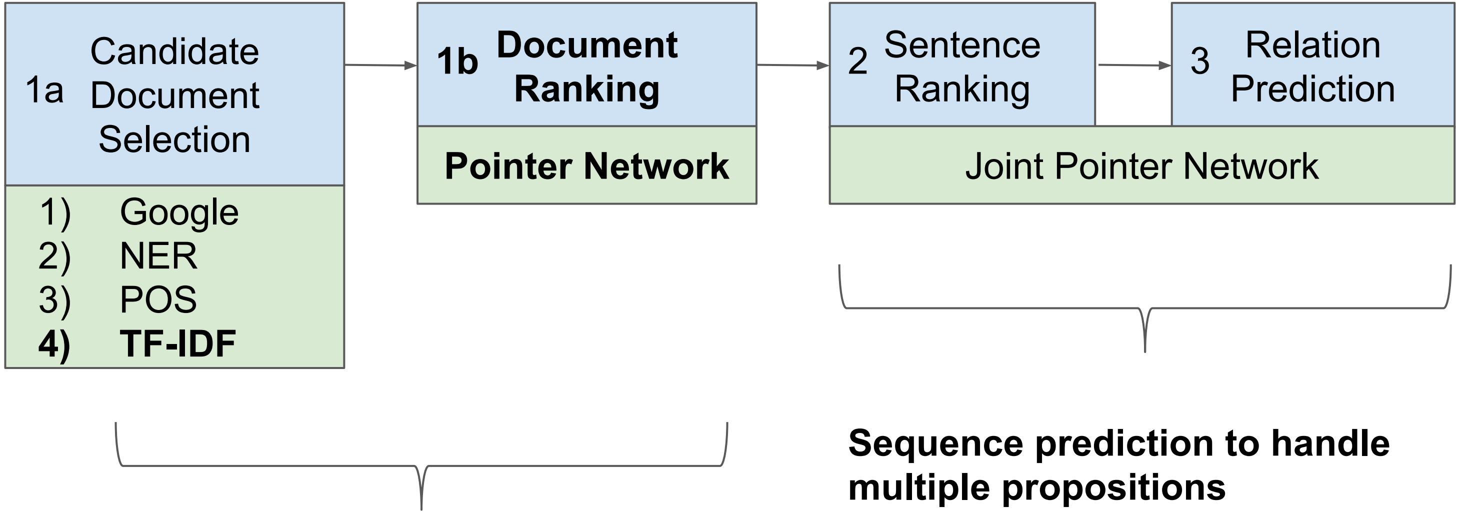
2 Sentence Ranking

3 Relation Prediction

Joint Pointer Network

Overgenerate and re-rank to handle ambiguity

Fixers



1a Candidate Document Selection

1) Google
2) NER
3) POS
4) TF-IDF

1b Document Ranking

Pointer Network

2 Sentence Ranking

3 Relation Prediction

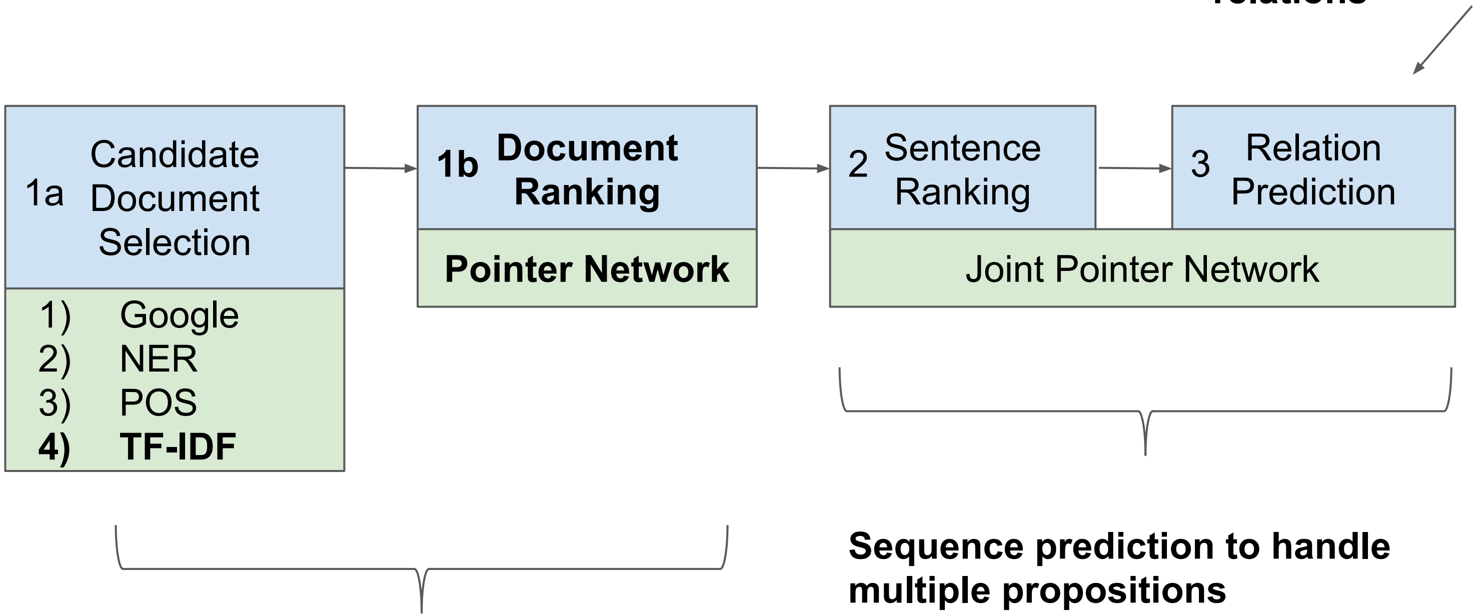
Joint Pointer Network

Sequence prediction to handle multiple propositions

Overgenerate and re-rank to handle ambiguity

Fixers

Post-processing to handle temporal relations



1a Candidate Document Selection

1) Google
2) NER
3) POS
4) **TF-IDF**

1b Document Ranking

Pointer Network

2 Sentence Ranking

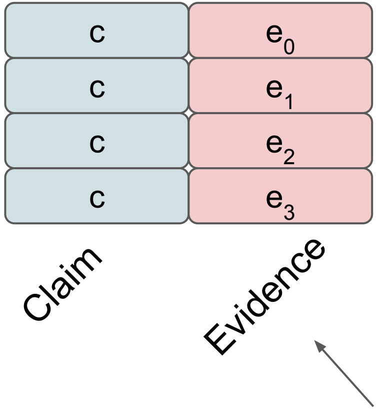
3 Relation Prediction

Joint Pointer Network

Sequence prediction to handle multiple propositions

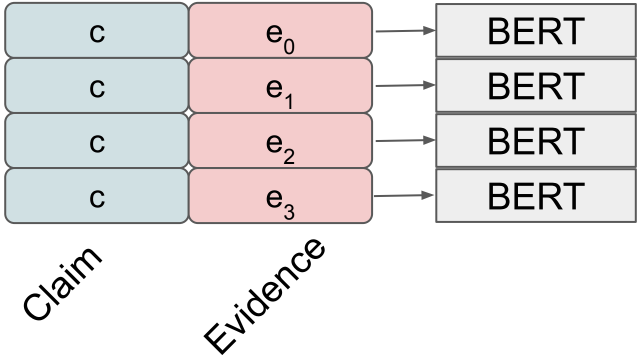
Overgenerate and re-rank to handle ambiguity

Pointer Network



Candidate sentences

Pointer Network



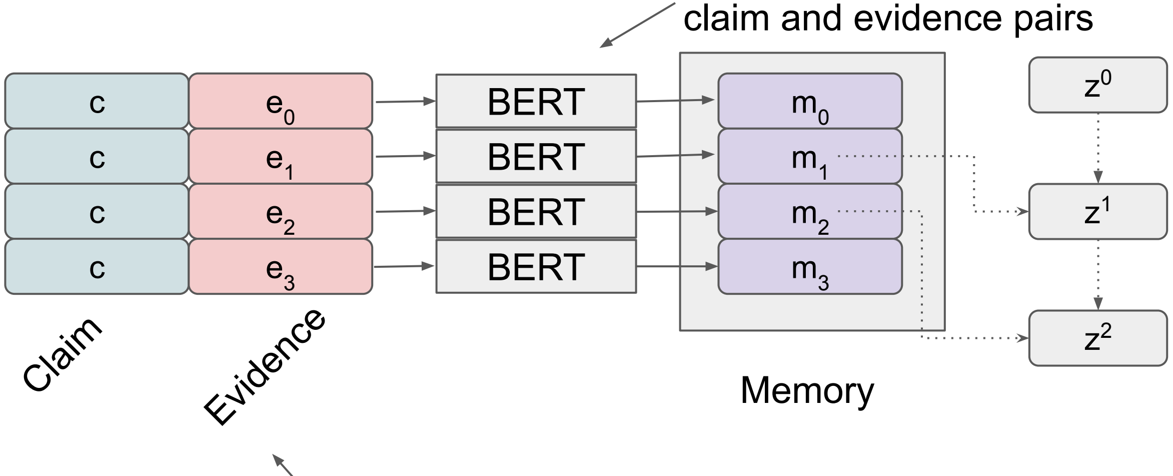
Model fine-tuned on gold claim and evidence pairs

Candidate sentences

Pointer Network - Builders

c=The Nice Guys is a 2016 film directed by a Danish screenwriter known for Lethal Weapon.

Model fine-tuned on gold claim and evidence pairs



Candidate sentences

LSTM decoder

e^0=The Nice Guys is a 2016 American neo-noir crime black comedy film directed by Shane Black …

e^1=Shane Black… is an American filmmaker… written such films as Lethal Weapon…

e^2=He made his directorial debut with the film Kiss Kiss Bang Bang...

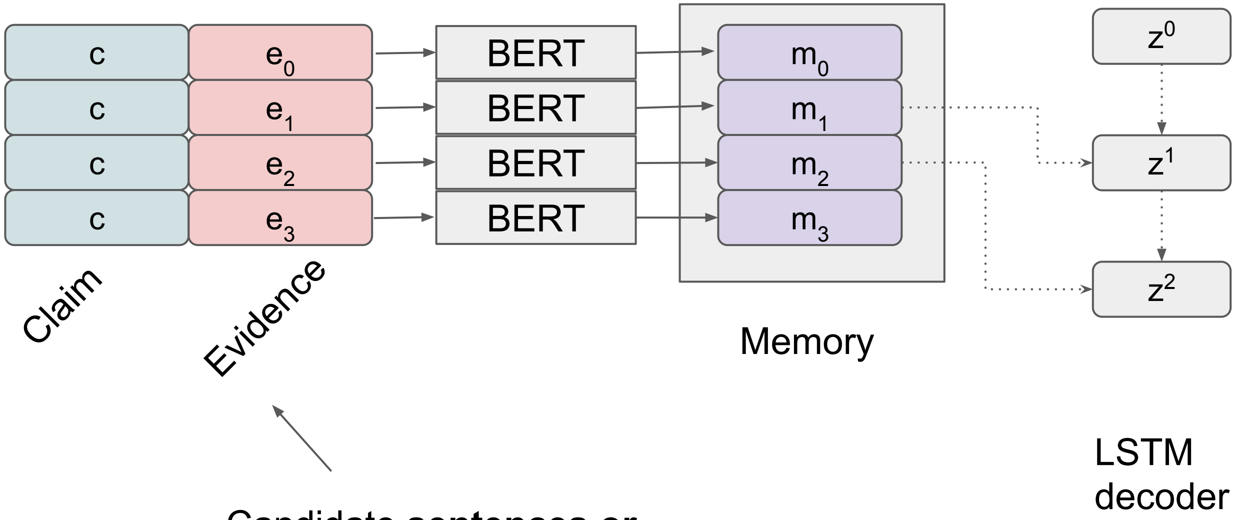
l=REFUTES

Concatenate evidence to make label prediction, train using RL

Pointer Network - Fixers

c=The Nice Guys is a 2016 film directed by a Danish screenwriter known for Lethal Weapon.

Model fine-tuned on gold claim and evidence pairs



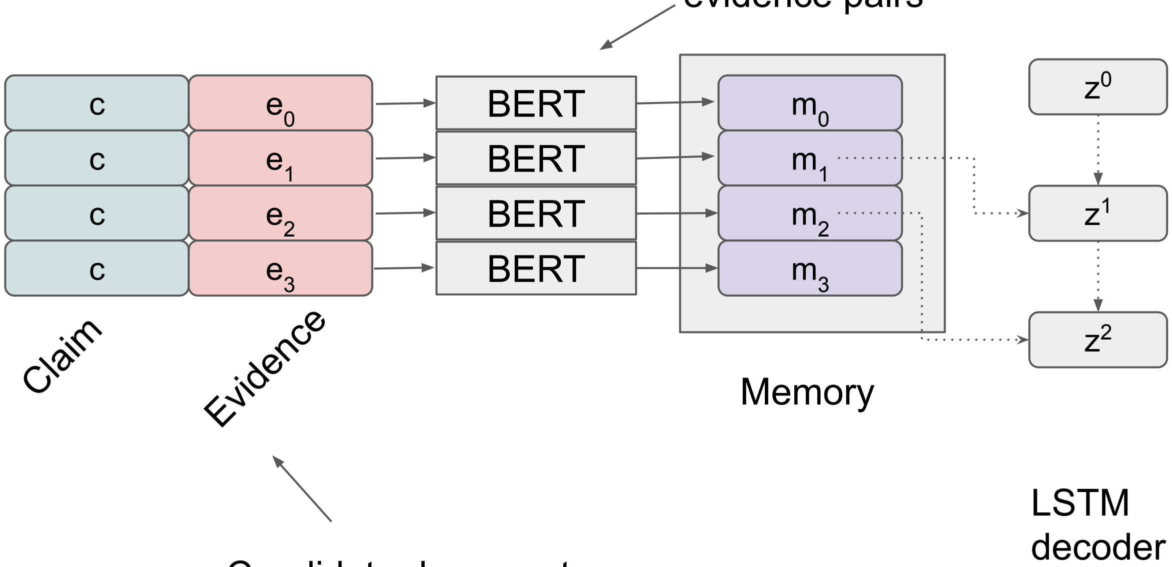
e^0=Nice Guys

e^1=Shane Black

Candidate **sentences or documents**

Pointer Network

Model fine-tuned on gold claim and evidence pairs



Candidate documents or sentences

c=The Nice Guys is a 2016 film directed by a Danish screenwriter known for Lethal Weapon.

e^0=The Nice Guys is a 2016 American neo-noir crime black comedy film directed by Shane Black …
l^0=NEI

e^1=Shane Black… is an American filmmaker… written such films as Lethal Weapon…
l^1=REFUTES

e^2=He made his directorial debut with the film Kiss Kiss Bang Bang...
l^2=REFUTES

Post-processing for Temporal Relations

1. Extract temporal expressions:
   The Latvian Soviet Socialist Republic was a republic of the Soviet Union **3 years after 2009**.
   1. Open IE -> **3 years after 2009**
   2. Normalize -> **2012**

2. Compare only dates in retrieved evidence:
   The Soviet Union … existed **from 1922 to 1991**.
   1991 < 2012 -> Refutes

Results - Breakers

Team	#	Raw Potency	Correctness
Baseline	498	60.34	82.33
NbAuzDrLqg	102	79.66	64.71
Ours	501	68.51	81.44
TMLab	79	79.97	84.81

Results - Builders

Team	FEVER 1.0	FEVER 2.0
Athene	61.58	25.35
UNC	64.21	30.47
Builders	67.08	32.92
Dominiks	**68.46**	35.82
UCL MR	62.52	35.83
Papelo	57.36	**37.31**

Results - Fixers

Team	FEVER 1.0	FEVER 2.0
Athene	61.58	25.35
UNC	64.21	30.47
Builders	67.08	32.92
Dominiks	68.46	35.82
UCL MR	62.52	35.83
Fixers	**68.8**	36.61
Papelo	57.36	**37.31**

Questions?