Neural Models of Factuality

Rachel Rudinger
Johns Hopkins University
Aaron Steven White
University of Rochester
Benjamin Van Durme
Johns Hopkins University

JOHNS HOPKINS UNIVERSITY

UNIVERSITY of ROCHESTER

Rachel
Rudinger

Aaron Steven
White

Ben
Van Durme

Student First Author

Rachel Rudinger

Slides at aaronstevenwhite.io

What is (event) factuality?

What is (event) factuality?

Did that **event/state** happen?

Nairn et al., 2006; Sauri and Pustejovsky, 2009, 2012; de Marneffe et al., 2012

What is (event) factuality?

Did that event/state happen?

Nairn et al., 2006; Sauri and Pustejovsky, 2009, 2012; de Marneffe et al., 2012

Unfortunately for LeBron James and the Cavs, though, **none** of the underdogs on the list above ended up **winning** the championship.

https://fivethirtyeight.com/features/six-key-stats-on-the-warriors-vs-cavs-re-re-rematch/

What is (event) factuality?

Did that **event/state** happen?

Nairn et al., 2006; Sauri and Pustejovsky, 2009, 2012; de Marneffe et al., 2012

Unfortunately for LeBron James and the Cavs, though, **none** of the underdogs on the list above ended up **winning** the championship.

DIDN'T HAPPEN!

https://fivethirtyeight.com/features/six-key-stats-on-the-warriors-vs-cavs-re-re-rematch/

What is (event) factuality?

Did that **event/state** happen?

Nairn et al., 2006; Sauri and Pustejovsky, 2009, 2012; de Marneffe et al., 2012

Unfortunately for LeBron James and the Cavs, though, **none** of the underdogs on the list above ended up **winning** the championship.

DIDN'T HAPPEN!

https://fivethirtyeight.com/features/six-key-stats-on-the-warriors-vs-cavs-re-re-rematch/

Why care as a linguist?

Why care as a linguist?

Event factuality is a window into **complex interactions** between **semantic operators**.

Kiparsky and Kiparsky, 1970; Karttunen, 1971a,b; Horn, 1972; Karttunen and Peters, 1979; Heim, 1992; Simons, 2001, 2007; Simons et al., 2010; Abusch, 2002, 2010; Gajewski, 2007; Anand and Hacquard, 2013, 2014

Why care as a linguist?

Event factuality is a window into **complex interactions** between **semantic operators.**

Kiparsky and Kiparsky, 1970; Karttunen, 1971a,b; Horn, 1972; Karttunen and Peters, 1979; Heim, 1992; Simons, 2001, 2007; Simons et al., 2010; Abusch, 2002, 2010; Gajewski, 2007; Anand and Hacquard, 2013, 2014

Because he did**n't** **remember** **that** he **was the author of the words**, he would pooh-pooh some passages...

http://mentalfloss.com/article/82018/10-charming-facts-about-eb-white

Why care as a linguist?

Event factuality is a window into **complex interactions** between **semantic operators.**

Kiparsky and Kiparsky, 1970; Karttunen, 1971a,b; Horn, 1972; Karttunen and Peters, 1979; Heim, 1992; Simons, 2001, 2007; Simons et al., 2010; Abusch, 2002, 2010; Gajewski, 2007; Anand and Hacquard, 2013, 2014

HAPPENED!

Because he did**n't** **remember** **that** he **was the author of the words**, he would pooh-pooh some passages...

http://mentalfloss.com/article/82018/10-charming-facts-about-eb-white

Why care as a linguist?

Event factuality is a window into **complex interactions** between **semantic operators**.

Kiparsky and Kiparsky, 1970; Karttunen, 1971a,b; Horn, 1972; Karttunen and Peters, 1979; Heim, 1992; Simons, 2001, 2007; Simons et al., 2010; Abusch, 2002, 2010; Gajewski, 2007; Anand and Hacquard, 2013, 2014

What I did wrong was I forgot and did**n't** **remember** **to** **declare** an interest in that I am a part of the co-operative.

https://www.pressandjournal.co.uk/fp/news/politics/holyrood/1476786/north-east-msp-quits-partys-front-bench-after-failing-to-declare-interest-while-lobbying-councillors-to-support-planning-application/

Why care as a linguist?

Event factuality is a window into **complex interactions** between **semantic operators**.

Kiparsky and Kiparsky, 1970; Karttunen, 1971a,b; Horn, 1972; Karttunen and Peters, 1979; Heim, 1992; Simons, 2001, 2007; Simons et al., 2010; Abusch, 2002, 2010; Gajewski, 2007; Anand and Hacquard, 2013, 2014

DIDN'T HAPPEN!

What I did wrong was I forgot and did**n't** **remember** **to** **declare** an interest in that I am a part of the co-operative.

https://www.pressandjournal.co.uk/fp/news/politics/holyrood/1476786/north-east-msp-quits-partys-front-bench-after-failing-to-declare-interest-while-lobbying-councillors-to-support-planning-application/

Why care as an NLPer?

Why care as an NLPer?

Event factuality is important for **information extraction**, **KB population**, ...

Why care as an NLPer?

Event factuality is important for **information extraction**, **KB population**, ...

North Korea, South Korea agree to end war, denuclearize peninsula

By **HAKYUNG KATE LEE** and **JOOHEE CHO** Apr 27, 2018, 6:31 AM ET

Share Tweet

Why care as an NLPer?

Event factuality is important for **information extraction**, **KB population**, ...

North Korea, South Korea agree to end war, denuclearize peninsula

agreement-between(NK, SK)

Why care as an NLPer?

Event factuality is important for **information extraction**, **KB population**, ...

North Korea, South Korea agree to end war, denuclearize peninsula

agreement-between(NK, SK)
end-war-between(NK, SK)

Why care as an NLPer?

Event factuality is important for **information extraction**, **KB population**, ...

North Korea, South Korea agree to end war, denuclearize peninsula

agreement-between(NK, SK)
end-war-between(NK, SK)
denuclearize(NK+SK, KoreanPeninsula)

Why care as an NLPer?

Event factuality is important for **information extraction**, **KB population**, ...

North Korea, South Korea agree to end war, denuclearize peninsula

Our contributions

- **New event factuality dataset** on Universal Dependencies-English Web TreeBank

Our contributions

- **New event factuality dataset** on Universal Dependencies-English Web TreeBank
- Evaluation of **simple, linguistically motivated neural models** for event factuality prediction, yielding SOTA

Outline

- Data
- Models
- Results
- Analysis
- Conclusion

Outline

- Data

Outline

- Data
- Models

Outline

- Data
- Models
- Results

Outline

- Data
- Models
- Results
- Analysis

Outline

- Data
- Models
- Results
- Analysis
- Conclusion

Existing Datasets

- Focus on three existing factuality datasets:

Existing Datasets

- Focus on three existing factuality datasets:
    1. **FactBank** (9,761 predicates) Saurí & Pustejovsky 2009, 2012

Existing Datasets

- Focus on three existing factuality datasets:
  1. FactBank (9,761 predicates) Saurí & Pustejovsky 2009, 2012
  2. UW (13,644 predicates) Lee et al., 2015

Existing Datasets

- Focus on three existing factuality datasets:
  1. FACTBANK (9,761 predicates) Saurí & Pustejovsky 2009, 2012
  2. UW (13,644 predicates) Lee et al., 2015
  3. MEANTIME (1,395 predicates) Minard et al., 2016

Existing Datasets

- Focus on three existing factuality datasets:
  1. FAC ... ky 2009, 2012
  2. UW
  3. MEANTIME (1,395 predicates) Minard et al., 2016

All collected under slightly different protocols

Existing Datasets

- Focus on three existing factuality datasets:
  1. FactBank (9,761 predicates) Saurí & Pustejovsky 2009, 2012
  2. UW (13,644 predicates) Lee et al., 2015
  3. MEANTIME (1,395 predicates) Minard et al., 2016
- Unified Factuality Dataset: map factuality labels to [-3, 3] scale Stanovsky et al. 2017, following Lee et al., 2015

Existing Datasets

- Focus on three existing factuality datasets:
  1. FactBank (9,761 predicates) Saurí & Pustejovsky 2009, 2012
  2. UW (13,644 predicates) Lee et al., 2015
  3. MEANTIME (1,395 predicates) Minard et al., 2016
- Unified Factuality Dataset: map factuality labels to [-3, 3] scale Stanovsky et al. 2017, following Lee et al., 2015
  - Only top-level source for FactBank

New Dataset: **It Happened**

- **Largest** English factuality dataset to date

New Dataset: **It Happened**

- **Largest** English factuality dataset to date
  - **27,289** predicates(+args) from PredPatt White et al. 2016

New Dataset: **It Happened**

- **Largest** English factuality dataset to date
  - **27,289** predicates(+args) from PredPatt White et al. 2016
- Covers all of **Universal Dependencies** English Web Treebank v1.2 extends White et al. 2016

New Dataset: **It Happened**

- **Largest** English factuality dataset to date
  - **27,289** predicates(+args) from PredPatt White et al. 2016
- Covers all of **Universal Dependencies** English Web Treebank v1.2 (extends White et al. 2016)
- Part of the **Decompositional Semantics Initiative** (decomp.net)

Collecting **It Happened** Dataset

Do n't take that deal out until I look at it .

The sentence -------- understandable, and take -------- refer to a predicate.

Collecting **It Happened** Dataset

Do n't take that deal out until I look at it .

The sentence [is] understandable, and take [does] refer to a predicate.

According to the author, the situation referred to by take [-------] happen, and you are [-------] about that.

Collecting **It Happened** Dataset

Do n't take that deal out until I look at it .

The sentence is understandable, and take does refer to a predicate.

According to the author, the situation referred to by take did not (or will not) happen, and you are totally confident about that.

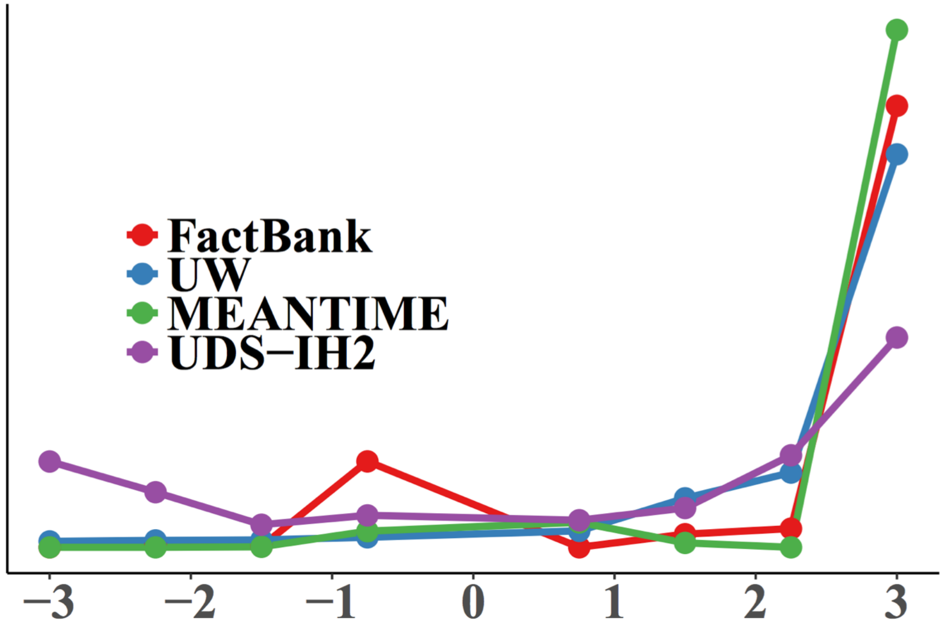
Existing Datasets

- Focus on three existing factuality datasets:
  1. FACTBANK (top-level source only) Saurí & Pustejovsky 2009, 2012
  2. UW Lee et al., 2015
  3. MEANTIME Minard et al., 2016
- Unified Factuality Dataset: map factuality labels to [-3, 3] scale Stanovsky et al. 2017, following Lee et al., 2015

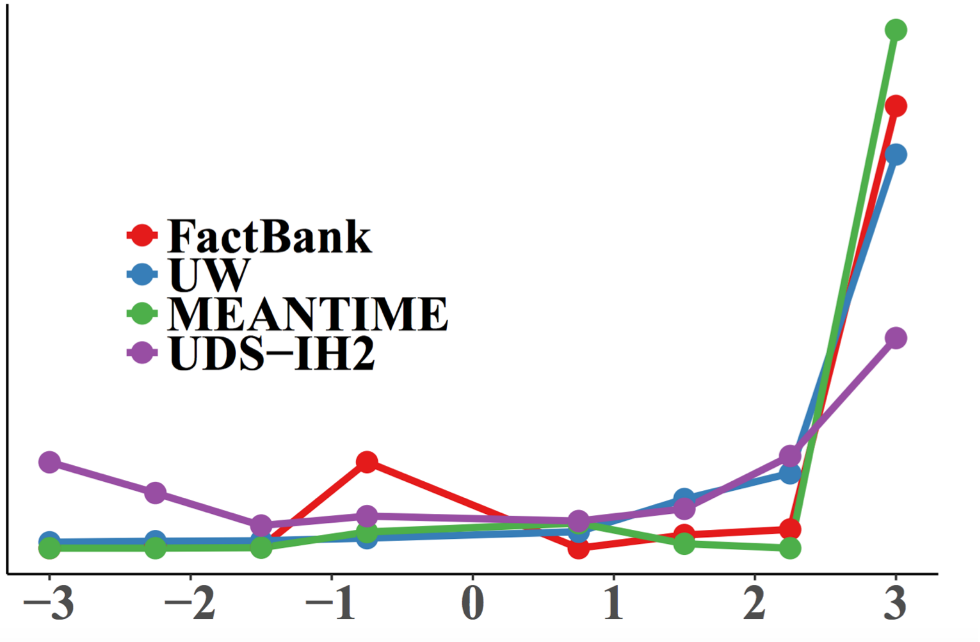
Existing Datasets

- Focus on three existing factuality datasets:
  1. FactBank (top-level source only) Saurí & Pustejovsky 2009, 2012
  2. UW Lee et al., 2015
  3. MEANTIME Minard et al., 2016
- Unified Factuality Dataset: map factuality labels to [-3, 3] scale Stanovsky et al. 2017, following Lee et al., 2015
- Map UD-It Happened to unified labels
  - Happened {yes -> +, no -> –} * ¾ * Confidence

Relative Frequency of Factuality Labels

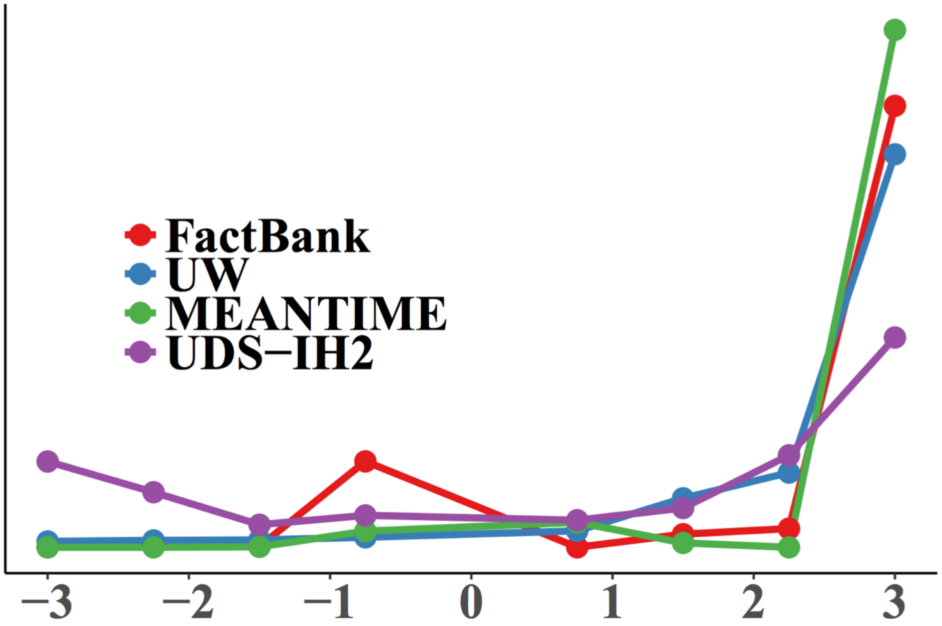


Relative Frequency of Factuality Labels



It-Happened shows more entropy in the distribution of labels

Relative Frequency of Factuality Labels



It-Happened shows more entropy in the distribution of labels

Higher entropy likely due to better genre distribution in UD

Examples from UDS-IH2

“Give me a call Tuesday afternoon to discuss (gone to Kelowna golfing for the weekend)”

Examples from UDS-IH2

DIDN'T HAPPEN!

$\boxed{\text{Give}}$ me a call Tuesday afternoon to discuss (gone to Kelowna golfing for the weekend)

Examples from UDS-IH2

DIDN'T HAPPEN! DIDN'T HAPPEN!

Give me a call Tuesday afternoon to discuss (gone to Kelowna golfing for the weekend)

Examples from UDS-IH2

DIDN'T HAPPEN! **DIDN'T HAPPEN!**

Give me a call Tuesday afternoon to discuss (gone to Kelowna golfing for the weekend)

HAPPENED!

Examples from UDS-IH2

DIDN'T HAPPEN! **DIDN'T HAPPEN!**

Give me a call Tuesday afternoon to discuss (gone to Kelowna golfing for the weekend)

HAPPENED! **HAPPENED!**

Examples from UDS-IH2

I <3 Max’s

Examples from UDS-IH2

I <3 Max’s

Models

Prior work

- Hand-engineered feature (templates)

Prior work

- Hand-engineered feature (templates)
  - Rule-based factuality computation based on type-level operator lexicon

    Nairn et al. 2006, Saurí 2008, Lotan et al. 2013

Signature Features

(+) Pat **failed** to eat lunch. → (–) Pat did not eat lunch.

(–) Pat did not **fail** to eat lunch. → (+) Pat ate lunch.

Signatures

fail to: –|+

Nairn et al. 2006, Karttunen 2012

Signature Features

(+) Pat **failed** to eat lunch. → (–) Pat did not eat lunch.

(–) Pat did not **fail** to eat lunch. → (+) Pat ate lunch.

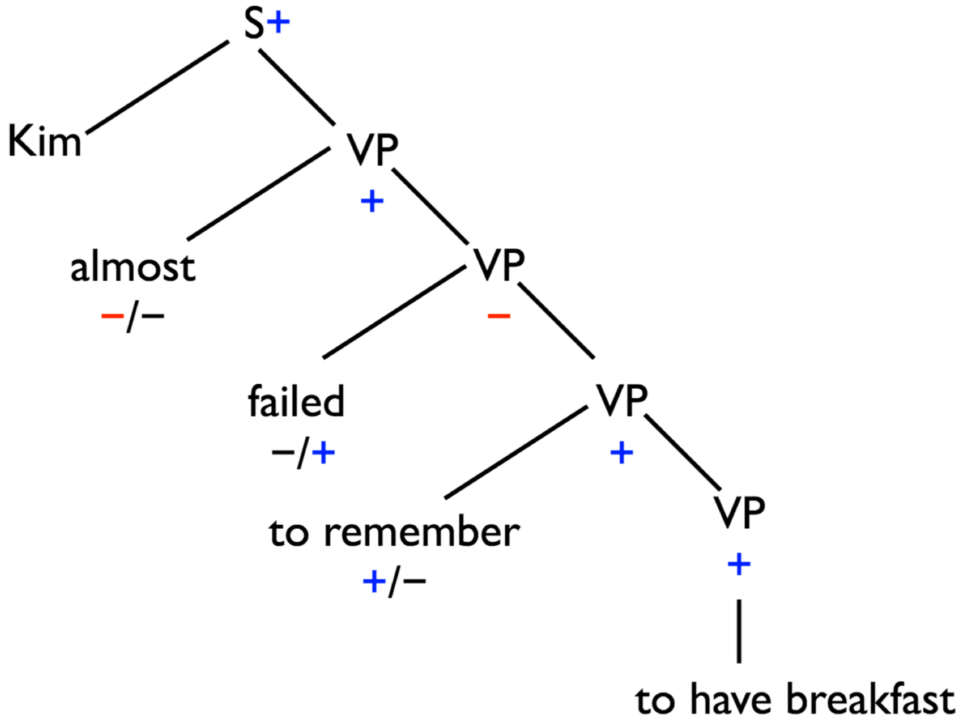
(+) Pat **managed** to eat lunch. → (+) Pat ate lunch.

(–) Pat did not **manage** to eat lunch. → (–) Pat did not eat lunch.

Signatures	
fail to:	–|+
manage to:	+|–
...	

Nairn et al. 2006, Karttunen 2012

Recursive Signature Application



https://web.stanford.edu/~laurik/presentations/CICLing.pdf

Prior work

- Hand-engineered feature (templates)
  - Rule-based factuality computation based on type-level operator lexicon
    Nairn et al. 2006, Saurí 2008, Lotan et al. 2013
  - Automatically extracted features + ML model;
    de Marneffe et al. 2012, Lee et al. 2016

Prior work

- Hand-engineered feature (templates)
  - Rule-based factuality computation based on type-level operator lexicon
    Nairn et al. 2006, Saurí 2008, Lotan et al. 2013
  - Automatically extracted features + ML model;
    de Marneffe et al. 2012, Lee et al. 2016
  - Combination of both strategies Stanovsky et al. 2017

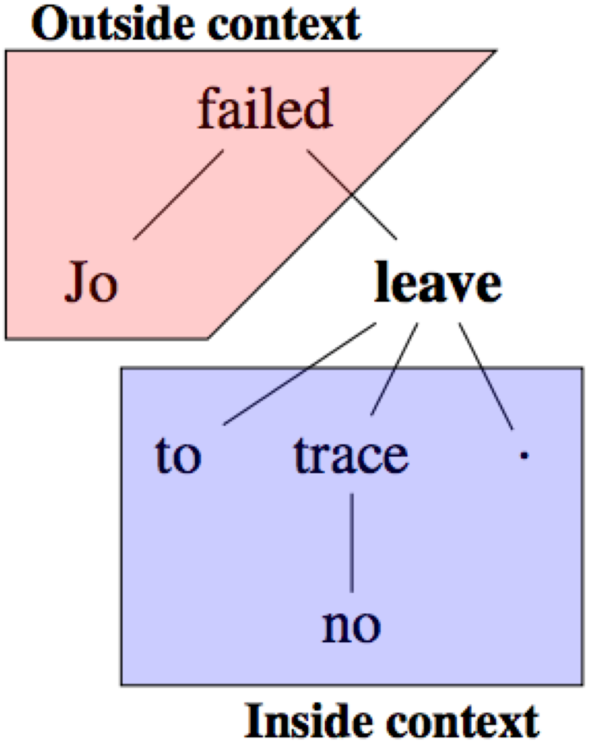
Our approach

1. **Learned features** using neural model w/ access to **inside** and **outside context**

Inside and outside context

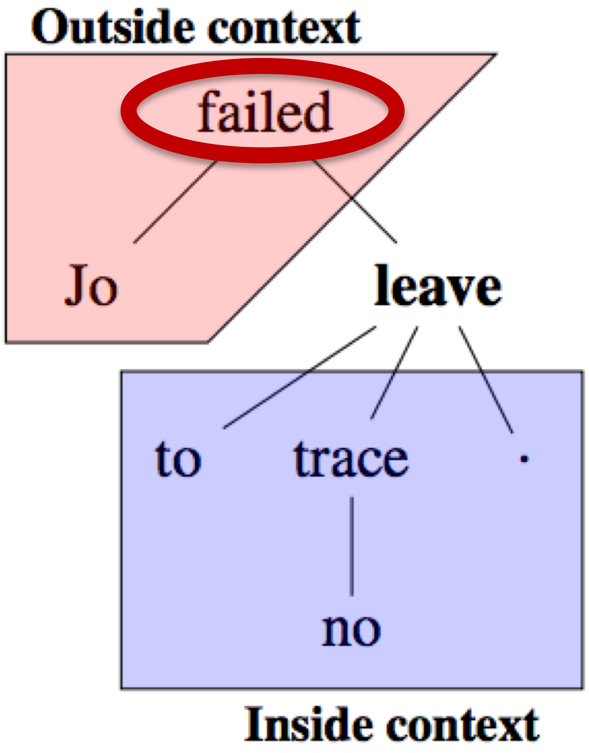
Lexical items and structures in both the **inside and outside context** matter for **factuality**

Inside and outside context



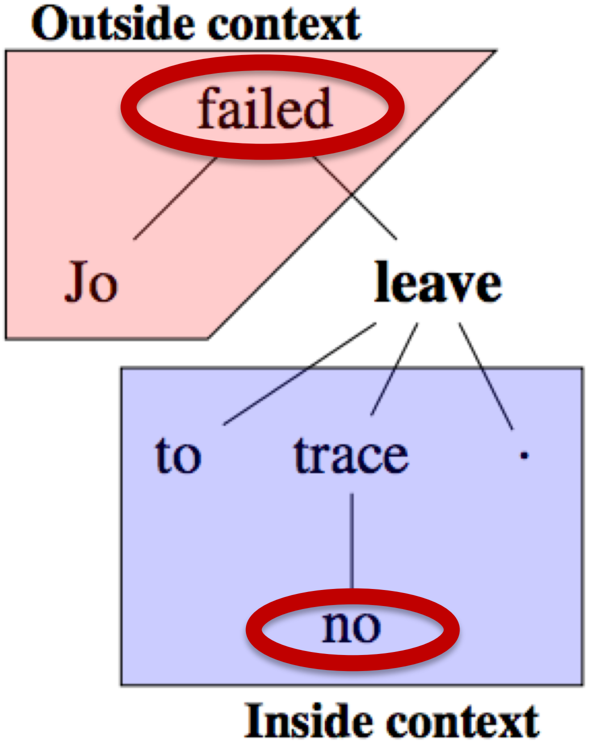
Lexical items and structures in both the **inside and outside context** matter for **factuality**

Inside and outside context



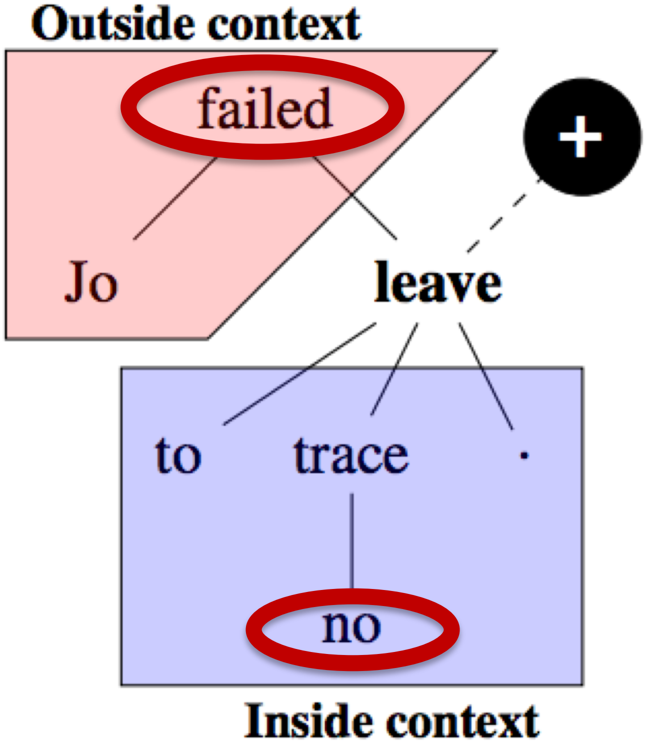
Lexical items and structures in both the **inside and outside context** matter for **factuality**

Inside and outside context



Lexical items and structures in both the **inside and outside context** matter for **factuality**

Inside and outside context



Lexical items and structures in both the **inside and outside context** matter for **factuality**

Our approach

1. **Learned features** with access to both **inside** and **outside context**

Our approach

1. **Learned features** with access to both **inside** and **outside context**

   (using bidirectional LSTMs)

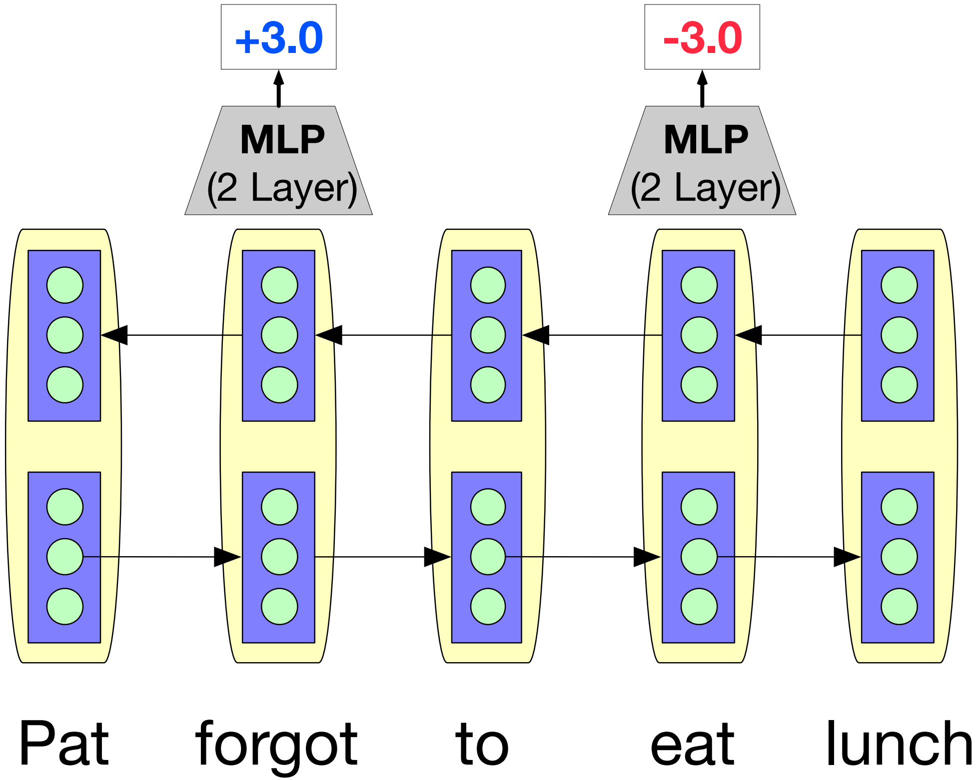
Our approach

1. **Learned features** with access to both **inside** and **outside context**
   (using bidirectional LSTMs)
2. **Push simple neural models** as far as they can go with **various training regimes** and addition of **linguistically motivated type-level features**

Our Models

- **L(inear chain)-biLSTM**

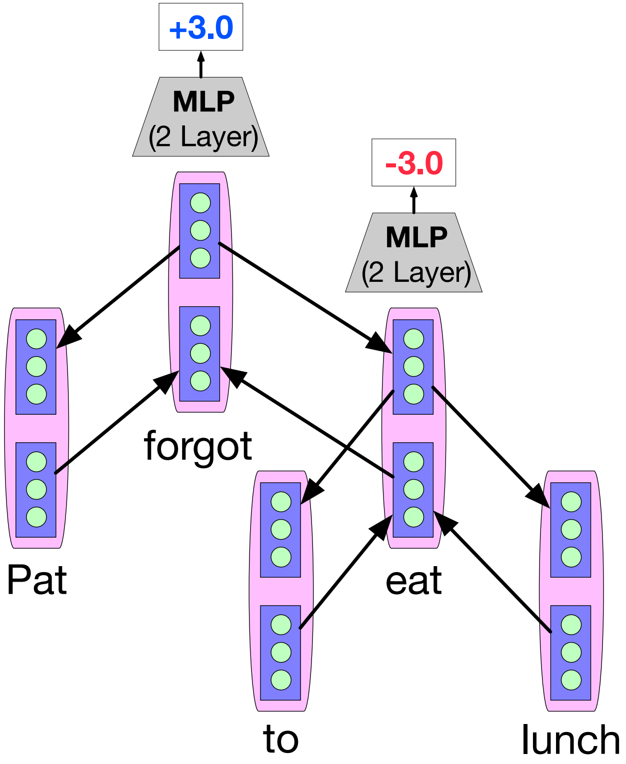
Model 1: Linear biLSTM + Regression



Our Models

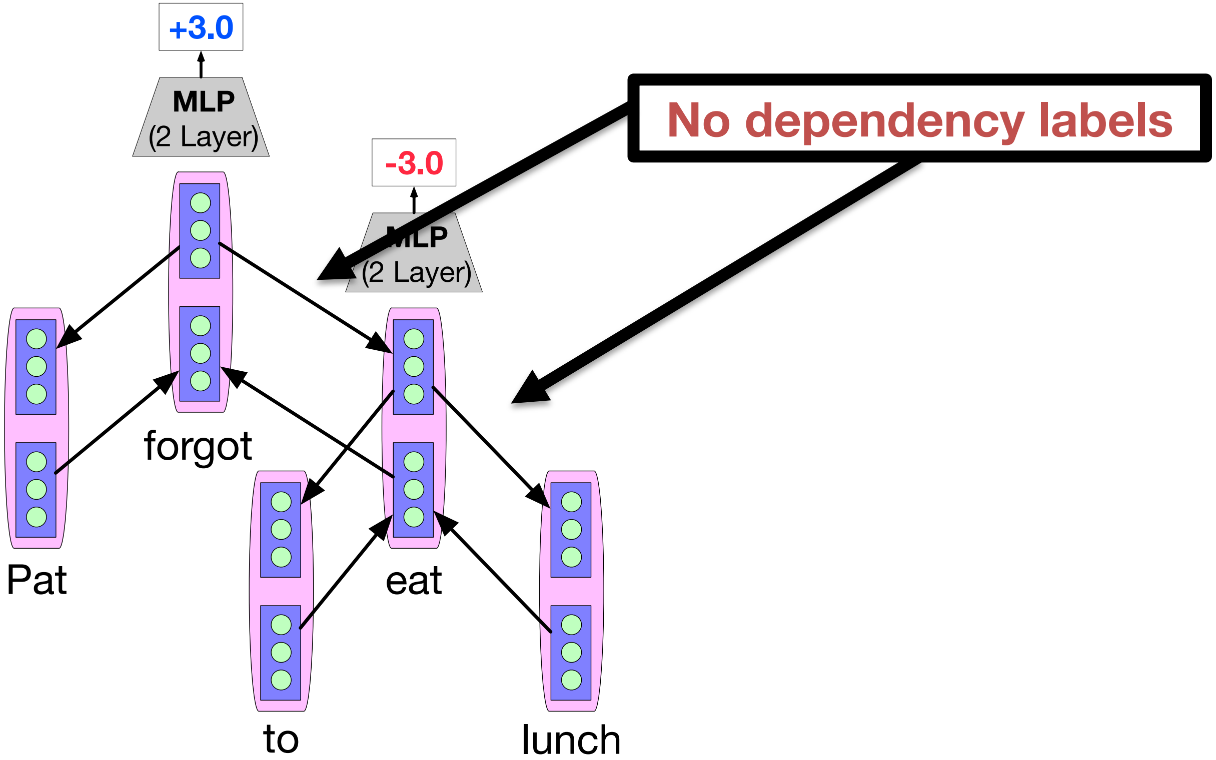
- **L(inear chain)-biLSTM**
- **(Dependency) T(ree)-biLSTM**

Model 2: Child Sum Tree biLSTM + Regression



Extension of Tai et al. 2015; see also Miwa & Bansal 2016

Model 2: Child Sum Tree biLSTM + Regression

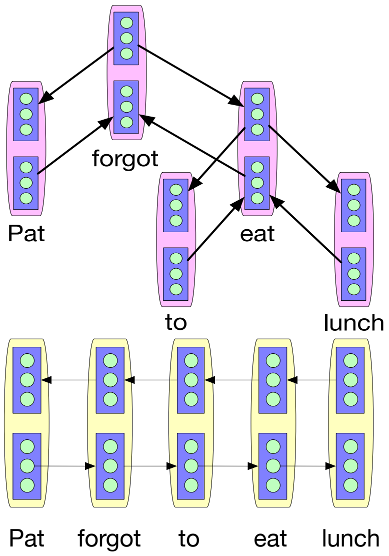


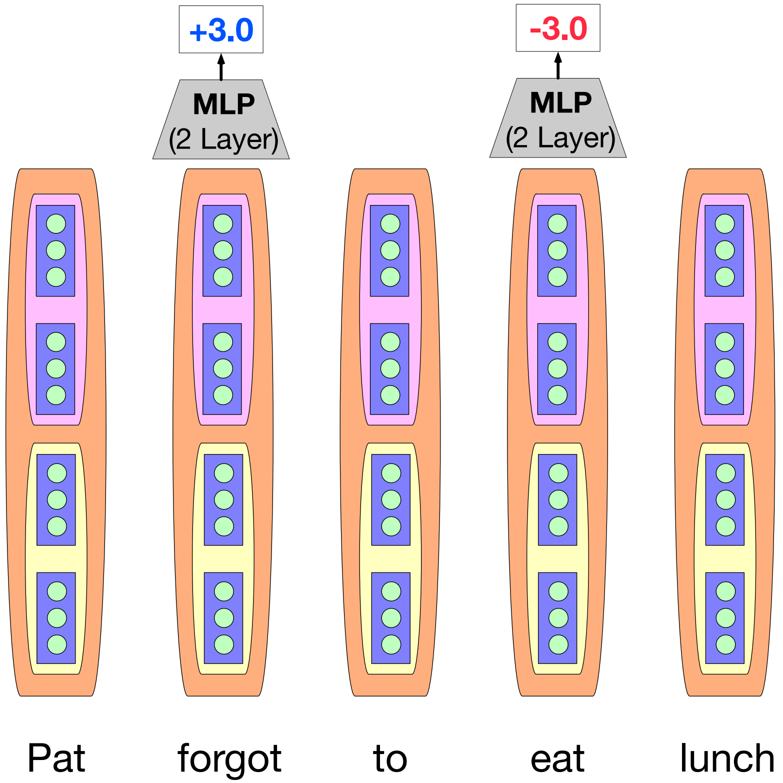
Extension of Tai et al. 2015; see also Miwa & Bansal 2016

Our Models

- **L(inear chain)-biLSTM**
- **(Dependency) T(ree)-biLSTM**
- **H(ybrid)-biLSTM** (parallel L- & T-biLSTMs)

Model 3: Hybrid (Linear + Tree)





Our Models

- **L(inear chain)-biLSTM**
- **(Dependency) T(ree)-biLSTM**
- **H(ybrid)-biLSTM** (parallel L- & T-biLSTMs)

Our Models

- **L(inear chain)-biLSTM**
- **(Dependency) T(ree)-biLSTM**
- **H(ybrid)-biLSTM** (parallel L- & T-biLSTMs)

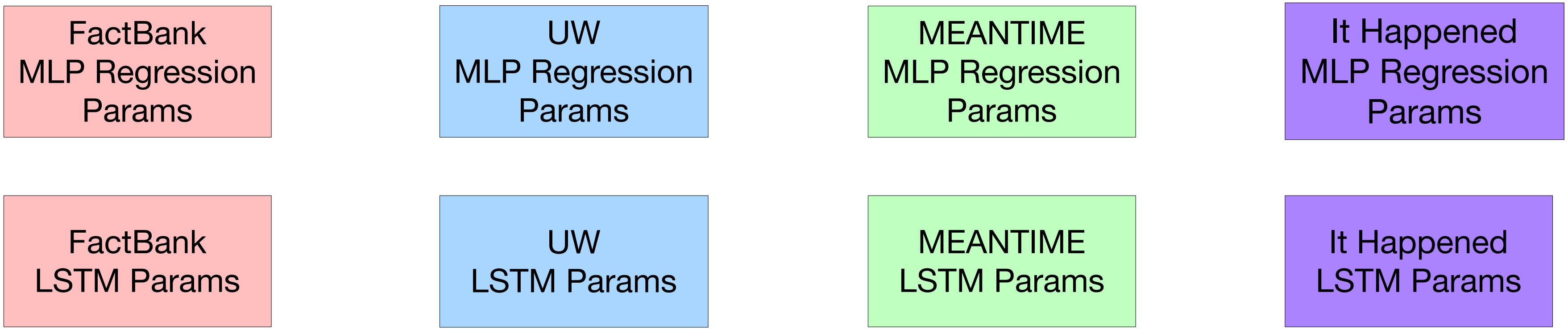
Aim: barebones models that can capture features in both contexts.

Training Regimes

- **Two settings**
  - Single-task

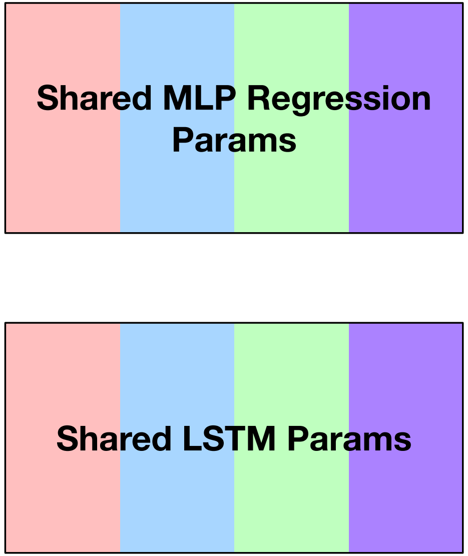
Single-task Specific

A separate network for each dataset.



Single-task General

A single network.



Training Regimes

- **Two settings**
  - Single-task

Training Regimes

- **Two settings**
  - Single-task
  - Multi-task

“Multi-task” Training Regimes

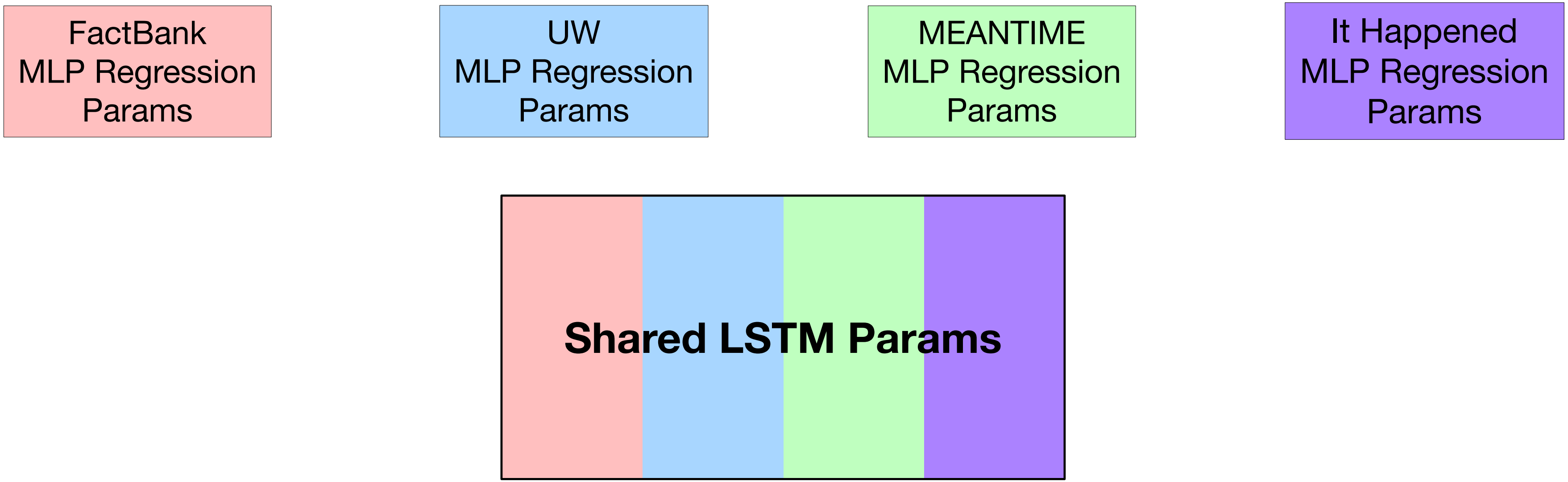
Each dataset collected under slightly **different protocols** and may capture slightly **different aspects of factuality**

Idea: treat each factuality dataset as a task.

FactBank **UW** **Meantime** **It Happened**

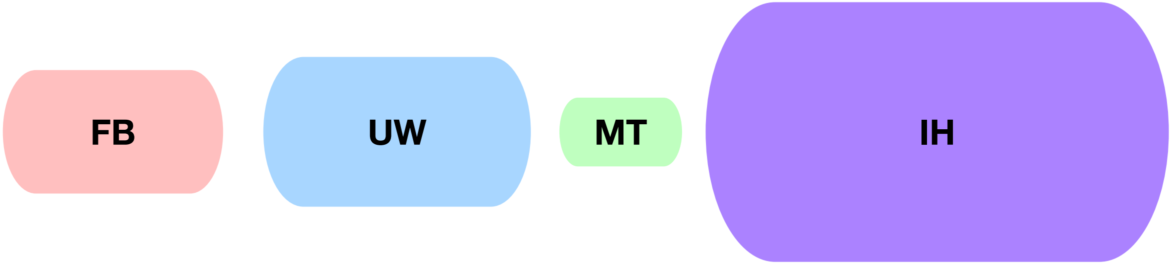
Multi-task

A single network with separate regression parameters for each dataset.



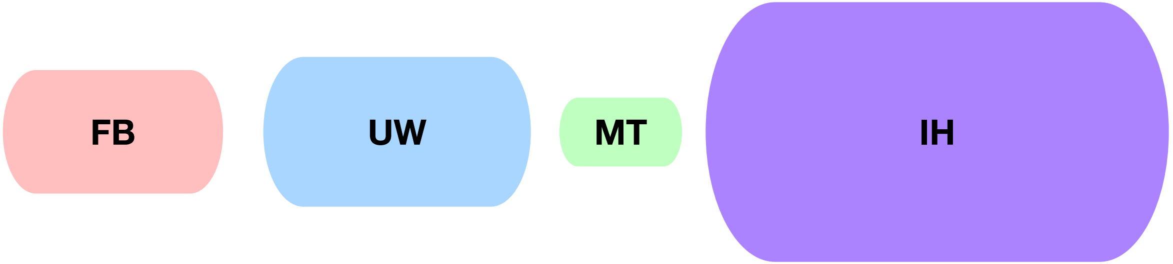
Multi-task Sampling Strategies

1. Simple. Concatenate the datasets, no upsampling.

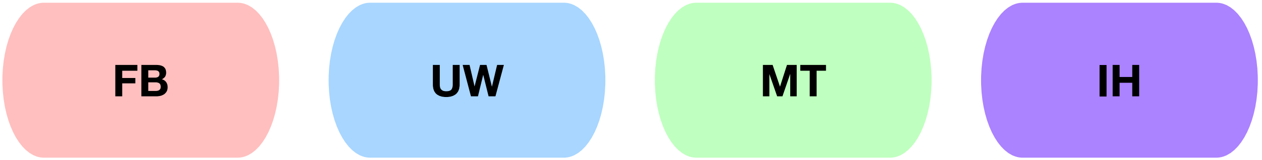


Multi-task Sampling Strategies

1. Simple.
Concatenate the datasets, no upsampling.

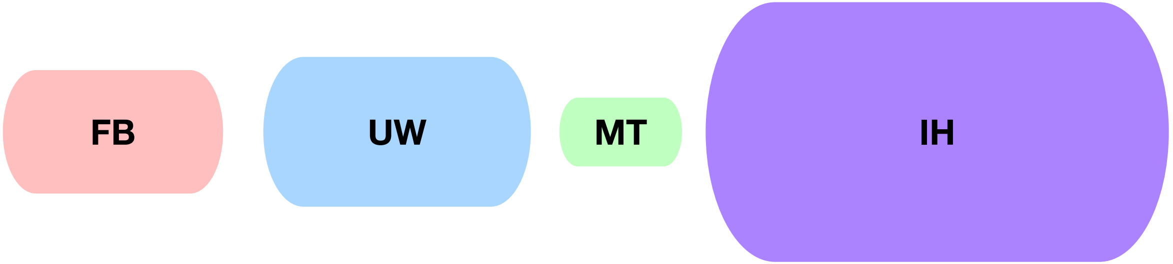


2. Balanced.
Upsample smaller datasets until uniform.

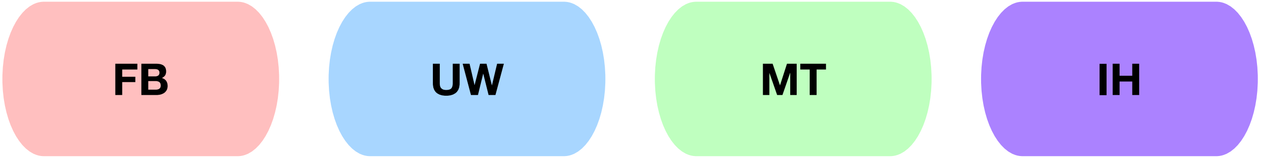


Multi-task Sampling Strategies

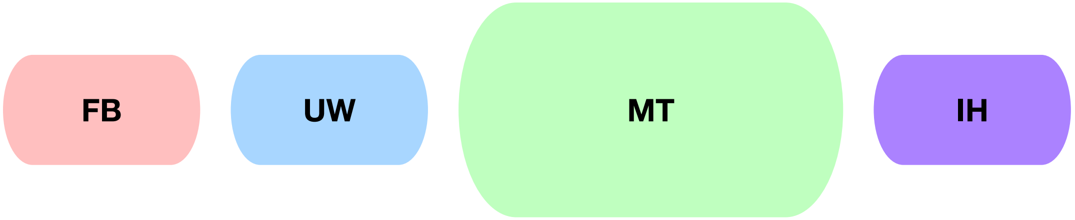
1. Simple.
Concatenate the datasets, no upsampling.



2. Balanced.
Upsample smaller datasets until uniform.



3. Focused.
Target dataset is 50% of all samples. Other datasets are divided uniformly.



Linguistically-Motivated Features

- Type-level, appended to *input* embeddings.

Linguistically-Motivated Features

- Type-level, appended to *input* embeddings.
- May propagate across hidden states.

Linguistically-Motivated Features

- Type-level, appended to *input* embeddings.
- May propagate across hidden states.
- Two kinds of features:

Linguistically-Motivated Features

- Type-level, appended to *input* embeddings.
- May propagate across hidden states.
- Two kinds of features:
  - Signature features (described earlier)

Linguistically-Motivated Features

- Type-level, appended to *input* embeddings.
- May propagate across hidden states.
- Two kinds of features:
  - Signature features (described earlier)
  - Mined features: built using tense agreement score Pavlick and Callison-Burch, 2016

Mined Features

"There is a curious restriction that the main sentence containing an implicative predicate and the complement sentence necessarily agree in tense."

Karttunen, 1971

Pat managed to eat lunch yesterday.

Pat managed to eat lunch tomorrow.

Pat wanted to eat lunch yesterday.

Pat wanted to eat lunch tomorrow.

Mined Features

Pavlick and Callison-Burch, 2016

- Mine implicatives from text based on Karttunen's tense constraint, using NLP pipeline.
- Tense agreement score =

  #(agree) / #(agree+disagree)

venture to	1.00	try to	0.42
forget to	0.80	agree to	0.34
manage to	0.79	promise to	0.22
bother to	0.61	want to	0.14
happen to	0.59	intend to	0.12
get to	0.52	plan to	0.10
decide to	0.45	hope to	0.03
dare to	0.44		

Our replication of P&C

- Simple text-matching patters over Common Crawl (3B sentences):

  `I $VERB to ___ $TIME`

dare to	1.00	intend to	0.83
bother to	1.00	want to	0.77
happen to	0.99	decide to	0.75
forget to	0.99	promise to	0.75
manage to	0.97	agree to	0.35
try to	0.96	plan to	0.20
get to	0.90	hope to	0.05
venture to	0.85		

Results

Summary results

	FactBank		UW		Meantime		UDS-IH2	
	MAE	r	MAE	r	MAE	r	MAE	r
All-3.0	0.8	NAN	0.78	NAN	0.31	NAN	2.255	NAN
Lee et al. 2015	-	-	0.511	0.708	-	-	-	-
Stanovsky et al. 2017	0.59	0.71	**0.42**†	0.66	0.34	0.47	-	-
L-biLSTM(2)-S	**0.427**	**0.826**	0.508	**0.719**	0.427	0.335	**0.960**†	**0.768**
T-biLSTM(2)-S	**0.577**	**0.752**	0.600	0.645	0.428	0.094	**1.101**	**0.704**
L-biLSTM(2)-G	**0.412**	**0.812**	0.523	0.703	0.409	0.462	-	-
T-biLSTM(2)-G	**0.455**	**0.809**	0.567	0.688	0.396	0.368	-	-
L-biLSTM(2)-S+lexfeats	**0.429**	**0.796**	0.495	**0.730**	0.427	0.322	**1.000**	**0.755**
T-biLSTM(2)-S+lexfeats	**0.542**	**0.744**	0.567	0.676	0.375	0.242	**1.087**	**0.719**
L-biLSTM(2)-MultiSimp	**0.353**	**0.843**	0.503	**0.725**	0.345	**0.540**	-	-
T-biLSTM(2)-MultiSimp	**0.482**	**0.803**	0.599	0.645	0.545	0.237	-	-
L-biLSTM(2)-MultiBal	**0.391**	**0.821**	0.496	**0.724**	**0.278**	**0.613**†	-	-
T-biLSTM(2)-MultiBal	**0.517**	**0.788**	0.573	0.659	0.400	0.405	-	-
L-biLSTM(1)-MultiFoc	**0.343**	**0.823**	0.516	0.698	**0.229**†	**0.599**	-	-
L-biLSTM(2)-MultiFoc	**0.314**	**0.846**	0.502	**0.710**	**0.305**	0.377	-	-
T-biLSTM(2)-MultiFoc	1.100	0.234	0.615	0.616	0.395	0.300	-	-
L-biLSTM(2)-MultiSimp w/UDS-IH2	**0.377**	**0.828**	0.508	**0.722**	0.367	0.469	**0.965**	**0.771**†
T-biLSTM(2)-MultiSimp w/UDS-IH2	0.595	**0.716**	0.598	0.609	0.467	0.345	**1.072**	**0.723**
H-biLSTM(2)-S	0.488	**0.775**	0.526	**0.714**	0.442	0.255	**0.967**	**0.768**
H-biLSTM(1)-MultiSimp	**0.313**†	**0.857**†	0.528	0.704	0.314	0.545	-	-
H-biLSTM(2)-MultiSimp	**0.431**	**0.808**	0.514	**0.723**	0.401	0.461	-	-
H-biLSTM(2)-MultiBal	**0.386**	**0.825**	0.502	**0.713**	0.352	**0.564**	-	-
H-biLSTM(2)-MultiSimp w/UDS-IH2	**0.393**	**0.820**	0.481	**0.749**†	0.374	**0.495**	**0.969**	**0.760**

Table 4: All 2-layer systems and overall best systems (shaded in purple). State-of-the-art results in bold. † indicates best in column. Key: L=linear, T=tree, H=hybrid, (1,2)=# layers, S=single-task specific, G=single-task general, +lexfeats=with all lexical features, MultiSimp=multi-task simple, MultiBal=multi-task balanced, MultiFoc=multi-task focused, w/UDS-IH2=trained on all data including UDS-IH2. All-3.0 is a constant baseline, always predicting 3.0.

Summary results

	FactBank		UW		Meantime		UDS-IH2	
	MAE	r	MAE	r	MAE	r	MAE	r
All-3.0	0.8	NAN	0.78	NAN	0.31	NAN	2.255	NAN
Lee et al. 2015	-	-	0.511	0.708	-	-	-	-
Stanovsky et al. 2017	0.59	0.71	**0.42**†	0.66	0.34	0.47	-	-
L-biLSTM(2)-S	**0.427**	**0.826**	0.508	**0.719**	0.427	0.335	**0.960**†	**0.768**
T-biLSTM(2)-S	**0.577**	**0.752**	0.600	0.645	0.428	0.094	**1.101**	**0.704**
L-biLSTM(2)-G	**0.412**	**0.812**	0.523	0.703	0.409	0.462	-	-
T-biLSTM(2)-G	**0.455**	**0.809**	0.567	0.688	0.396	0.368	-	-
L-biLSTM(2)-S+lexfeats	**0.429**	**0.796**	0.495	**0.730**	0.427	0.322	**1.000**	**0.755**

TOO MUCH INFO!

L-biLSTM(2)-MultiFoc	**0.314**	**0.846**	0.502	**0.710**	**0.305**	0.377	-	-
T-biLSTM(2)-MultiFoc	1.100	0.234	0.615	0.616	0.395	0.300	-	-
L-biLSTM(2)-MultiSimp w/UDS-IH2	**0.377**	**0.828**	0.508	**0.722**	0.367	0.469	**0.965**	**0.771**†
T-biLSTM(2)-MultiSimp w/UDS-IH2	0.595	**0.716**	0.598	0.609	0.467	0.345	**1.072**	**0.723**
H-biLSTM(2)-S	0.488	**0.775**	0.526	**0.714**	0.442	0.255	**0.967**	**0.768**
H-biLSTM(1)-MultiSimp	**0.313**†	**0.857**†	0.528	0.704	0.314	0.545	-	-
H-biLSTM(2)-MultiSimp	**0.431**	**0.808**	0.514	**0.723**	0.401	0.461	-	-
H-biLSTM(2)-MultiBal	**0.386**	**0.825**	0.502	**0.713**	0.352	**0.564**	-	-
H-biLSTM(2)-MultiSimp w/UDS-IH2	**0.393**	**0.820**	0.481	**0.749**†	0.374	**0.495**	**0.969**	**0.760**

Table 4: All 2-layer systems and overall best systems (shaded in purple). State-of-the-art results in bold. † indicates best in column. Key: L=linear, T=tree, H=hybrid, (1,2)=# layers, S=single-task specific, G=single-task general, +lexfeats=with all lexical features, MultiSimp=multi-task simple, MultiBal=multi-task balanced, MultiFoc=multi-task focused, w/UDS-IH2=trained on all data including UDS-IH2. All-3.0 is a constant baseline, always predicting 3.0.

Summary results

	FactBank		UW		Meantime		UDS-IH2	
	MAE	r	MAE	r	MAE	r	MAE	r
All-3.0	0.8	NAN	0.78	NAN	0.31	NAN	2.255	NAN
Lee et al. 2015	-	-	0.511	0.708	-	-	-	-
Stanovsky et al. 2017	0.59	0.71	**0.42**†	0.66	0.34	0.47	-	-
L-biLSTM(2)-S	**0.427**	**0.826**	0.508	**0.719**	0.427	0.335	**0.960**†	**0.768**
T-biLSTM(2)-S	**0.577**	**0.752**	0.600	0.645	0.428	0.094	**1.101**	**0.704**
L-biLSTM(2)-G	**0.412**	**0.812**	0.523	0.703	0.409	0.462	-	-
T-biLSTM(2)-G	**0.455**	**0.809**	0.567	0.688	0.396	0.368	-	-
L-biLSTM(2)-S+lexfeats	**0.429**	**0.796**	0.495	**0.730**	0.427	0.322	**1.000**	**0.755**
T-biLSTM(2)-S+lexfeats	**0.542**	**0.744**	0.567	0.676	0.375	0.242	**1.087**	**0.719**
L-biLSTM(2)-MultiSimp	**0.353**	**0.843**	0.503	**0.725**	0.345	**0.540**	-	-
T-biLSTM(2)-MultiSimp	**0.482**	**0.803**	0.599	0.645	0.545	0.237	-	-
L-biLSTM(2)-MultiBal	**0.391**	**0.821**	0.496	**0.724**	**0.278**	**0.613**†	-	-
T-biLSTM(2)-MultiBal	**0.517**	**0.788**	0.573	0.659	0.400	0.405	-	-
L-biLSTM(1)-MultiFoc	**0.343**	**0.823**	0.516	0.698	**0.229**†	**0.599**	-	-
L-biLSTM(2)-MultiFoc	**0.314**	**0.846**	0.502	**0.710**	**0.305**	0.377	-	-
T-biLSTM(2)-MultiFoc	1.100	0.234	0.615	0.616	0.395	0.300	-	-
L-biLSTM(2)-MultiSimp w/UDS-IH2	**0.377**	**0.828**	0.508	**0.722**	0.367	0.469	**0.965**	**0.771**†
T-biLSTM(2)-MultiSimp w/UDS-IH2	0.595	**0.716**	0.598	0.609	0.467	0.345	**1.072**	**0.723**
H-biLSTM(2)-S	0.488	**0.775**	0.526	**0.714**	0.442	0.255	**0.967**	**0.768**
H-biLSTM(1)-MultiSimp	**0.313**†	**0.857**†	0.528	0.704	0.314	0.545	-	-
H-biLSTM(2)-MultiSimp	**0.431**	**0.808**	0.514	**0.723**	0.401	0.461	-	-
H-biLSTM(2)-MultiBal	**0.386**	**0.825**	0.502	**0.713**	0.352	**0.564**	-	-
H-biLSTM(2)-MultiSimp w/UDS-IH2	**0.393**	**0.820**	0.481	**0.749**†	0.374	**0.495**	**0.969**	**0.760**

Table 4: All 2-layer systems and overall best systems (shaded in purple). State-of-the-art results in bold. † indicates best in column. Key: L=linear, T=tree, H=hybrid, (1,2)=# layers, S=single-task specific, G=single-task general, +lexfeats=with all lexical features, MultiSimp=multi-task simple, MultiBal=multi-task balanced, MultiFoc=multi-task focused, w/UDS-IH2=trained on all data including UDS-IH2. All-3.0 is a constant baseline, always predicting 3.0.

Summary results

	FactBank		UW		Meantime		UDS-IH2	
	MAE	r	MAE	r	MAE	r	MAE	r
All-3.0	0.8	NAN	0.78	NAN	0.31	NAN	2.255	NAN
Lee et al. 2015	-	-	0.511	0.708	-	-	-	-
Stanovsky et al. 2017	0.59	0.71	**0.42**†	0.66	0.34	0.47	-	-
L-biLSTM(2)-S	**0.427**	**0.826**	0.508	**0.719**	0.427	0.335	**0.960**†	**0.768**
T-biLSTM(2)-S	**0.577**	**0.752**	0.600	0.645	0.428	0.094	**1.101**	**0.704**
L-biLSTM(2)-G	**0.412**	**0.812**	0.523	0.703	0.409	0.462	-	-
T-biLSTM(2)-G	**0.455**	**0.809**	0.567	0.688	0.396	0.368	-	-
L-biLSTM(2)-S+lexfeats	**0.429**	**0.796**	0.495	**0.730**	0.427	0.322	**1.000**	**0.755**
T-biLSTM(2)-S+lexfeats	**0.542**	**0.744**	0.567	0.676	0.375	0.242	**1.087**	**0.719**
L-biLSTM(2)-MultiSimp	**0.353**	**0.843**	0.503	**0.725**	0.345	**0.540**	-	-
T-biLSTM(2)-MultiSimp	**0.482**	**0.803**	0.599	0.645	0.545	0.237	-	-
L-biLSTM(2)-MultiBal	**0.391**	**0.821**	0.496	**0.724**	**0.278**	**0.613**†	-	-
T-biLSTM(2)-MultiBal	**0.517**	**0.788**	0.573	0.659	0.400	0.405	-	-
L-biLSTM(1)-MultiFoc	**0.343**	**0.823**	0.516	0.698	**0.229**†	**0.599**	-	-
L-biLSTM(2)-MultiFoc	**0.314**	**0.846**	0.502	**0.710**	**0.305**	0.377	-	-
T-biLSTM(2)-MultiFoc	1.100	0.234	0.615	0.616	0.395	0.300	-	-
L-biLSTM(2)-MultiSimp w/UDS-IH2	**0.377**	**0.828**	0.508	**0.722**	0.367	0.469	**0.965**	**0.771**†
T-biLSTM(2)-MultiSimp w/UDS-IH2	0.595	**0.716**	0.598	0.609	0.467	0.345	**1.072**	**0.723**
H-biLSTM(2)-S	0.488	**0.775**	0.526	**0.714**	0.442	0.255	**0.967**	**0.768**
H-biLSTM(1)-MultiSimp	**0.313**†	**0.857**†	0.528	0.704	0.314	0.545	-	-
H-biLSTM(2)-MultiSimp	**0.431**	**0.808**	0.514	**0.723**	0.401	0.461	-	-
H-biLSTM(2)-MultiBal	**0.386**	**0.825**	0.502	**0.713**	0.352	**0.564**	-	-
H-biLSTM(2)-MultiSimp w/UDS-IH2	**0.393**	**0.820**	0.481	**0.749**†	0.374	**0.495**	**0.969**	**0.760**

Table 4: All 2-layer systems and overall best systems (shaded in purple). State-of-the-art results in bold. † indicates best in column. Key: L=linear, T=tree, H=hybrid, (1,2)=# layers, S=single-task specific, G=single-task general, +lexfeats=with all lexical features, MultiSimp=multi-task simple, MultiBal=multi-task balanced, MultiFoc=multi-task focused, w/UDS-IH2=trained on all data including UDS-IH2. All-3.0 is a constant baseline, always predicting 3.0.

Summary results

	FactBank		UW		Meantime		UDS-IH2	
	MAE	r	MAE	r	MAE	r	MAE	r
All-3.0	0.8	NAN	0.78	NAN	0.31	NAN	2.255	NAN
Lee et al. 2015	-	-	0.511	0.708	-	-	-	-
Stanovsky et al. 2017	0.59	0.71	**0.42**†	0.66	0.34	0.47	-	-
L-biLSTM(2)-S	**0.427**	**0.826**	0.508	**0.719**	0.427	0.335	**0.960**†	**0.768**
T-biLSTM(2)-S	**0.577**	**0.752**	0.600	0.645	0.428	0.094	**1.101**	**0.704**
L-biLSTM(2)-G	**0.412**	**0.812**	0.523	0.703	0.409	0.462	-	-
T-biLSTM(2)-G	**0.455**	**0.809**	0.567	0.688	0.396	0.368	-	-
L-biLSTM(2)-S+lexfeats	**0.429**	**0.796**	0.495	**0.730**	0.427	0.322	**1.000**	**0.755**
T-biLSTM(2)-S+lexfeats	**0.542**	**0.744**	0.567	0.676	0.375	0.242	**1.087**	**0.719**
L-biLSTM(2)-MultiSimp	**0.353**	**0.843**	0.503	**0.725**	0.345	**0.540**	-	-
T-biLSTM(2)-MultiSimp	**0.482**	**0.803**	0.599	0.645	0.545	0.237	-	-
L-biLSTM(2)-MultiBal	**0.391**	**0.821**	0.496	**0.724**	**0.278**	**0.613**†	-	-
T-biLSTM(2)-MultiBal	**0.517**	**0.788**	0.573	0.659	0.400	0.405	-	-
L-biLSTM(1)-MultiFoc	**0.343**	**0.823**	0.516	0.698	**0.229**†	**0.599**	-	-
L-biLSTM(2)-MultiFoc	**0.314**	**0.846**	0.502	**0.710**	**0.305**	0.377	-	-
T-biLSTM(2)-MultiFoc	1.100	0.234	0.615	0.616	0.395	0.300	-	-
L-biLSTM(2)-MultiSimp w/UDS-IH2	**0.377**	**0.828**	0.508	**0.722**	0.367	0.469	**0.965**	**0.771**†
T-biLSTM(2)-MultiSimp w/UDS-IH2	0.595	**0.716**	0.598	0.609	0.467	0.345	**1.072**	**0.723**
H-biLSTM(2)-S	0.488	**0.775**	0.526	**0.714**	0.442	0.255	**0.967**	**0.768**
H-biLSTM(1)-MultiSimp	**0.313**†	**0.857**†	0.528	0.704	0.314	0.545	-	-
H-biLSTM(2)-MultiSimp	**0.431**	**0.808**	0.514	**0.723**	0.401	0.461	-	-
H-biLSTM(2)-MultiBal	**0.386**	**0.825**	0.502	**0.713**	0.352	**0.564**	-	-
H-biLSTM(2)-MultiSimp w/UDS-IH2	**0.393**	**0.820**	0.481	**0.749**†	0.374	**0.495**	**0.969**	**0.760**

Table 4: All 2-layer systems and overall best systems (shaded in purple). State-of-the-art results in bold. † indicates best in column. Key: L=linear, T=tree, H=hybrid, (1,2)=# layers, S=single-task specific, G=single-task general, +lexfeats=with all lexical features, MultiSimp=multi-task simple, MultiBal=multi-task balanced, MultiFoc=multi-task focused, w/UDS-IH2=trained on all data including UDS-IH2. All-3.0 is a constant baseline, always predicting 3.0.

Summary results

	FactBank		UW		Meantime		UDS-IH2	
	MAE	r	MAE	r	MAE	r	MAE	r
All-3.0	0.8	NAN	0.78	NAN	0.31	NAN	2.255	NAN
Lee et al. 2015	-	-	0.511	0.708	-	-	-	-
Stanovsky et al. 2017	0.59	0.71	**0.42**†	0.66	0.34	0.47	-	-
L-biLSTM(2)-S	**0.427**	**0.826**	0.508	**0.719**	0.427	0.335	**0.960**†	**0.768**
T-biLSTM(2)-S	**0.577**	**0.752**	0.600	0.645	0.428	0.094	**1.101**	**0.704**
L-biLSTM(2)-G	**0.412**	**0.812**	0.523	0.703	0.409	0.462	-	-
T-biLSTM(2)-G	**0.455**	**0.809**	0.567	0.688	0.396	0.368	-	-
L-biLSTM(2)-S+lexfeats	**0.429**	**0.796**	0.495	**0.730**	0.427	0.322	**1.000**	**0.755**
T-biLSTM(2)-S+lexfeats	**0.542**	**0.744**	0.567	0.676	0.375	0.242	**1.087**	**0.719**
L-biLSTM(2)-MultiSimp	**0.353**	**0.843**	0.503	**0.725**	0.345	**0.540**	-	-
T-biLSTM(2)-MultiSimp	**0.482**	**0.803**	0.599	0.645	0.545	0.237	-	-
L-biLSTM(2)-MultiBal	**0.391**	**0.821**	0.496	**0.724**	**0.278**	**0.613**†	-	-
T-biLSTM(2)-MultiBal	**0.517**	**0.788**	0.573	0.659	0.400	0.405	-	-
L-biLSTM(1)-MultiFoc	**0.343**	**0.823**	0.516	0.698	**0.229**†	**0.599**	-	-
L-biLSTM(2)-MultiFoc	**0.314**	**0.846**	0.502	**0.710**	**0.305**	0.377	-	-
T-biLSTM(2)-MultiFoc	1.100	0.234	0.615	0.616	0.395	0.300	-	-
L-biLSTM(2)-MultiSimp w/UDS-IH2	**0.377**	**0.828**	0.508	**0.722**	0.367	0.469	**0.965**	**0.771**†
T-biLSTM(2)-MultiSimp w/UDS-IH2	0.595	**0.716**	0.598	0.609	0.467	0.345	**1.072**	**0.723**
H-biLSTM(2)-S	0.488	**0.775**	0.526	**0.714**	0.442	0.255	**0.967**	**0.768**
H-biLSTM(1)-MultiSimp	**0.313**†	**0.857**†	0.528	0.704	0.314	0.545	-	-
H-biLSTM(2)-MultiSimp	**0.431**	**0.808**	0.514	**0.723**	0.401	0.461	-	-
H-biLSTM(2)-MultiBal	**0.386**	**0.825**	0.502	**0.713**	0.352	**0.564**	-	-
H-biLSTM(2)-MultiSimp w/UDS-IH2	**0.393**	**0.820**	0.481	**0.749**†	0.374	**0.495**	**0.969**	**0.760**

Table 4: All 2-layer systems and overall best systems (shaded in purple). State-of-the-art results in bold. † indicates best in column. Key: L=linear, T=tree, H=hybrid, (1,2)=# layers, S=single-task specific, G=single-task general, +lexfeats=with all lexical features, MultiSimp=multi-task simple, MultiBal=multi-task balanced, MultiFoc=multi-task focused, w/UDS-IH2=trained on all data including UDS-IH2. All-3.0 is a constant baseline, always predicting 3.0.

Summary results

	FactBank		UW		Meantime		UDS-IH2	
	MAE	r	MAE	r	MAE	r	MAE	r
All-3.0	0.8	NAN	0.78	NAN	0.31	NAN	2.255	NAN
Lee et al. 2015	-	-	0.511	0.708	-	-	-	-
Stanovsky et al. 2017	0.59	0.71	**0.42**†	0.66	0.34	0.47	-	-
L-biLSTM(2)-S	**0.427**	**0.826**	0.508	**0.719**	0.427	0.335	**0.960**†	**0.768**
T-biLSTM(2)-S	**0.577**	**0.752**	0.600	0.645	0.428	0.094	**1.101**	**0.704**
L-biLSTM(2)-G	**0.412**	**0.812**	0.523	0.703	0.409	0.462	-	-
T-biLSTM(2)-G	**0.455**	**0.809**	0.567	0.688	0.396	0.368	-	-
L-biLSTM(2)-S+lexfeats	**0.429**	**0.796**	0.495	**0.730**	0.427	0.322	**1.000**	**0.755**
T-biLSTM(2)-S+lexfeats	**0.542**	**0.744**	0.567	0.676	0.375	0.242	**1.087**	**0.719**
L-biLSTM(2)-MultiSimp	**0.353**	**0.843**	0.503	**0.725**	0.345	**0.540**	-	-
T-biLSTM(2)-MultiSimp	**0.482**	**0.803**	0.599	0.645	0.545	0.237	-	-
L-biLSTM(2)-MultiBal	**0.391**	**0.821**	0.496	**0.724**	**0.278**	**0.613**†	-	-
T-biLSTM(2)-MultiBal	**0.517**	**0.788**	0.573	0.659	0.400	0.405	-	-
L-biLSTM(1)-MultiFoc	**0.343**	**0.823**	0.516	0.698	**0.229**†	**0.599**	-	-
L-biLSTM(2)-MultiFoc	**0.314**	**0.846**	0.502	**0.710**	**0.305**	0.377	-	-
T-biLSTM(2)-MultiFoc	1.100	0.234	0.615	0.616	0.395	0.300	-	-
L-biLSTM(2)-MultiSimp w/UDS-IH2	**0.377**	**0.828**	0.508	**0.722**	0.367	0.469	**0.965**	**0.771**†
T-biLSTM(2)-MultiSimp w/UDS-IH2	0.595	**0.716**	0.598	0.609	0.467	0.345	**1.072**	**0.723**
H-biLSTM(2)-S	0.488	**0.775**	0.526	**0.714**	0.442	0.255	**0.967**	**0.768**
H-biLSTM(1)-MultiSimp	**0.313**†	**0.857**†	0.528	0.704	0.314	0.545	-	-
H-biLSTM(2)-MultiSimp	**0.431**	**0.808**	0.514	**0.723**	0.401	0.461	-	-
H-biLSTM(2)-MultiBal	**0.386**	**0.825**	0.502	**0.713**	0.352	**0.564**	-	-
H-biLSTM(2)-MultiSimp w/UDS-IH2	**0.393**	**0.820**	0.481	**0.749**†	0.374	**0.495**	**0.969**	**0.760**

Table 4: All 2-layer systems and overall best systems (shaded in purple). State-of-the-art results in bold. † indicates best in column. Key: L=linear, T=tree, H=hybrid, (1,2)=# layers, S=single-task specific, G=single-task general, +lexfeats=with all lexical features, MultiSimp=multi-task simple, MultiBal=multi-task balanced, MultiFoc=multi-task focused, w/UDS-IH2=trained on all data including UDS-IH2. All-3.0 is a constant baseline, always predicting 3.0.

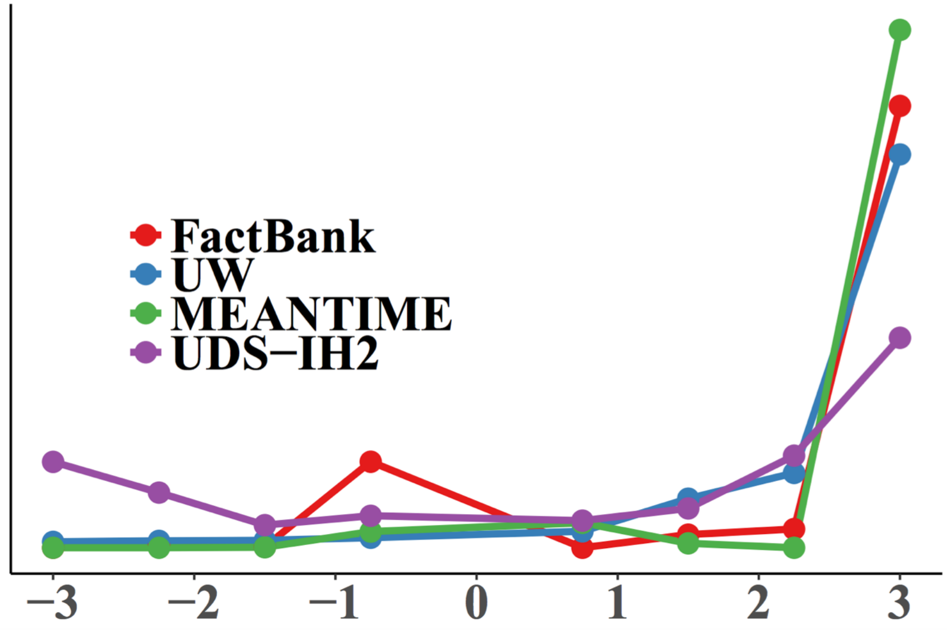
Summary results

	FactBank MAE	r	UW MAE	r	Meantime MAE	r	UDS-IH2 MAE	r
All-3.0	0.8	NAN	0.78	NAN	0.31	NAN	2.255	NAN
Lee et al. 2015	-	-	0.511	0.708	-	-	-	-
Stanovsky et al. 2017	0.59	0.71	**0.42**†	0.66	0.34	0.47	-	-
H-biLSTM(2)-S	0.488	**0.775**	0.526	**0.714**	0.442	0.255	**0.967**	**0.768**
H-biLSTM(1)-MultiSimp	**0.313**†	**0.857**†	0.528	0.704	0.314	0.545	-	-
H-biLSTM(2)-MultiSimp	**0.431**	**0.808**	0.514	**0.723**	0.401	0.461	-	-
H-biLSTM(2)-MultiBal	**0.386**	**0.825**	0.502	**0.713**	0.352	**0.564**	-	-
H-biLSTM(2)-MultiSimp w/UDS-IH2	**0.393**	**0.820**	0.481	**0.749**†	0.374	**0.495**	**0.969**	**0.760**

Better controls for (lack of) variance in rating distributions

Table 4: All 2-layer systems and overall best systems (shaded in purple). State-of-the-art results in bold. † indicates best in column. Key: L=linear, T=tree, H=hybrid, (1,2)=# layers, S=single-task specific, G=single-task general, +lexfeats=with all lexical features, MultiSimp=multi-task simple, MultiBal=multi-task balanced, MultiFoc=multi-task focused, w/UDS-IH2=trained on all data including UDS-IH2. All-3.0 is a constant baseline, always predicting 3.0.

Relative Frequency of Factuality Labels

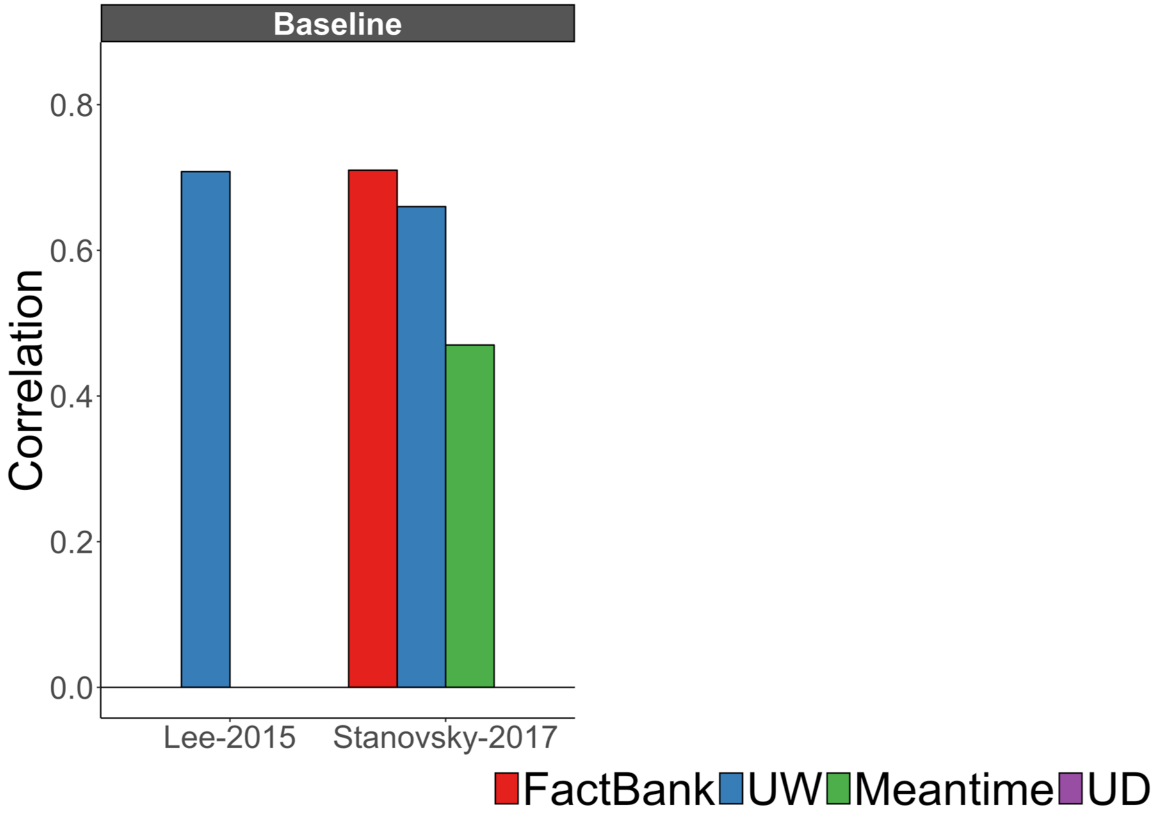


It-Happened shows more entropy in the distribution of labels

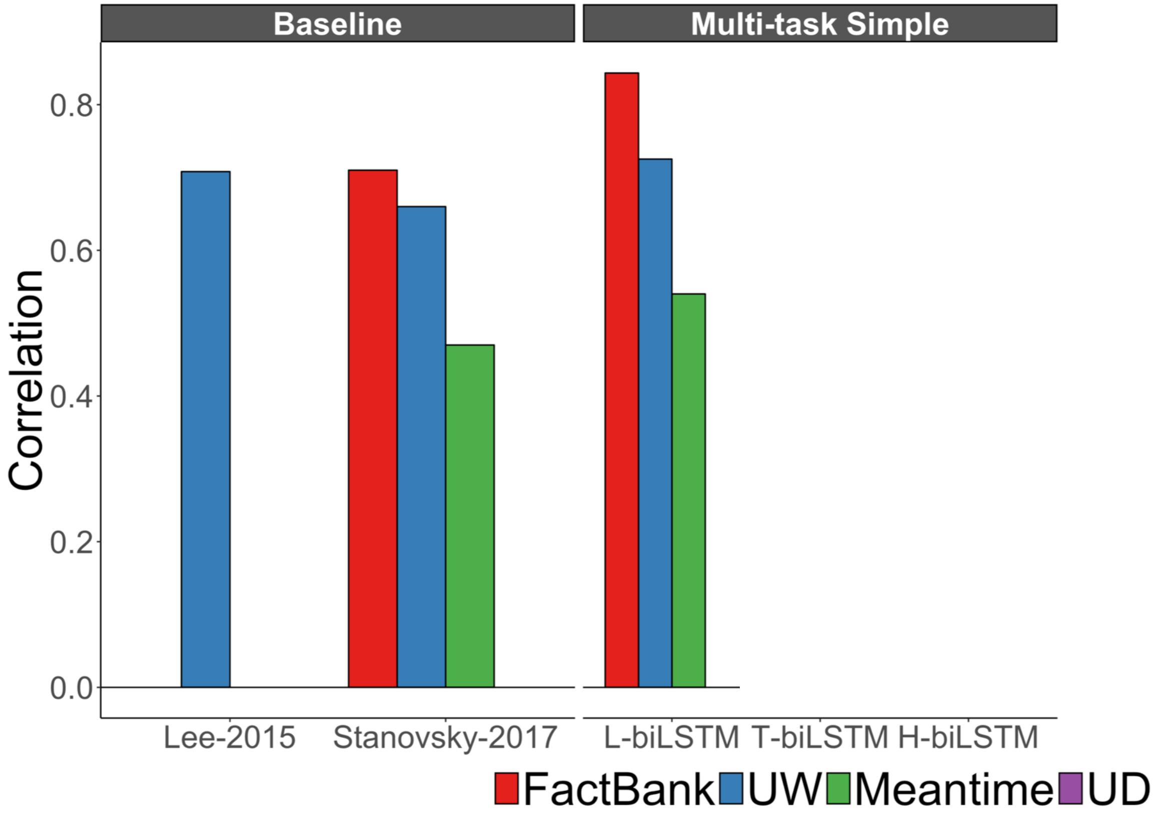
Higher entropy likely due to better genre distribution in UD

Single-task simple w/ features

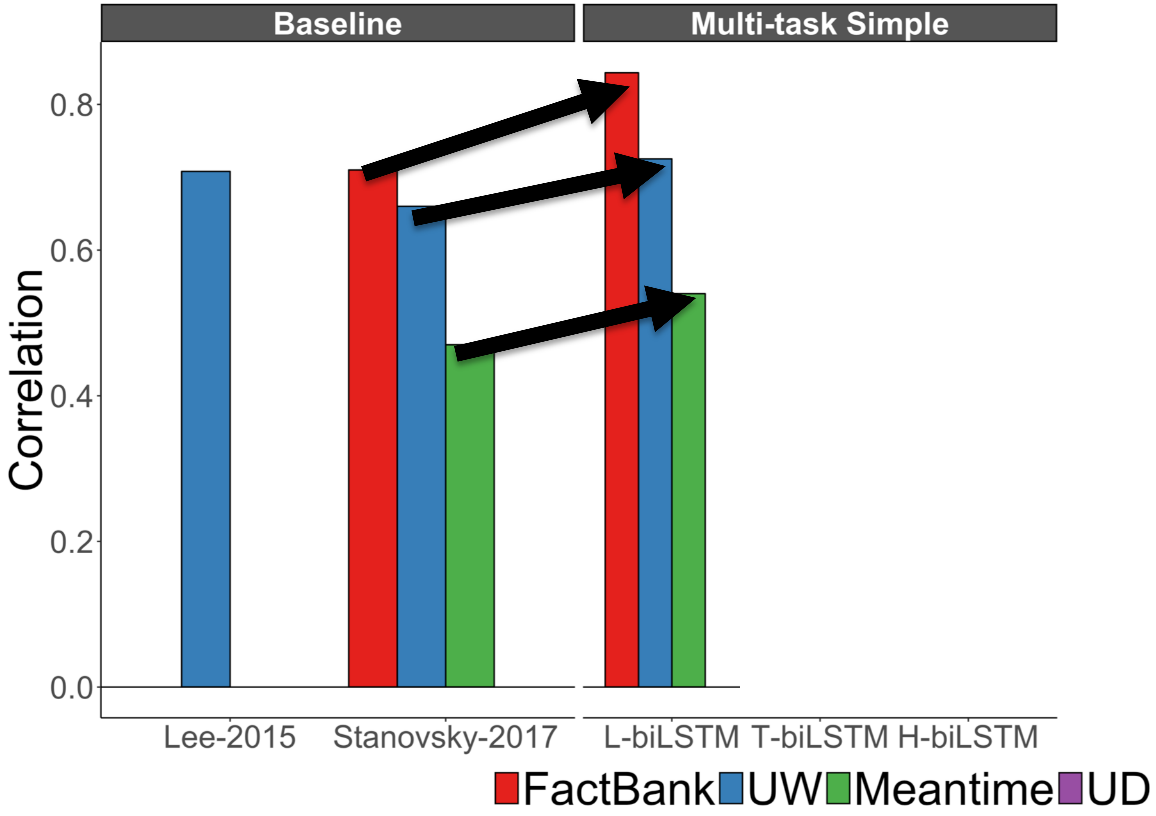
Single-task simple w/ features



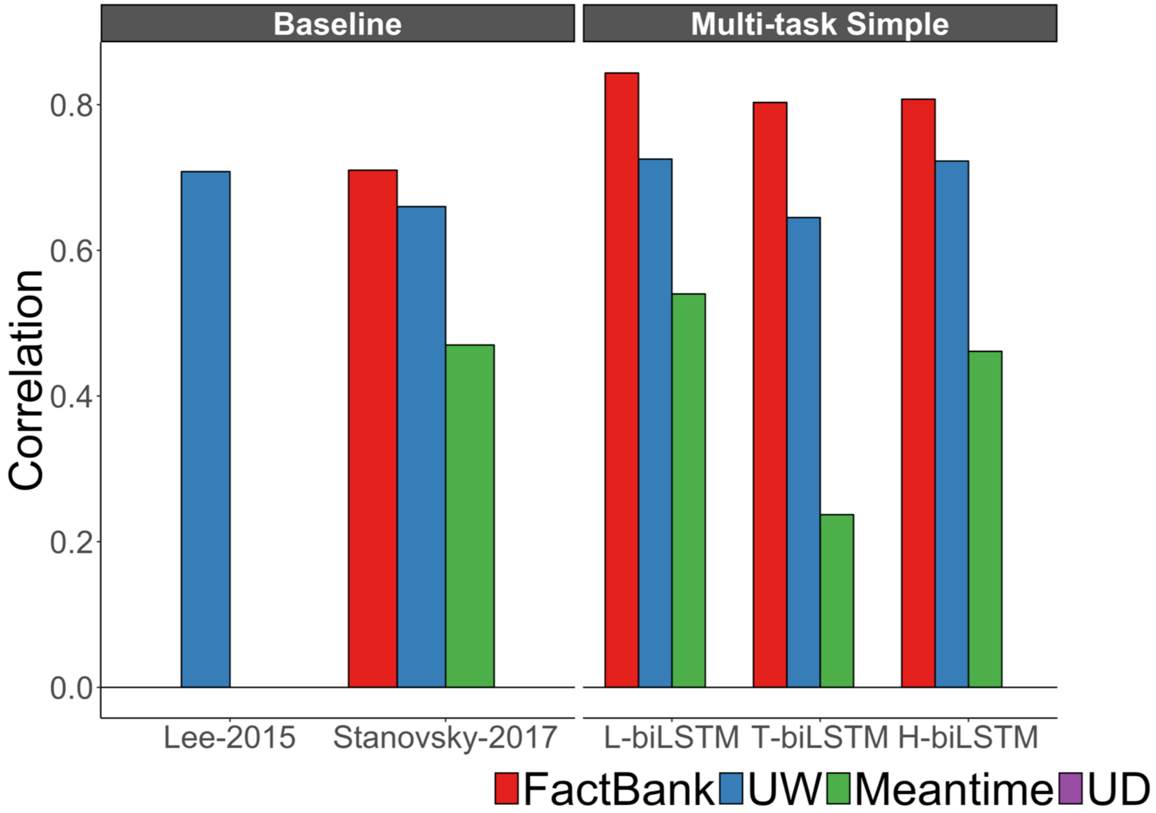
Single-task simple w/ features



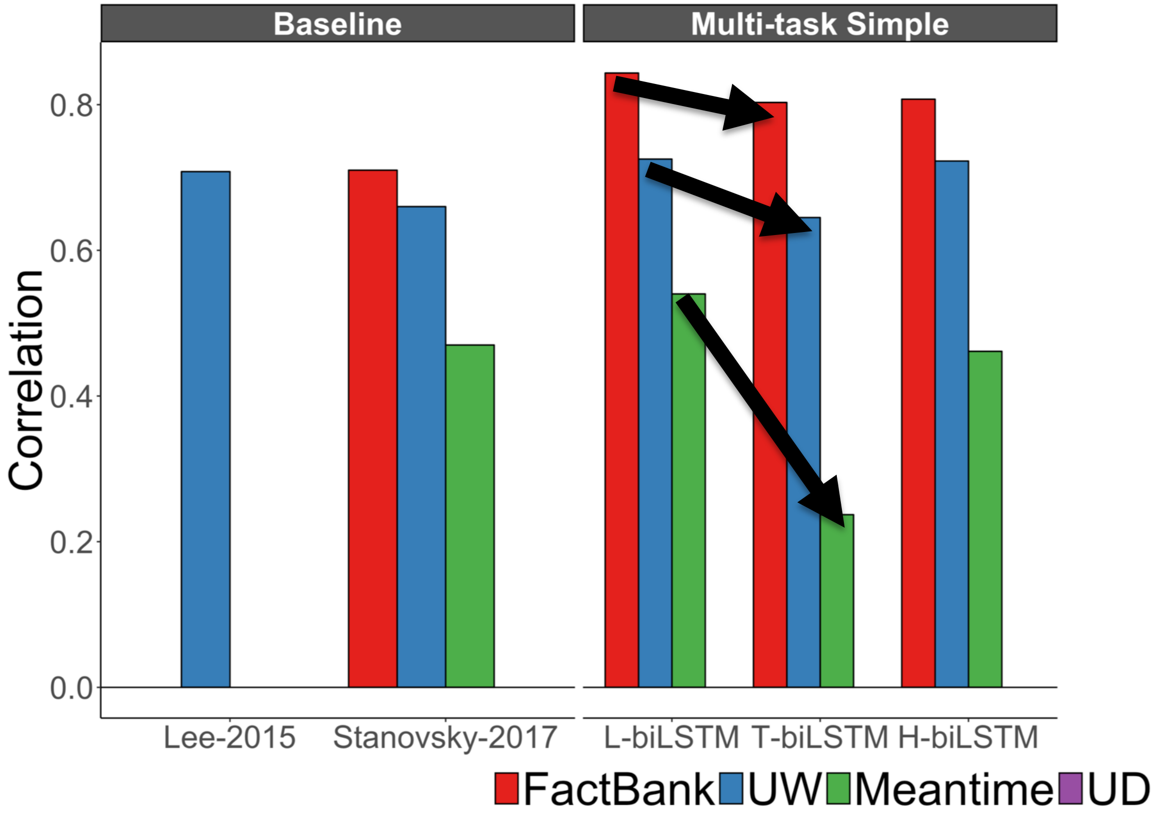
Single-task simple w/ features



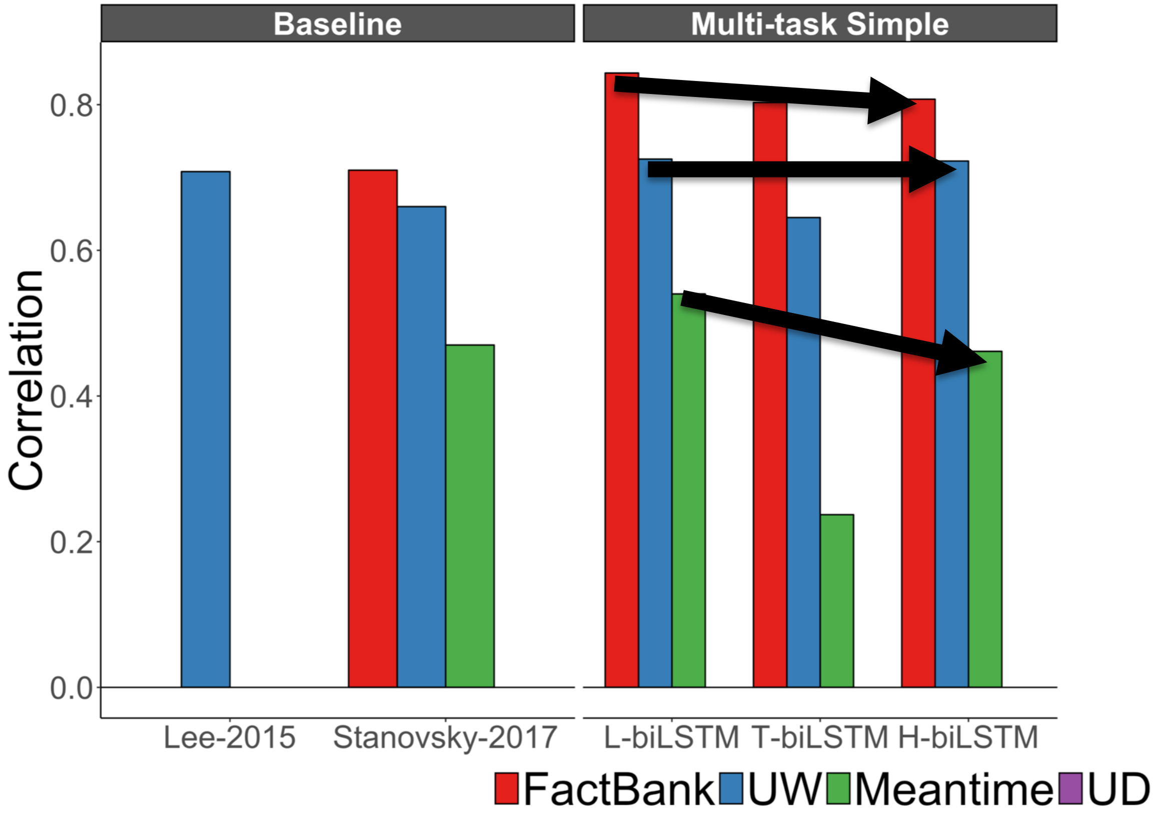
Single-task simple w/ features



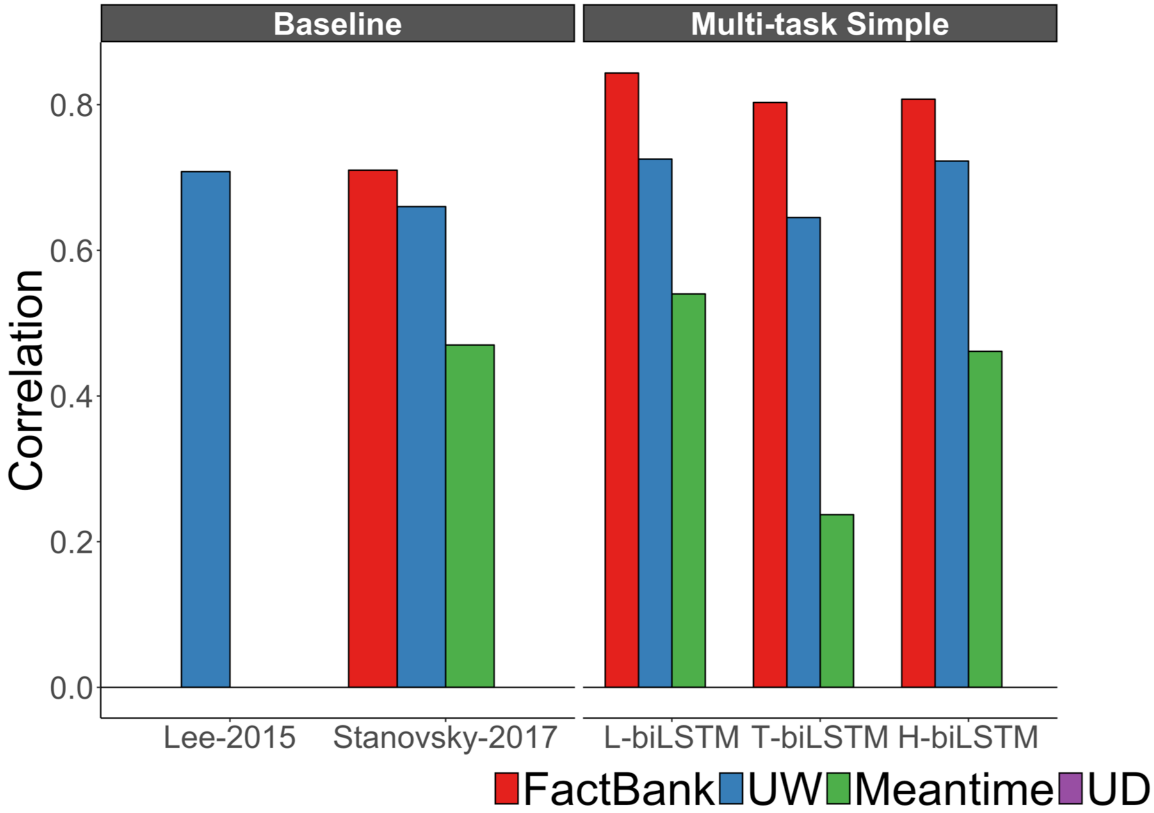
Single-task simple w/ features



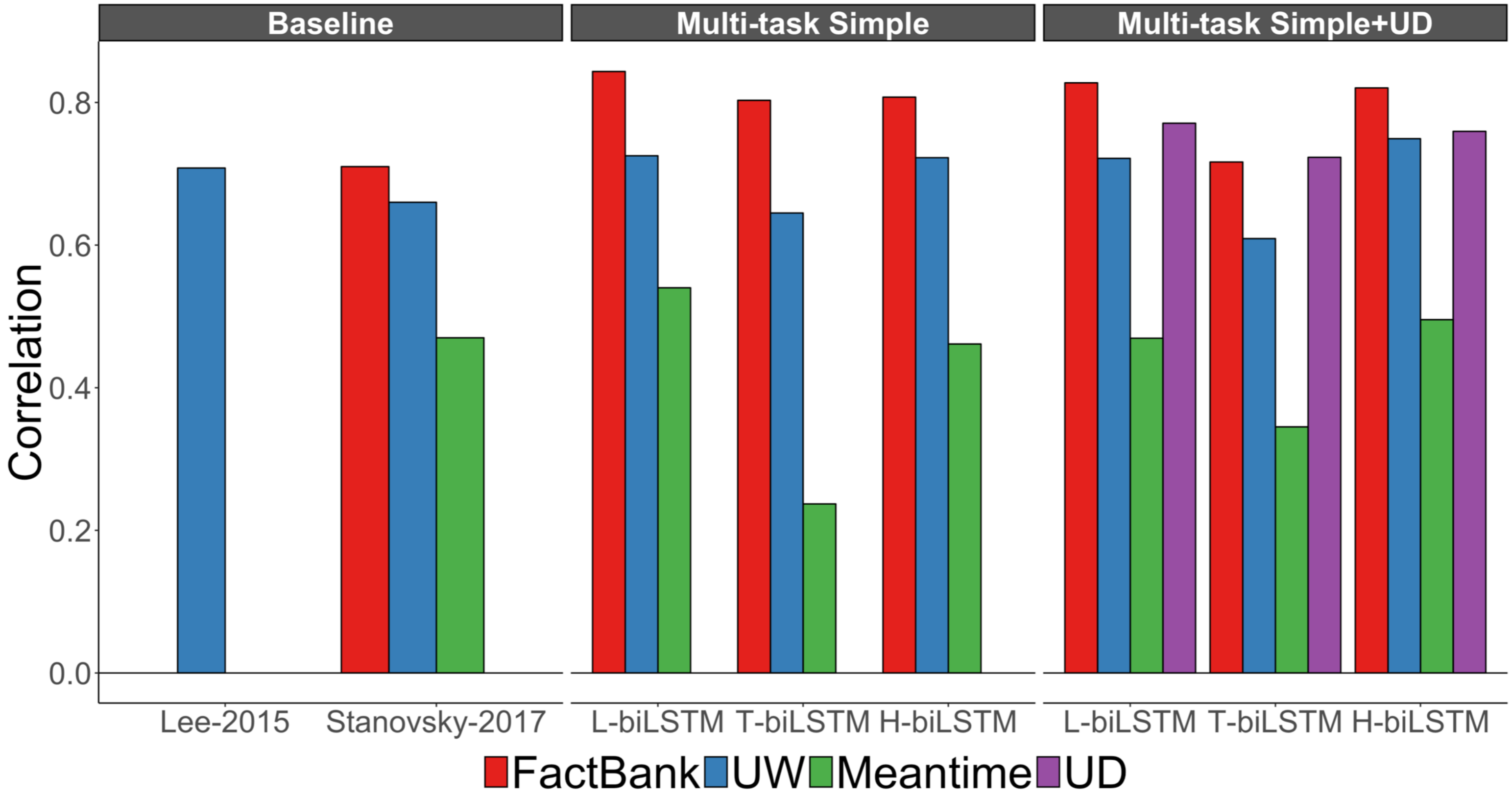
Single-task simple w/ features



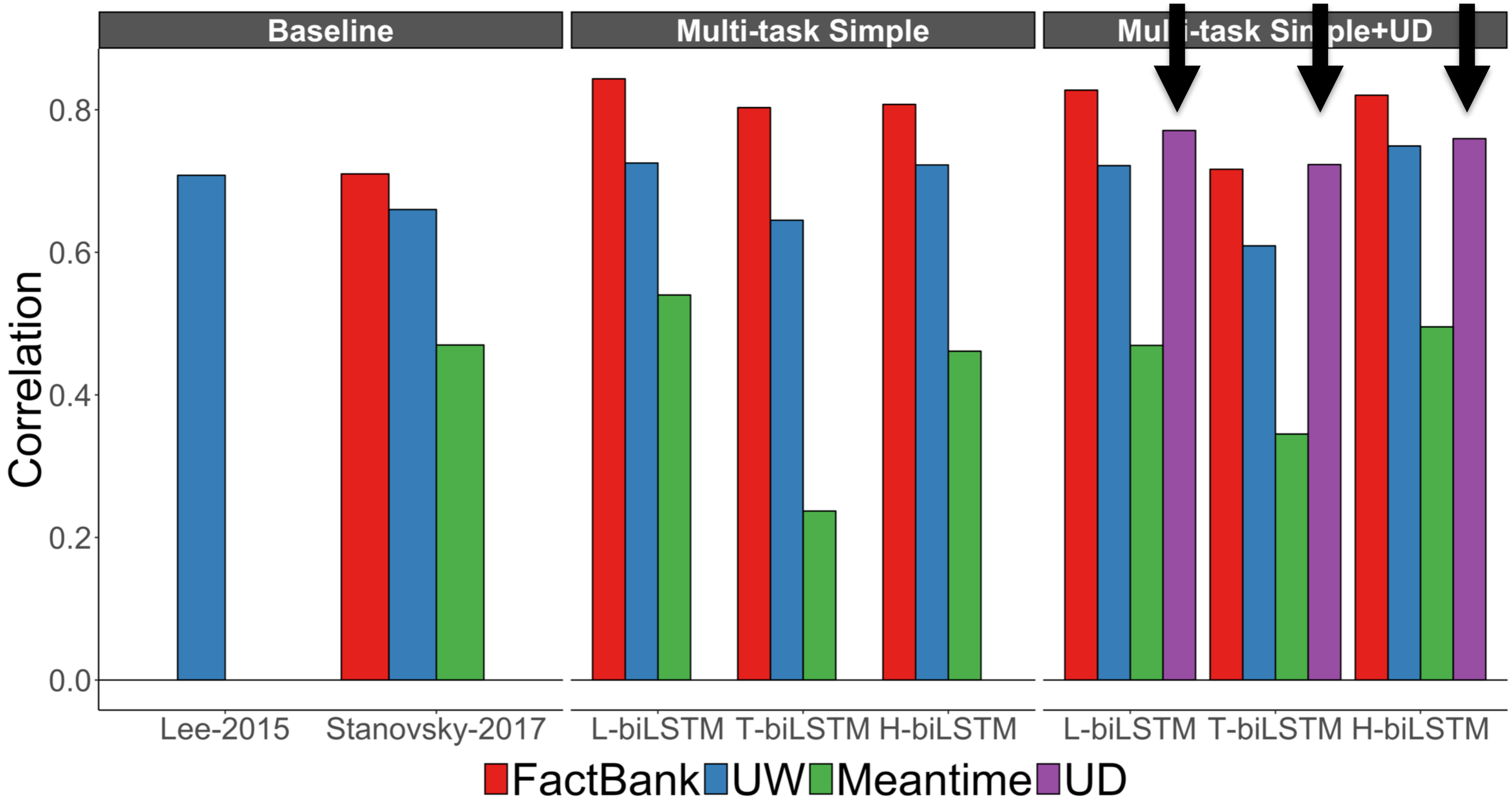
Single-task simple w/ features



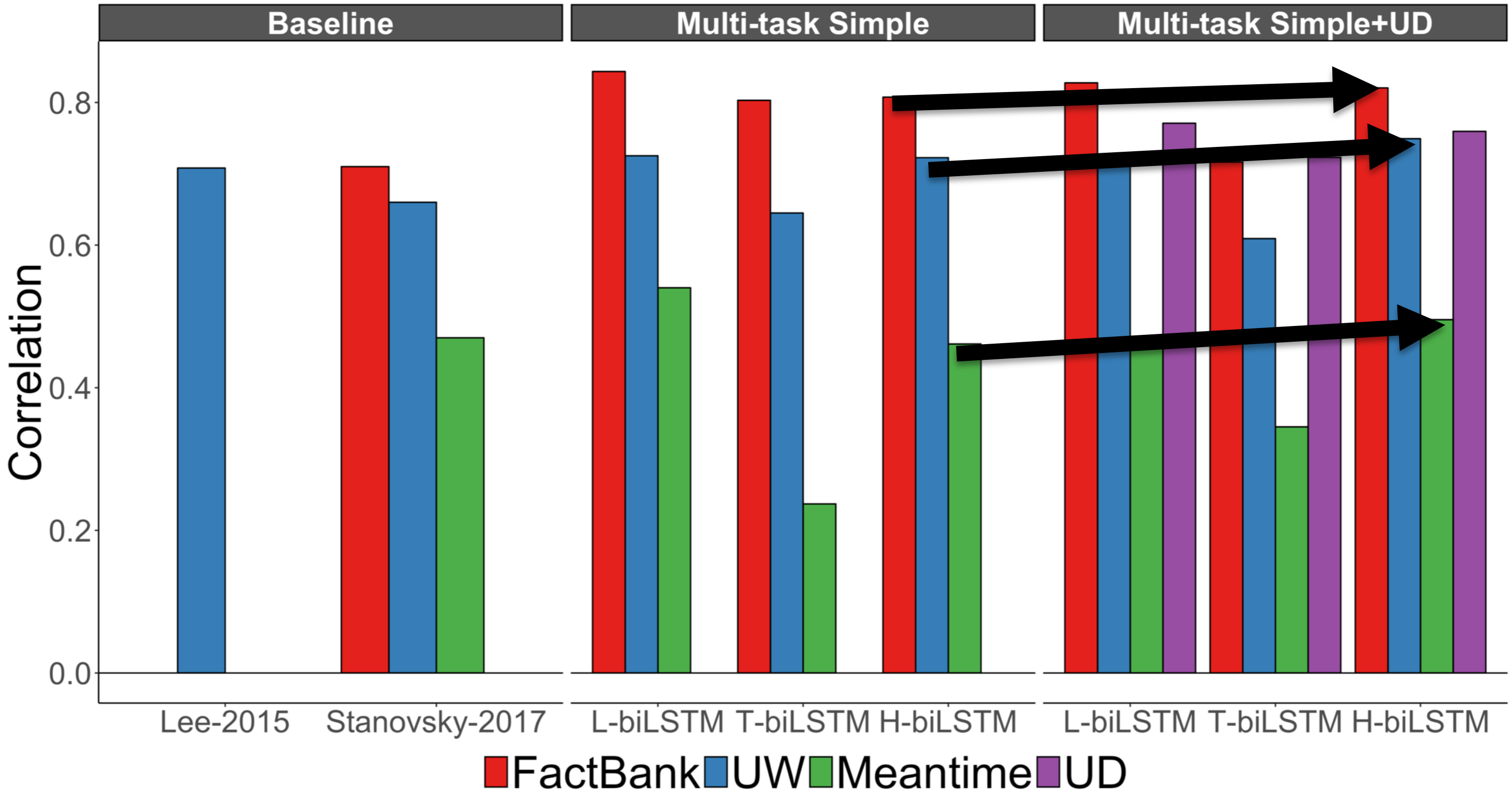
Single-task simple w/ features



Single-task simple w/ features



Single-task simple w/ features



Analysis

Analysis

- Conducted analyses on UD-It Happened
  - Predictability of factuality based on parent dependency of predicate

Error by parent dependency

Relation	Mean Label	L-biLSTM	T-biLSTM	#
root	1.07	1.03	0.96	949
conj	0.37	0.44	0.46	316
advcl	0.46	0.53	0.45	303
xcomp	-0.42	-0.57	-0.49	234
acl:relcl	1.28	1.40	1.31	193
ccomp	0.11	0.31	0.34	191
acl	0.77	0.59	0.58	159
parataxis	0.44	0.63	0.79	127
amod	1.92	1.88	1.81	76
csubj	0.36	0.38	0.27	37

Analysis

- Conducted analyses on UD-It Happened
  - Predictability of factuality based on parent dependency of predicate
  - Predictability of factuality based on modal or negation dependent

Error by presence of modal/neg

Modal	Negated	Mean Label	Linear MAE	Tree MAE	#
NONE	no	1.00	0.93	1.03	2244
NONE	yes	-0.19	1.40	1.69	98
may	no	-0.38	1.00	0.99	14
would	no	-0.61	0.85	0.99	39
ca(n't)	yes	-0.72	1.28	1.55	11
can	yes	-0.75	0.99	0.86	6
(wi)'ll	no	-0.94	1.47	1.14	8
could	no	-1.03	0.97	1.32	20
can	no	-1.25	1.02	1.21	73
might	no	-1.25	0.66	1.06	6
would	yes	-1.27	0.40	0.86	5
should	no	-1.31	1.20	1.01	22
will	no	-1.88	0.75	0.86	75

Analysis

- Conducted analyses on UD-It Happened
  - Predictability of factuality based on parent dependency of predicate
  - Predictability of factuality based on modal or negation dependent
  - Manual error analysis of 50 worst predicted

Manual error analysis

Attribute	#
Grammatical error present, incl. run-ons	16
Is an auxiliary or light verb	14
Annotation is incorrect	13
Future event	12
Is a question	5
Is an imperative	3
Is not an event or state	2
One or more of the above	43

Manual error analysis

Attribute	#
Grammatical error present, incl. run-ons	16
Is an auxiliary or light verb	14
Annotation is incorrect	13
Future event	12
Is a question	5
Is an imperative	3
Is not an event or state	2
One or more of the above	43

Manual error analysis

Attribute	#
Grammatical error present, incl. run-ons	16
Is an auxiliary or light verb	14
Annotation is incorrect	13
Future event	12
Is a question	5
[illegible]	[illegible]
One or more of the above	43

All labeled NOT HAPPENED

Manual error analysis

(We **check** in early afternoon and we fly next day.)

Manual error analysis

Before that , we are turned loose to **get** dinner .

Manual error analysis

Guerrillas threatened to **assassinate** Prime Minister Iyad Allawi and Minister of Defense Hazem Shaalan in retaliation for the attack .

Conclusion

Our contributions

- **New event factuality dataset** on Universal Dependencies-English Web TreeBank

Our contributions

- **New event factuality dataset** on Universal Dependencies-English Web TreeBank
- Evaluation of **simple, linguistically motivated neural models** for event factuality prediction, yielding SOTA

Thanks!

Research supported by the JHU HLTCOE, DARPA LORELEI + AIDA, and NSF-GRFP-1232825.

Rachel Rudinger

Aaron Steven White

Ben Van Durme