Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation

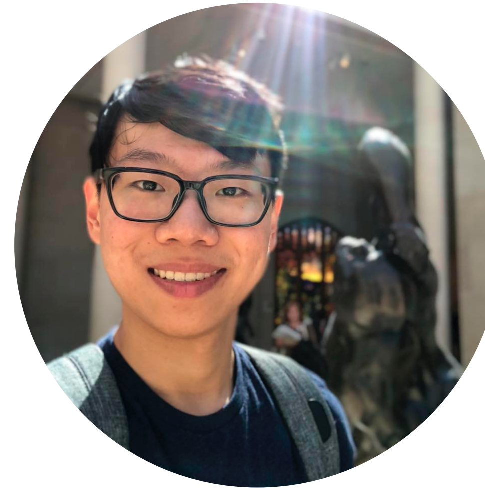
EMNLP 2018

Adam Poliak, Aparajita Haldar,
Rachel Rudinger, J. Edward Hu, Ellie Pavlick,
Aaron Steven White, Benjamin Van Durme

JOHNS HOPKINS UNIVERSITY

Collaborators






JOHNS HOPKINS
UNIVERSITY

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailed not-entailed

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

Multiple labeling schemas

entailed not-entailed

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

Multiple labeling schemas

entailed | not-entailed

entailment | neutral | contradiction

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailed   not-entailed

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailed

not-entailed

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailed   not-entailed

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailed   not-entailed

Why NLI as an NLP task?

Evaluation & Probing models

Historically

Historically

FraCas:

(Cooper et al., 1996)

Historically

FraCas: determine whether a model performs distinct types of reasoning

(Cooper et al., 1996)

Historically

FraCas: determine whether a model performs distinct types of reasoning

(Cooper et al., 1996)

Pascal RTE:

(Dagan et al., 2006)

JOHNS HOPKINS UNIVERSITY

Historically

FraCas: determine whether a model performs distinct types of reasoning

(Cooper et al., 1996)

Pascal RTE: “a generic evaluation framework” to compare models for distinct downstream tasks

(Dagan et al., 2006)

More recent

More recent

<u>SNLI & Multi-NLI:</u>

(Bowman et. al. 2015; Williams et. al. 2018)

More recent

<u>SNLI & Multi-NLI:</u> large scale datasets

(Bowman et. al. 2015; Williams et. al. 2018)

JOHNS HOPKINS UNIVERSITY

More recent

SNLI & Multi-NLI: large scale datasets

(Bowman et. al. 2015; Williams et. al. 2018)

Evaluate sentence representations

(Rep Eval 2017 Shared Task - Nangia et. al. 2017)

More recent

<u>*SNLI & Multi-NLI:*</u> large scale datasets

(Bowman et. al. 2015; Williams et. al. 2018)

Evaluate sentence representations

(Rep Eval 2017 Shared Task - Nangia et. al. 2017)

Training to improve models for downstream tasks

(Guo et. al. 2018)

JOHNS HOPKINS UNIVERSITY

Prior Dataset Characteristics

JOHNS HOPKINS
UNIVERSITY

Prior Dataset Characteristics

NLU Insights

Prior Dataset Characteristics

NLU Insights

Generation Methods

Prior Dataset Characteristics

NLU Insights

Generation Methods

Small Probing Sets

Characteristic 1: NLU Insights

Understanding our models' reasoning capabilities?

Characteristic 1: NLU Insights

Jianpeng Cheng et al. '16	450D LSTMN with deep attention fusion	3.4m	88.5	86.3
Parikh et al. '16	200D decomposable attention model	380k	89.5	86.3
Parikh et al. '16	200D decomposable attention model with intra-sentence attention	580k	90.5	86.8
Munkhdalai & Yu '16b	300D Full tree matching NTI-SLSTM-LSTM w/ global attention	3.2m	88.5	87.3
Zhiguo Wang et al. '17	BiMPM	1.6m	90.9	87.5
Lei Sha et al. '16	300D re-read LSTM	2.0m	90.7	87.5
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code)	4.4m	91.2	88.0
McCann et al. '17	Biattentive Classification Network + CoVe + Char	22m	88.5	88.1
Chuanqi Tan et al. '18	150D Multiway Attention Network	14m	94.5	88.3
Xiaodong Liu et al. '18	Stochastic Answer Network	3.5m	93.3	88.5
Ghaeini et al. '18	450D DR-BiLSTM	7.5m	94.1	88.5
Yi Tay et al. '18	300D CAFE	4.7m	89.8	88.5
Qian Chen et al. '17	KIM	4.3m	94.1	88.6
Qian Chen et al. '16	600D ESIM + 300D Syntactic TreeLSTM (code)	7.7m	93.5	88.6
Peters et al. '18	ESIM + ELMo	8.0m	91.6	88.7
Boyuan Pan et al. '18	300D DMAN	9.2m	95.4	88.8
Zhiguo Wang et al. '17	BiMPM **Ensemble**	6.4m	93.2	88.8
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code) **Ensemble**	17m	92.3	88.9
Seonhoon Kim et al. '18	Densely-Connected Recurrent and Co-Attentive Network	6.7m	93.1	88.9
Zhuosheng Zhang et al. '18	SLRC	6.1m	89.1	89.1
Qian Chen et al. '17	KIM **Ensemble**	43m	93.6	89.1
Ghaeini et al. '18	450D DR-BiLSTM **Ensemble**	45m	94.8	89.3
Peters et al. '18	ESIM + ELMo **Ensemble**	40m	92.1	89.3
Yi Tay et al. '18	300D CAFE **Ensemble**	17.5m	92.5	89.3
Chuanqi Tan et al. '18	150D Multiway Attention Network **Ensemble**	58m	95.5	89.4
Boyuan Pan et al. '18	300D DMAN **Ensemble**	79m	96.1	89.6
Radford et al. '18	Fine-Tuned LM-Pretrained Transformer	85m	96.6	**89.9**
Seonhoon Kim et al. '18	Densely-Connected Recurrent and Co-Attentive Network **Ensemble**	53.3m	95.0	**90.1**

Characteristic 2: Generation Methods

Characteristic 2: Generation Methods

Expensive

Characteristic 2: Generation Methods

Expensive

Leads to biases:

Characteristic 2: Generation Methods

Expensive

Leads to biases:

Stereotypical

(Rudinger et. al. 2017)

Characteristic 2: Generation Methods

Expensive

Leads to biases:

Stereotypical

(Rudinger et. al. 2017)

Class-based Statistical Irregularities

(Tsuchiya, 2018; Gururangan et al., 2018; Poliak et al., 2018)

Characteristic 3: Small Probing Sets

Characteristic 3: Small Probing Sets

FraCas is too small

Training neural network on 300 examples

Outline

- ~~Introduction~~
- The DNC: Diverse NLI Collection
- Constructing the DNC
- Experiments & Results

The DNC

The DNC

Diverse **N**atural Language Inference **C**ollection

The DNC

Diverse **N**atural Language Inference **C**ollection

Large scale collection of diverse NLI problems

The DNC

Diverse **N**atural Language Inference **C**ollection

Large scale collection of diverse NLI problems

Convert 7 semantic phenomena into NLI from 13 existing datasets

The DNC - Examples

Event Factuality	▶ Find him before he finds the dog food The finding did not happen	✓
	▶ I'll need to ponder The pondering happened	✗
Relation Extraction	▶ Ward joined Tom in their native Perth Ward was born in Perth	✓
	▶ Stefan had visited his son in Bulgaria Stefan was born in Bulgaria	✗
Puns	▶ Kim heard masks have no face value Kim heard a pun	✓
	▶ Tod heard that thrift is better than annuity Tod heard a pun	✗

The DNC

Sem. Phenomena/Annotations	Dataset	# pairs
Event Factuality	Decomp (Rudinger et al., 2018b)	42K (41,888)
	UW (Lee et al., 2015)	5K (5,094)
	MeanTime (Minard et al., 2016)	.7K (738)
Named Entity Recognition	Groningen (Bos et al., 2017)	260K (261,406)
	CoNLL (Tjong Kim Sang and De Meulder, 2003)	60K (59,970)
Gendered Anaphora	Winogender (Rudinger et al., 2018a)	.4K (464)
Lexicosyntactic Inference	VerbCorner (Hartshorne et al., 2013)	135K (138, 648)
	MegaVeridicality (White and Rawlins, 2018)	11K (11,814)
	VerbNet (Schuler, 2005)	2K (1, 950)
Puns	(Yang et al., 2015)	9K (9,492)
	SemEval 2017 Task 7 (Miller et al., 2017)	8K (8, 054)
Relationship Extraction	FACC1 (Gabrilovich et al., 2013)	30K (30,920)
Sentiment Analysis	(Kotzias et al., 2015)	6K (6,000)
Combined	Diverse NLI Collection (DNC)	575K (576,438)
—	SNLI (Bowman et al., 2015)	570K
—	Multi-NLI (Williams et al., 2017)	433K

JOHNS HOPKINS
UNIVERSITY

decompositional-semantics-initiative / **DNC**

Unwatch 4 | Unstar 9 | Fork 1

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Diverse Natural Language Inference Collection - NLI dataset that can used to evaluate how well models perform distinct types of reasoning (EMNLP 2018) http://decomp.io/projects/diverse-nat...

Edit

natural-language-processing | natural-language-inference | computational-semantics | emnlp2018 | Manage topics

6 commits | **1** branch | **1** release | **1** contributor

Branch: **master** | New pull request | Create new file | Upload files | Find file | Clone or download

azpoliak update README.md - inference is everything data ... Latest commit 6a8beee on Sep 14

dev	Released DNC and updated README	2 months ago
test	Released DNC and updated README	2 months ago
train	Released DNC and updated README	2 months ago
README.md	update README.md - inference is everything data	a month ago
additional_references.md	added bibs for original datasets	2 months ago
inference_is_everything.zip	included White et al's IJCNLP 2017 recast data	a month ago

README.md

DNC: *D*iverse *N*atural Language Inference *C*ollection

Dataset associated and released as part of *Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation* (EMNLP 2018).

decompositional-semantics-initiative / **DNC**

Unwatch 4 | Unstar 9 | Fork 1

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Diverse Natural Language Inference Collection - NLI dataset that can used to evaluate how well models perform distinct types of reasoning (EMNLP 2018) http://decomp.io/projects/diverse-nat...

Edit

natural-language-processing | natural-language-inference | computational-semantics | emnlp2018 | Manage topics

6 commits | 1 branch | 1 release | 1 contribut

Branch: master | New pull request | Create new file | Up files | Find | Clo r downlo

azpoliak update README.md - inference is everything data ... | st c a8beee on Sep 14

dev	Release READ	2 months ago
test	Rele d DNC and dat READM	2 months ago
train	Relea and updated README	2 months ago
EADME	pdate README.md - inference is everything data	a month ago
additi l_refe es.md	added bibs for original datasets	2 months ago
nference_is_everything.zip	included White et al's IJCNLP 2017 recast data	a month ago

README.md

DNC: *D*iverse *N*atural *L*anguage *I*nference *C*ollection

Dataset associated and released as part of *Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation* (EMNLP 2018).

decomp.io

The DNC

Sem. Phenomena/Annotations	Dataset	# pairs
Event Factuality	Decomp (Rudinger et al., 2018b) UW (Lee et al., 2015) MeanTime (Minard et al., 2016)	42K (41,888) 5K (5,094) .7K (738)
Named Entity Recognition	Groningen (Bos et al., 2017) CoNLL (Tjong Kim Sang and De Meulder, 2003)	260K (261,406) 60K (59,970)
Gendered Anaphora	Winogender (Rudinger et al., 2018a)	.4K (464)
Lexicosyntactic Inference	VerbCorner (Hartshorne et al., 2013) MegaVeridicality (White and Rawlins, 2018) VerbNet (Schuler, 2005)	135K (138, 648) 11K (11,814) 2K (1, 950)
Puns	(Yang et al., 2015) SemEval 2017 Task 7 (Miller et al., 2017)	9K (9,492) 8K (8, 054)
Relationship Extraction	FACC1 (Gabrilovich et al., 2013)	30K (30,920)
Sentiment Analysis	(Kotzias et al., 2015)	6K (6,000)
Combined	Diverse NLI Collection (DNC)	575K (576,438)
— —	SNLI (Bowman et al., 2015) Multi-NLI (Williams et al., 2017)	570K 433K

JOHNS HOPKINS
UNIVERSITY

Outline

- ~~Introduction~~
- ~~The DNC: Diverse NLI Collection~~
- Constructing the DNC
- Experiments & Results

The DNC

Sem. Phenomena/Annotations	Dataset	# pairs
Event Factuality	Decomp (Rudinger et al., 2018b) UW (Lee et al., 2015) MeanTime (Minard et al., 2016)	42K (41,888) 5K (5,094) .7K (738)
Named Entity Recognition	Groningen (Bos et al., 2017) CoNLL (Tjong Kim Sang and De Meulder, 2003)	260K (261,406) 60K (59,970)
Gendered Anaphora	Winogender (Rudinger et al., 2018a)	.4K (464)
Lexicosyntactic Inference	VerbCorner (Hartshorne et al., 2013) MegaVeridicality (White and Rawlins, 2018) VerbNet (Schuler, 2005)	135K (138, 648) 11K (11,814) 2K (1, 950)
Puns	(Yang et al., 2015) SemEval 2017 Task 7 (Miller et al., 2017)	9K (9,492) 8K (8, 054)
Relationship Extraction	FACC1 (Gabrilovich et al., 2013)	30K (30,920)
Sentiment Analysis	(Kotzias et al., 2015)	6K (6,000)
Combined	Diverse NLI Collection (DNC)	575K (576,438)
— —	SNLI (Bowman et al., 2015) Multi-NLI (Williams et al., 2017)	570K 433K

Recasting

Recasting

“*Leverage existing semantic annotations to create NLI datasets that probe different semantic phenomena*”

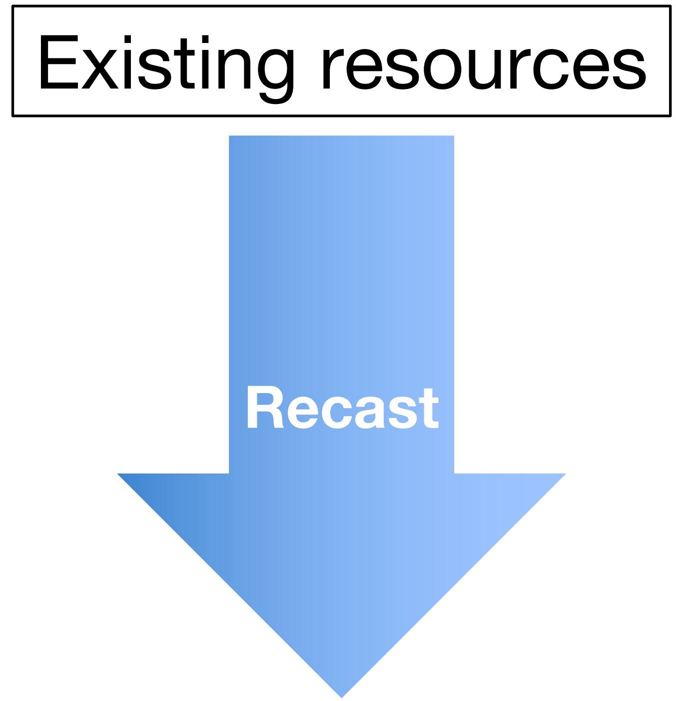
Recasting

"Leverage existing semantic annotations to create NLI datasets that probe different semantic phenomena"

Existing resources

Recasting

"Leverage existing semantic annotations to create NLI datasets that probe different semantic phenomena"



Recasting

"Leverage existing semantic annotations to create NLI datasets that probe different semantic phenomena"

Existing resources

↓ **Recast**

Focused Evaluation Datasets that probe different semantic phenomena

JOHNS HOPKINS UNIVERSITY

Event Factuality

Event Factuality

Create natural language template

Event Factuality

Create natural language template

Extract annotated preposition

Event Factuality

Create natural language template

Extract annotated preposition

Fill in template with preposition

Event Factuality

Create natural language template

Extract annotated preposition

Fill in template with preposition

Label example based on annotation

Event Factuality

I enjoyed studying here

Event Factuality

I enjoyed studying here

happened

Johns Hopkins University

Event Factuality

I enjoyed studying here

The studying happened

entailed　　　　not-entailed

Johns Hopkins University

Event Factuality

I enjoyed studying here

The studying did not happen

entailed　　　not-entailed

Event Factuality

I actually forgot to feed my chicken

JOHNS HOPKINS
UNIVERSITY

Event Factuality

I actually forgot to feed my chicken

did not happened

Event Factuality

I actually forgot to feed my chicken

The feeding happened

entailed

not-entailed

Event Factuality

I actually forgot to feed my chicken

The feeding did not happen

entailed not-entailed

Event Factuality

It Happened (White et. al. 2016; Rudinger et. al. 2018)

42K Examples

Event Factuality

It Happened (White et. al. 2016; Rudinger et. al. 2018)

42K Examples

UW (Lee et. al. 2015)

5K Examples

Event Factuality

It Happened (White et. al. 2016; Rudinger et. al. 2018)

42K Examples

UW (Lee et. al. 2015)

5K Examples

MeanTime (Minard et. al. 2016)

700 Examples

VerbNet Thematic Roles

No Comments

floss-41.2.1

Members: 4, Frames: 4

Members

BRUSH (FN 1; WN 3; G 1)

FLOSS (FN 1; WN 1)

SHAVE (FN 1; WN 2; G 1)

WASH (FN 1; WN 2, 3; G 1)

Roles

- Agent [+animate]
- Patient [+body_part]
- Instrument

Frames

NP V NP

EXAMPLE	"The hygienist flossed my teeth."
SYNTAX	Agent V Patient
SEMANTICS	take_care_of(during(E), Agent, Patient)

NP V

EXAMPLE	"I flossed."
SYNTAX	Agent V
SEMANTICS	take_care_of(during(E), Agent, ?Patient)

No Comments

floss-41.2.1

Members: 4, Frames: 4

Members

BRUSH (FN 1; WN 3; G 1)

FLOSS (FN 1; WN 1)

SHAVE (FN 1; WN 2; G 1)

WASH (FN 1; WN 2, 3; G 1)

1. Align tokens to Thematic Roles

Roles

- Agent [+ANIMATE]
- Patient [+BODY_PART]
- Instrument

Frames

NP V NP

EXAMPLE	"The hygienist flossed my teeth."
SYNTAX	Agent V Patient
SEMANTICS	TAKE_CARE_OF(DURING(E), Agent, Patient)

NP V

EXAMPLE	"I flossed."
SYNTAX	Agent V
SEMANTICS	TAKE_CARE_OF(DURING(E), Agent, ?Patient)

No Comments

floss-41.2.1

Members: 4, Frames: 4

Members

BRUSH (FN 1; WN 3; G 1)

FLOSS (FN 1; WN 1)

SHAVE (FN 1; WN 2; G 1)

WASH (FN 1; WN 2, 3; G 1)

1. Align tokens to Thematic Roles

Roles

- AGENT [+ANIMATE]
- PATIENT [+BODY_PART]
- INSTRUMENT

Frames

NP V NP

EXAMPLE	"The hygienist flossed my teeth."
SYNTAX	AGENT V PATIENT
SEMANTICS	TAKE_CARE_OF(DURING(E), AGENT, PATIENT)

hygienist ⟷ Agent

teeth ⟷ Patient

NP V

EXAMPLE	"I flossed."
SYNTAX	AGENT V
SEMANTICS	TAKE_CARE_OF(DURING(E), AGENT, ?PATIENT)

No Comments

floss-41.2.1

Members: 4, Frames: 4

Members

BRUSH (FN 1; WN 3; G 1)
FLOSS (FN 1; WN 1)
SHAVE (FN 1; WN 2; G 1)
WASH (FN 1; WN 2, 3; G 1)

1. Align tokens to Thematic Roles

Roles

- Agent [+animate]
- Patient [+body_part]
- Instrument

2. Convert semantics into natural language templates

Frames

NP V NP

EXAMPLE "The hygienist flossed my teeth."

SYNTAX Agent V Patient

SEMANTICS take_care_of(during(E), Agent, Patient)

NP V

EXAMPLE "I flossed."

SYNTAX Agent V

SEMANTICS take_care_of(during(E), Agent, ?Patient)

No Comments

floss-41.2.1

Members: 4, Frames: 4

Members

BRUSH (FN 1; WN 3; G 1)

FLOSS (FN 1; WN 1)

SHAVE (FN 1; WN 2; G 1)

WASH (FN 1; WN 2, 3; G 1)

1. Align tokens to Thematic Roles

Roles

- Agent [+animate]
- Patient [+body_part]
- Instrument

2. Convert semantics into natural language templates

Frames

NP V NP

EXAMPLE "The hygienist flossed my teeth."

Agent took care of Patient

SYNTAX Agent V Patient

SEMANTICS $\boxed{\text{take\_care\_of(during(E), Agent, Patient)}}$

NP V

EXAMPLE "I flossed."

SYNTAX Agent V

SEMANTICS take_care_of(during(E), Agent, ?Patient)

No Comments

floss-41.2.1

Members: 4, Frames: 4

Members

BRUSH (FN 1; WN 3; G 1)

FLOSS (FN 1; WN 1)

SHAVE (FN 1; WN 2; G 1)

WASH (FN 1; WN 2, 3; G 1)

1. Align tokens to Thematic Roles

Roles

- Agent [+ANIMATE]
- Patient [+BODY_PART]
- Instrument

2. Convert semantics into natural language templates

Frames

NP V NP

EXAMPLE "The hygienist flossed my teeth."

SYNTAX Agent V Patient

SEMANTICS TAKE_CARE_OF(DURING(E), Agent, Patient)

NP V

3. Fill in natural language templates

EXAMPLE "I flossed."

SYNTAX Agent V

SEMANTICS TAKE_CARE_OF(DURING(E), Agent, ?Patient)

VerbNet Thematic Roles

The hygienist flossed my teeth

<u>Agent</u> took care of <u>Patient</u>

entailed   not-entailed

VerbNet Thematic Roles

The hygienist flossed my teeth

<u>The hygienist</u> took care of <u>my teeth</u>

entailed not-entailed

VerbNet Thematic Roles

The hygienist flossed my teeth

Patient took care of Agent

entailed not-entailed

VerbNet Thematic Roles

The hygienist flossed my teeth

<u>My teeth</u> took care of <u>the hygienist</u>

entailed not-entailed

Outline

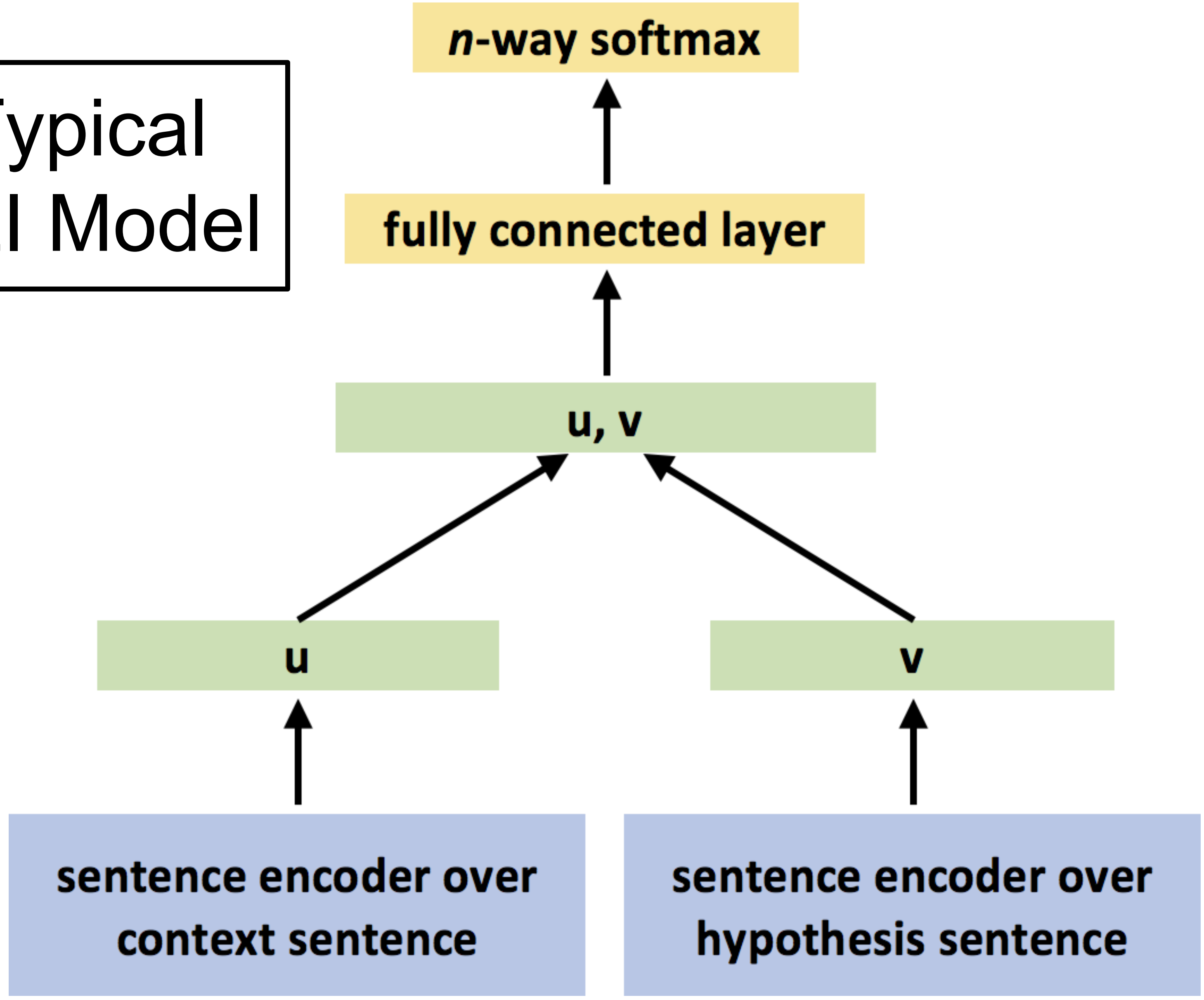
- ~~Introduction~~
- ~~The DNC: Diverse NLI Collection~~
- ~~Constructing the DNC~~
- Experiments & Results

Experimental Goal

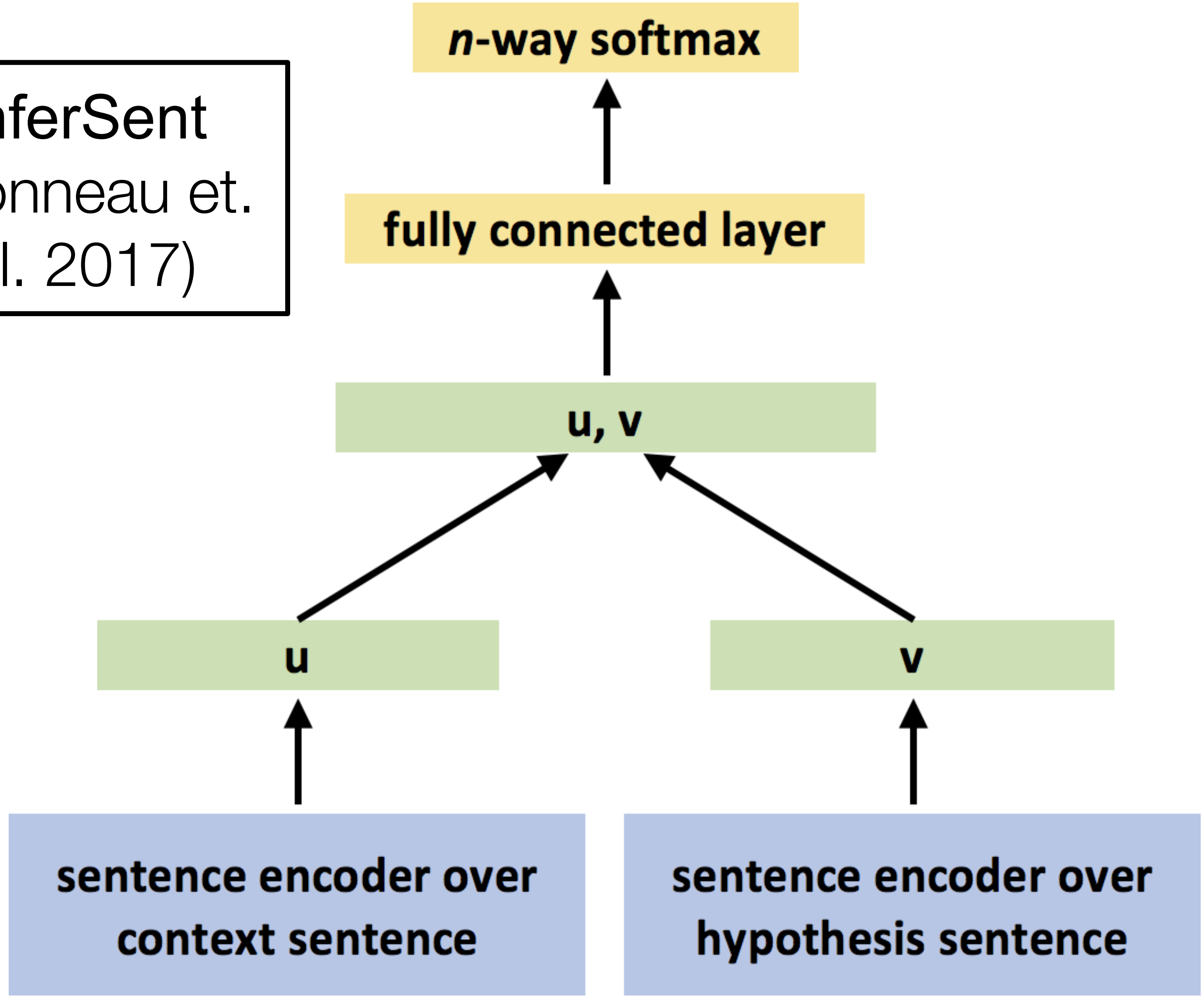
Experimental Goal

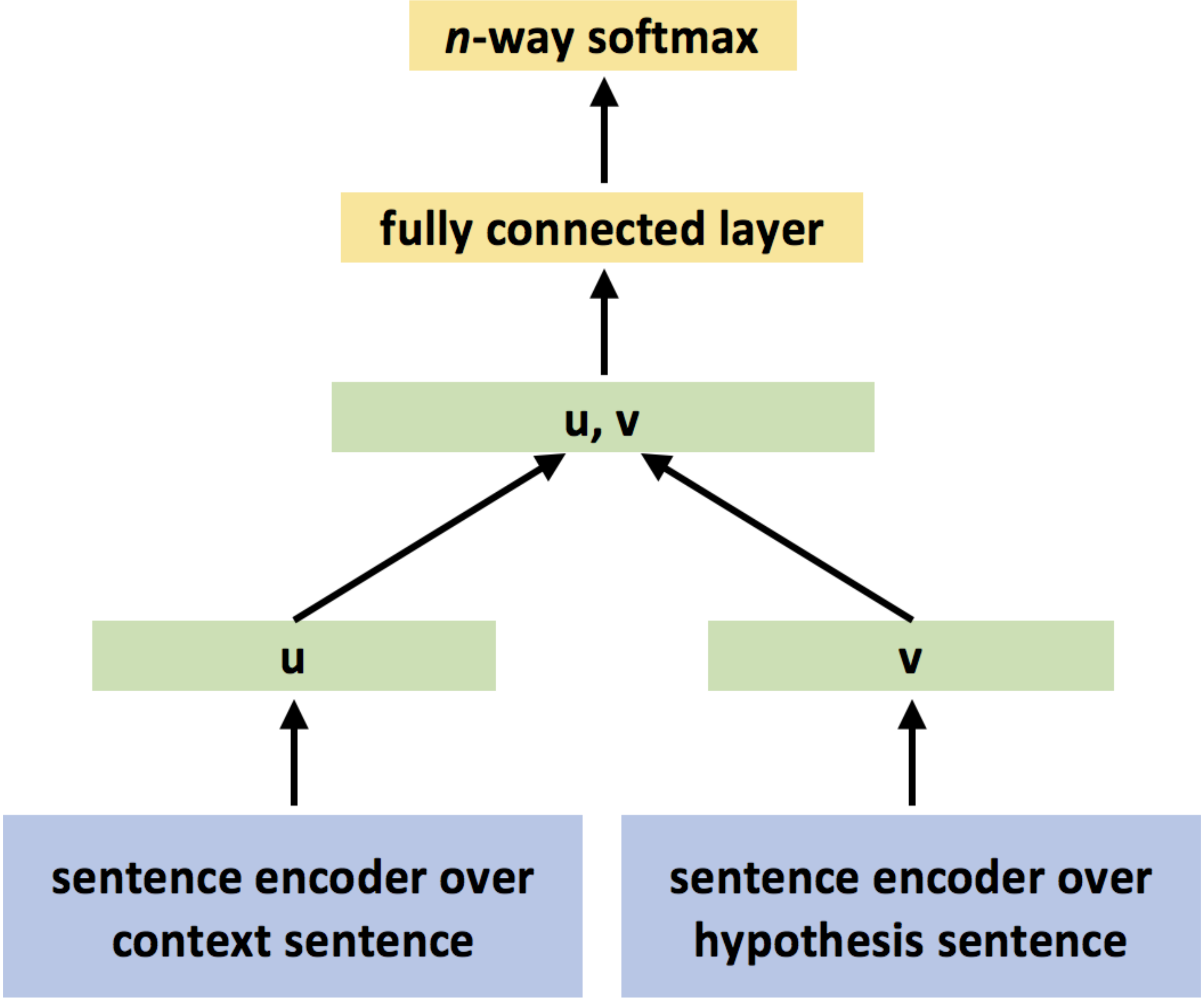
"demonstrate how the DNC can help to evaluate how well models capture different types of semantic reasoning necessary for general language understanding"

Typical NLI Model



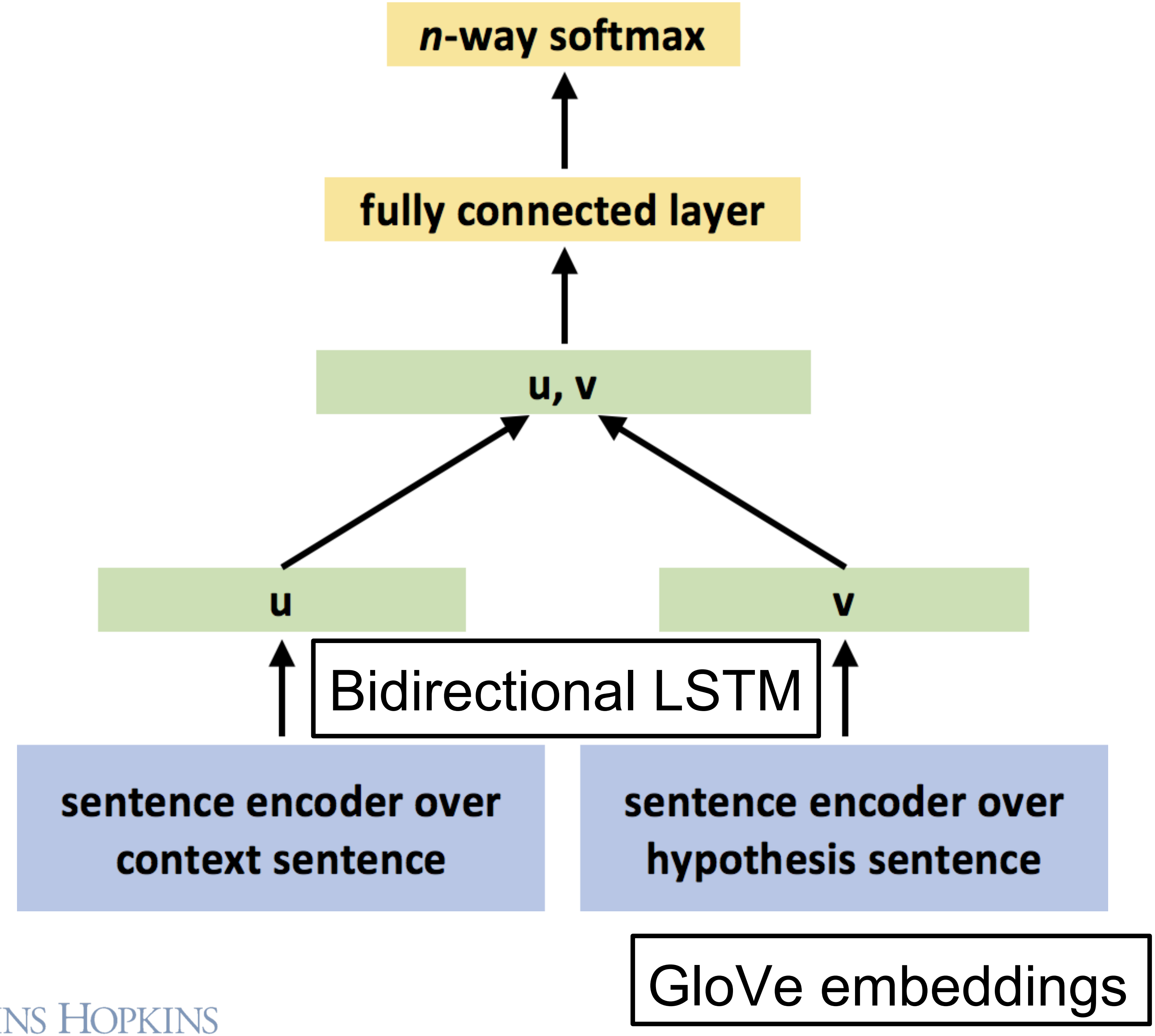
InferSent (Conneau et. al. 2017)



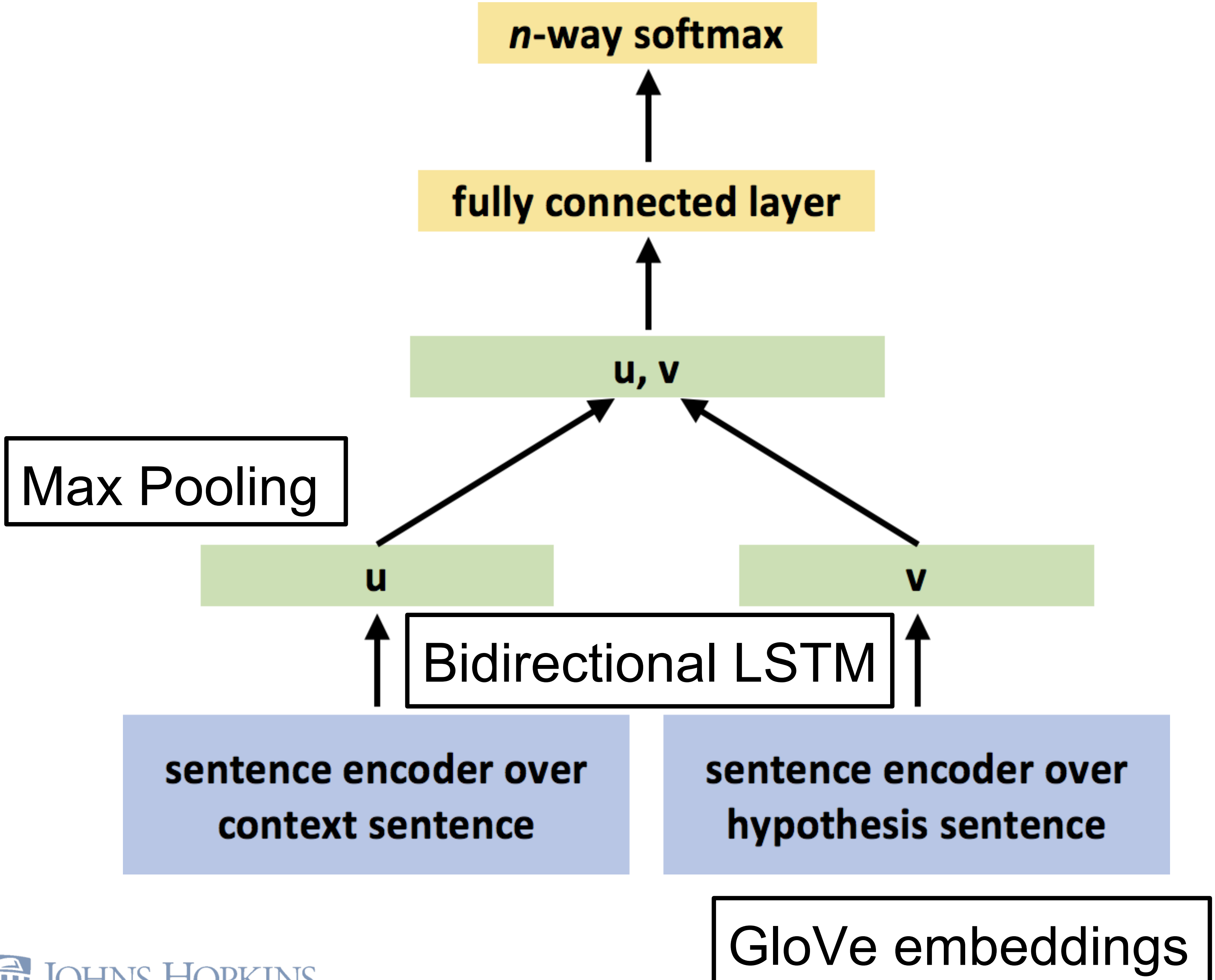


GloVe embeddings

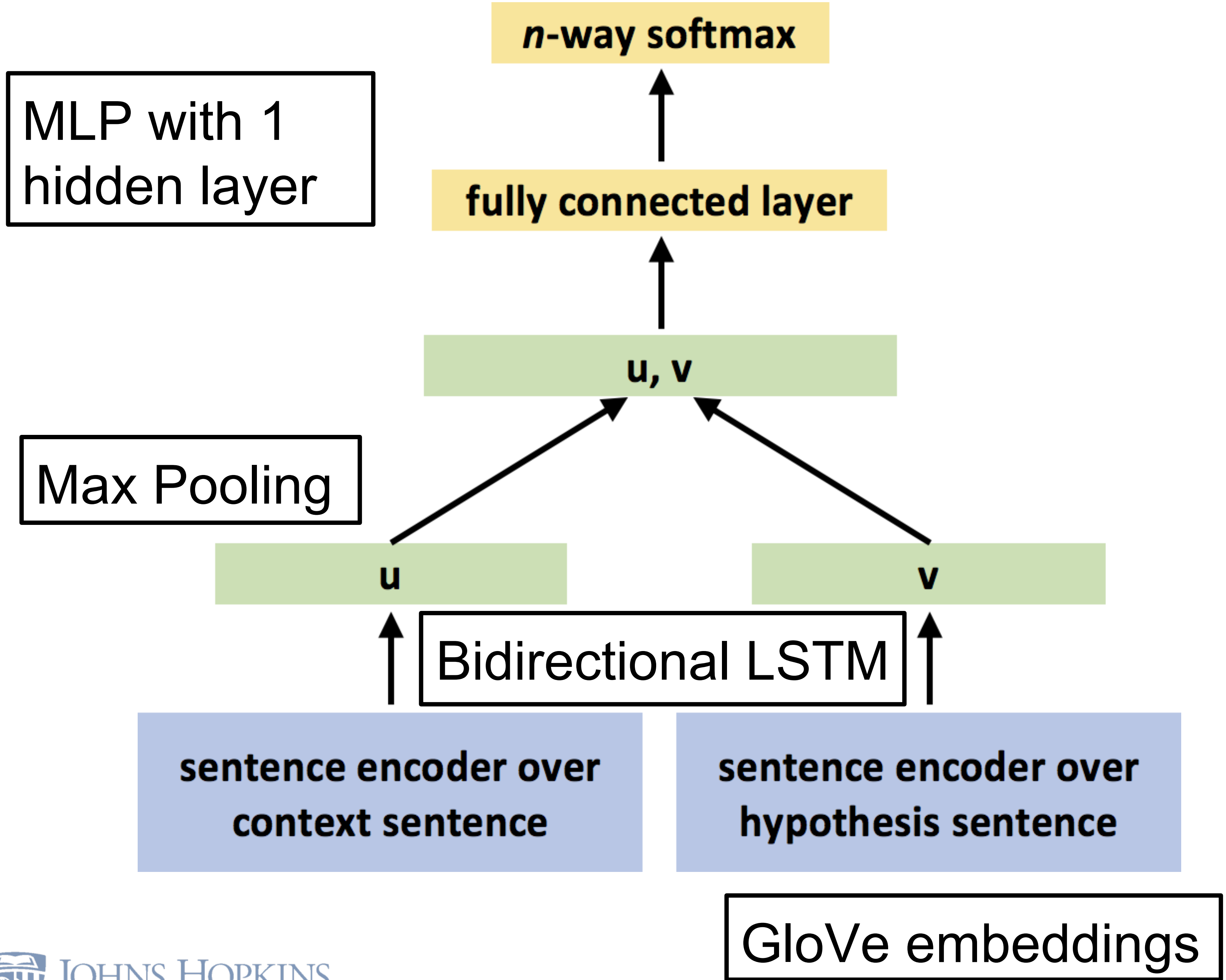
JOHNS HOPKINS UNIVERSITY

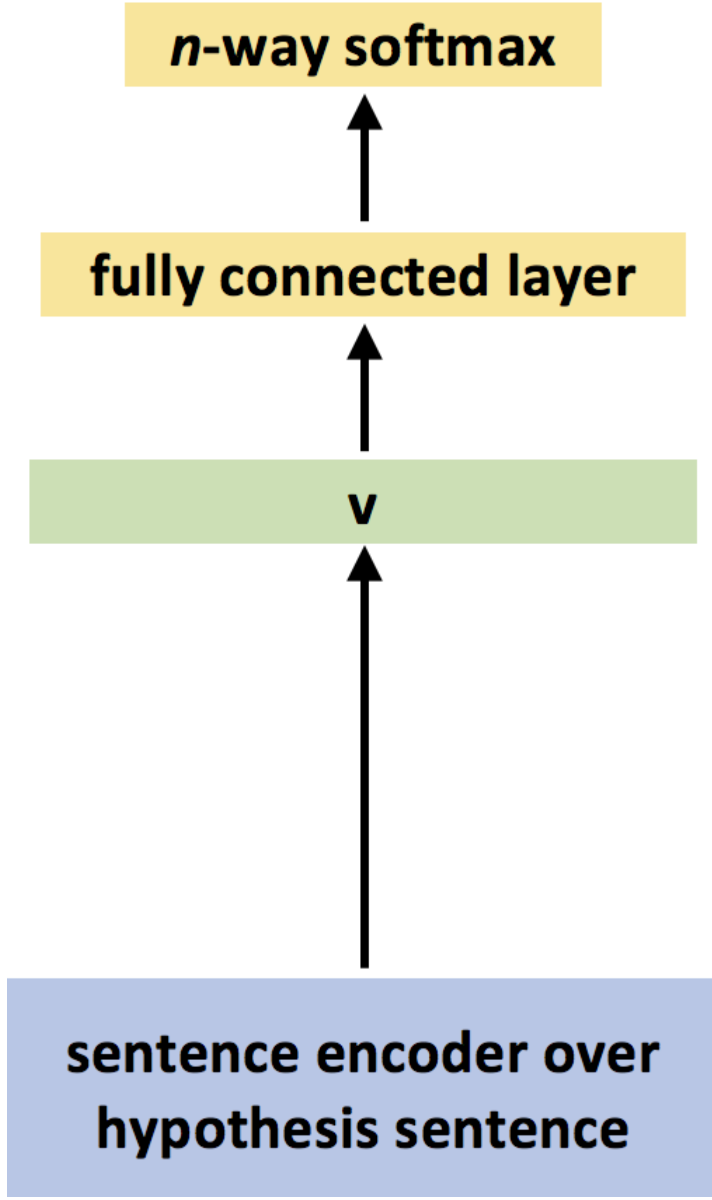


JOHNS HOPKINS UNIVERSITY



Johns Hopkins University


n-way softmax
MLP with 1 hidden layer
fully connected layer
u, v
Max Pooling
u
v
Bidirectional LSTM
sentence encoder over context sentence
sentence encoder over hypothesis sentence
GloVe embeddings



Hypothesis Only baseline (Poliak et. al. *SEM 2018)

Johns Hopkins University

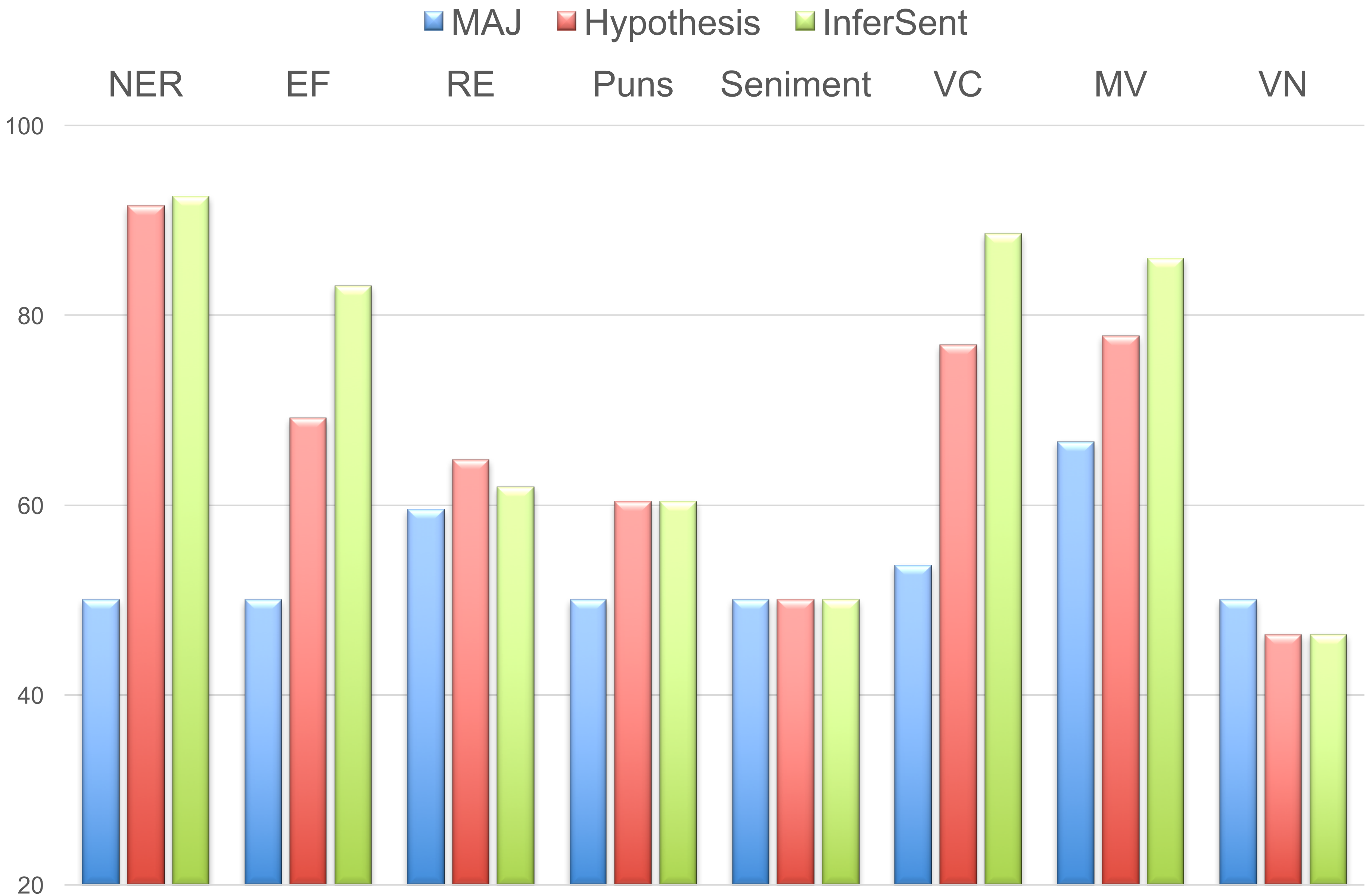
Results

Model \ Recast Data	NER	EF	RE	Puns	Sentiment	GAR	VC	MV	VN
Majority (MAJ)	50.00	50.00	59.53	50.00	50.00	50.00	50.00	66.67	53.66
No Pre-training									
InferSent	**92.50**	83.07	61.89	60.36	50.00	–	88.60	**85.96**	46.34
Hyp-only	91.48	69.14	64.78	60.36	50.00	–	76.82	77.83	46.34
Pre-trained DNC									
InferSent *(update)*	92.47	**83.86**	74.38	**93.17**	81.00	–	**89.00**	85.62	76.83
InferSent *(fixed)*	92.20	81.07	74.11	87.76	77.33	**50.65**	88.59	83.84	67.68
Hyp-only *(update)*	91.60	71.07	70.57	60.02	46.83	–	76.78	77.83	68.90
Hyp-only *(fixed)*	91.37	69.74	65.97	56.44	48.17	50.00	76.78	77.83	59.15
Pre-trained Multi-NLI									
InferSent *(update)*	92.37	83.03	**76.08**	92.48	**83.50**	–	88.45	85.11	**78.05**
InferSent *(fixed)*	52.99	54.88	66.75	56.04	56.50	**50.65**	45.33	55.92	45.73
Hyp-only *(update)*	91.62	70.64	69.91	60.36	49.33	–	76.82	77.83	68.29
Hyp-only *(fixed)*	52.55	66.33	52.96	60.59	50.00	50.43	41.31	46.28	48.78

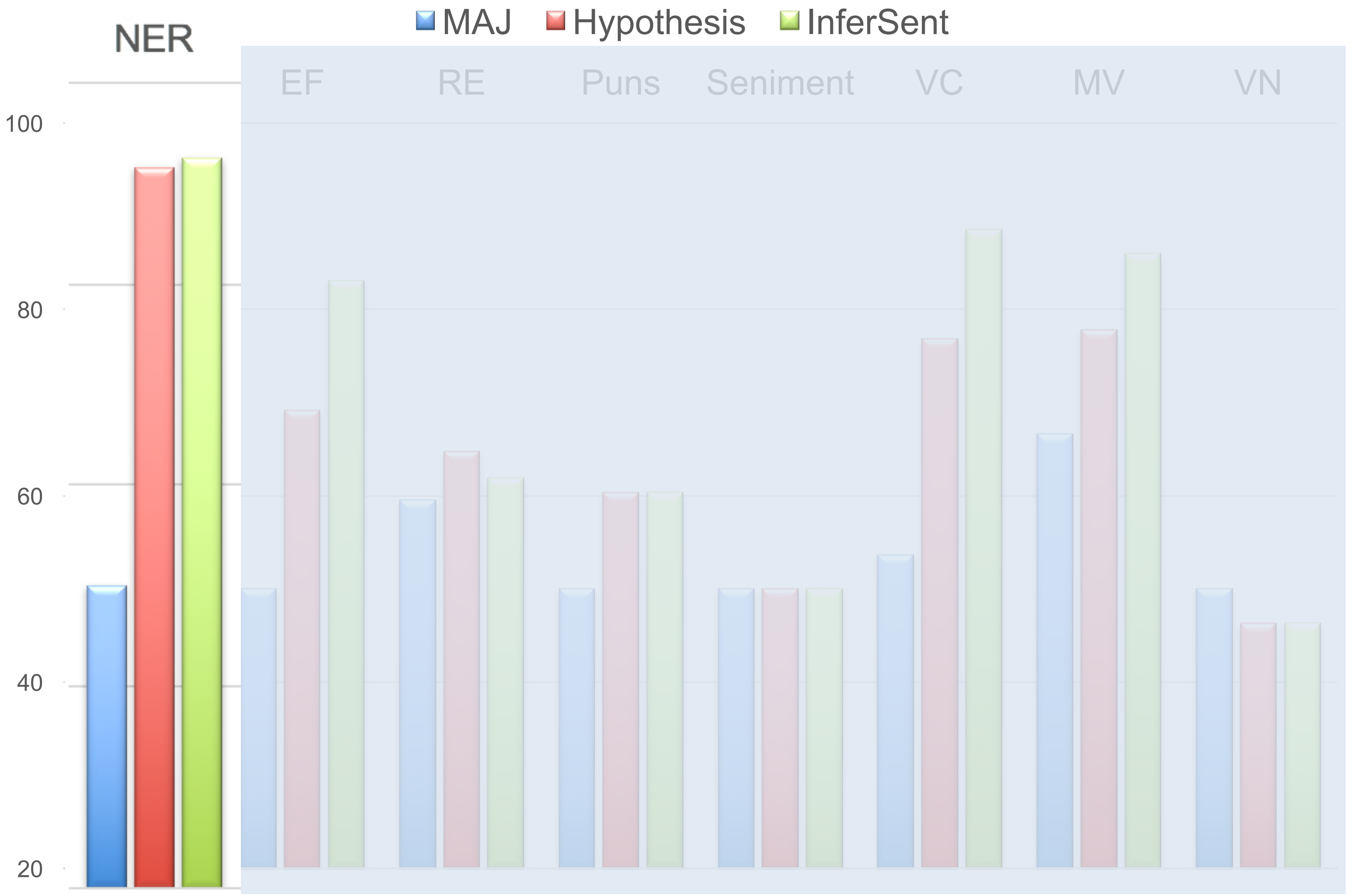
Experimental Setup

Experimental Setup

Train models on each DNC dataset



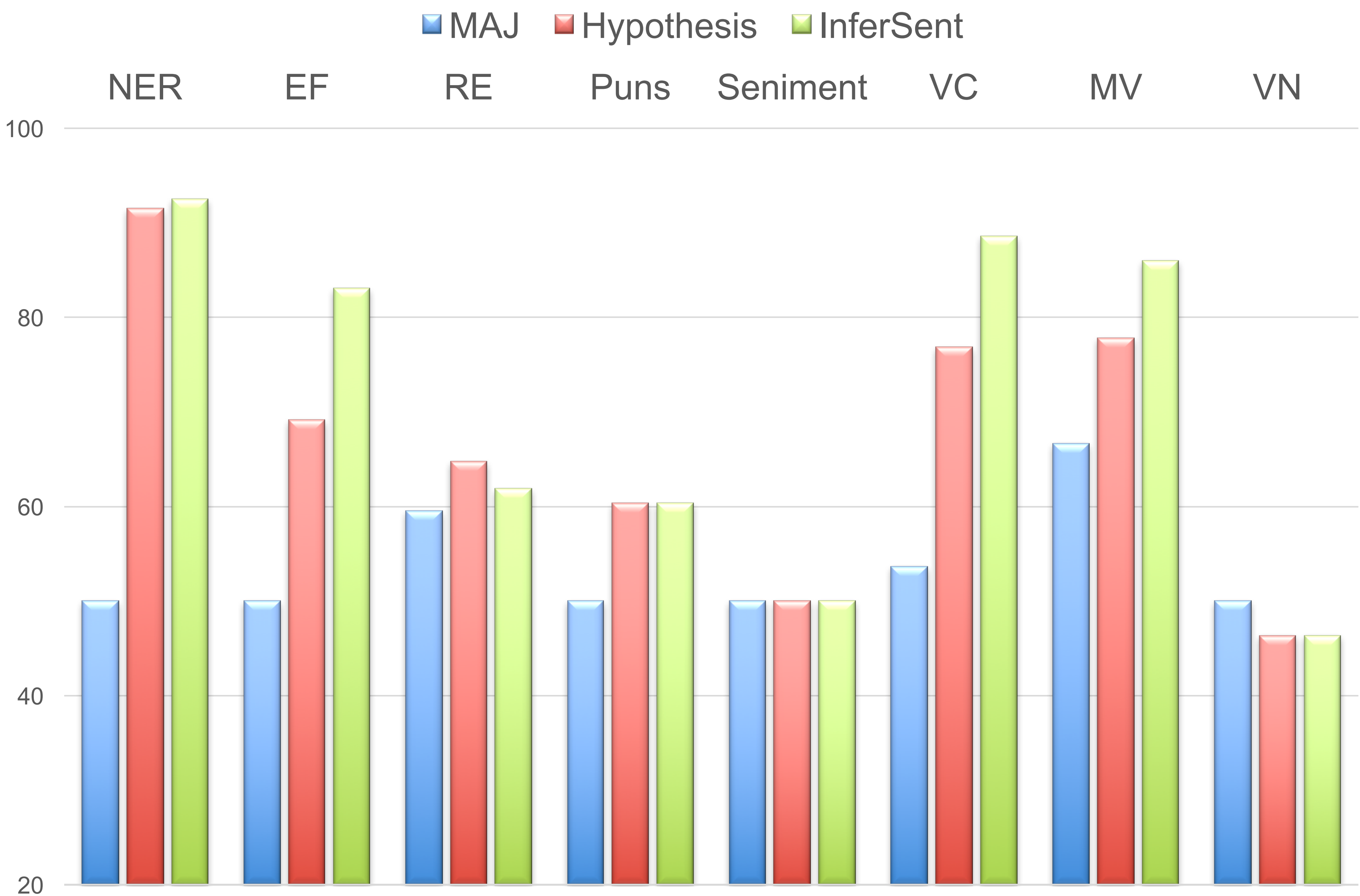
MAJ
Hypothesis
InferSent
NER
EF
RE
Puns
Seniment
VC
MV
VN
100
80
60
40
20

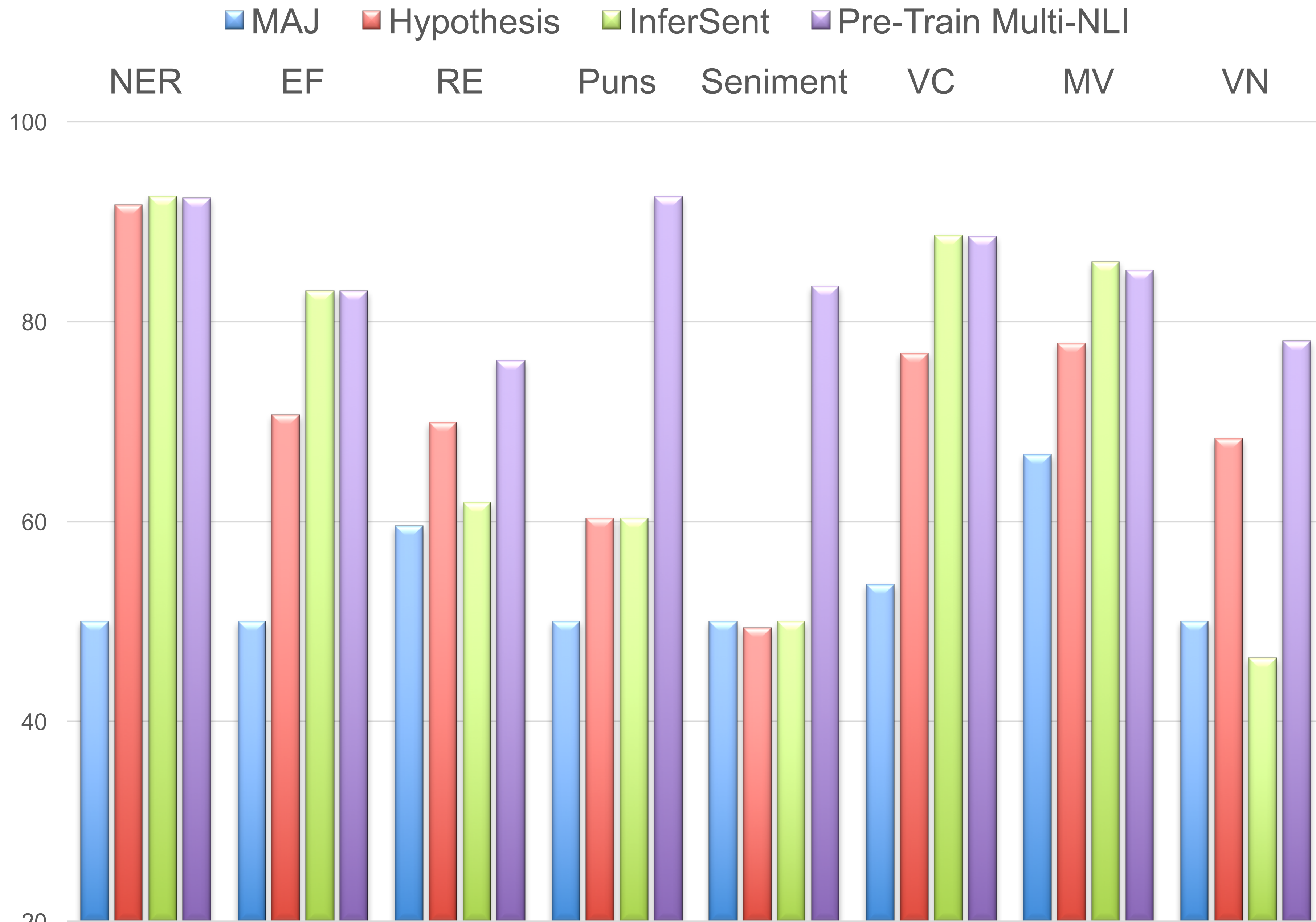

NER
MAJ
Hypothesis
InferSent
EF
RE
Puns
Seniment
VC
MV
VN
100
80
60
40
20

Experimental Setup

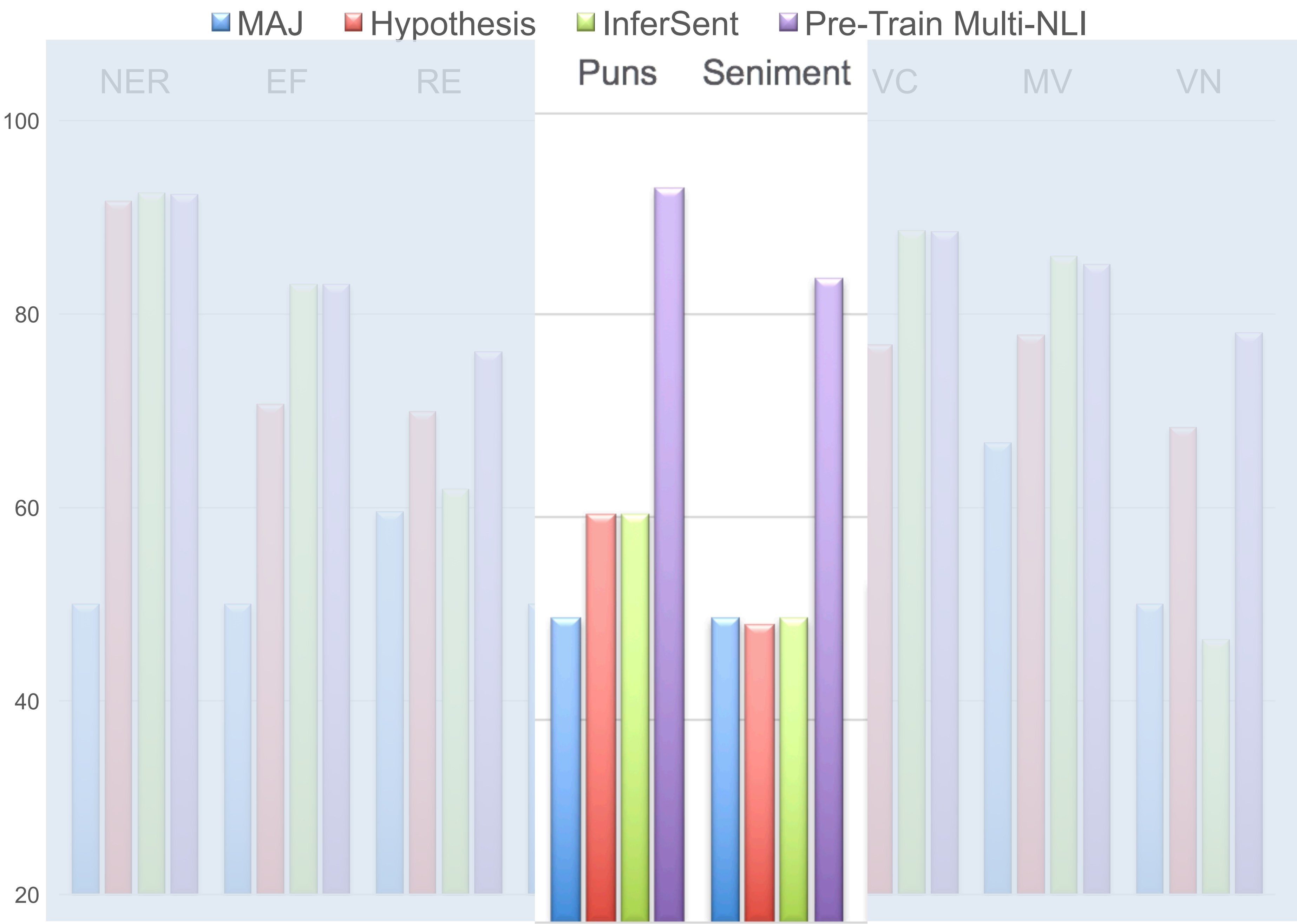
Train models on each DNC dataset

Pre-train models on all of DNC or Multi-NLI


MAJ
Hypothesis
InferSent
NER
EF
RE
Puns
Seniment
VC
MV
VN
100
80
60
40
20



JOHNS HOPKINS UNIVERSITY


MAJ
Hypothesis
InferSent
Pre-Train Multi-NLI
NER
EF
RE
Puns
Seniment
VC
MV
VN
100
80
60
40
20

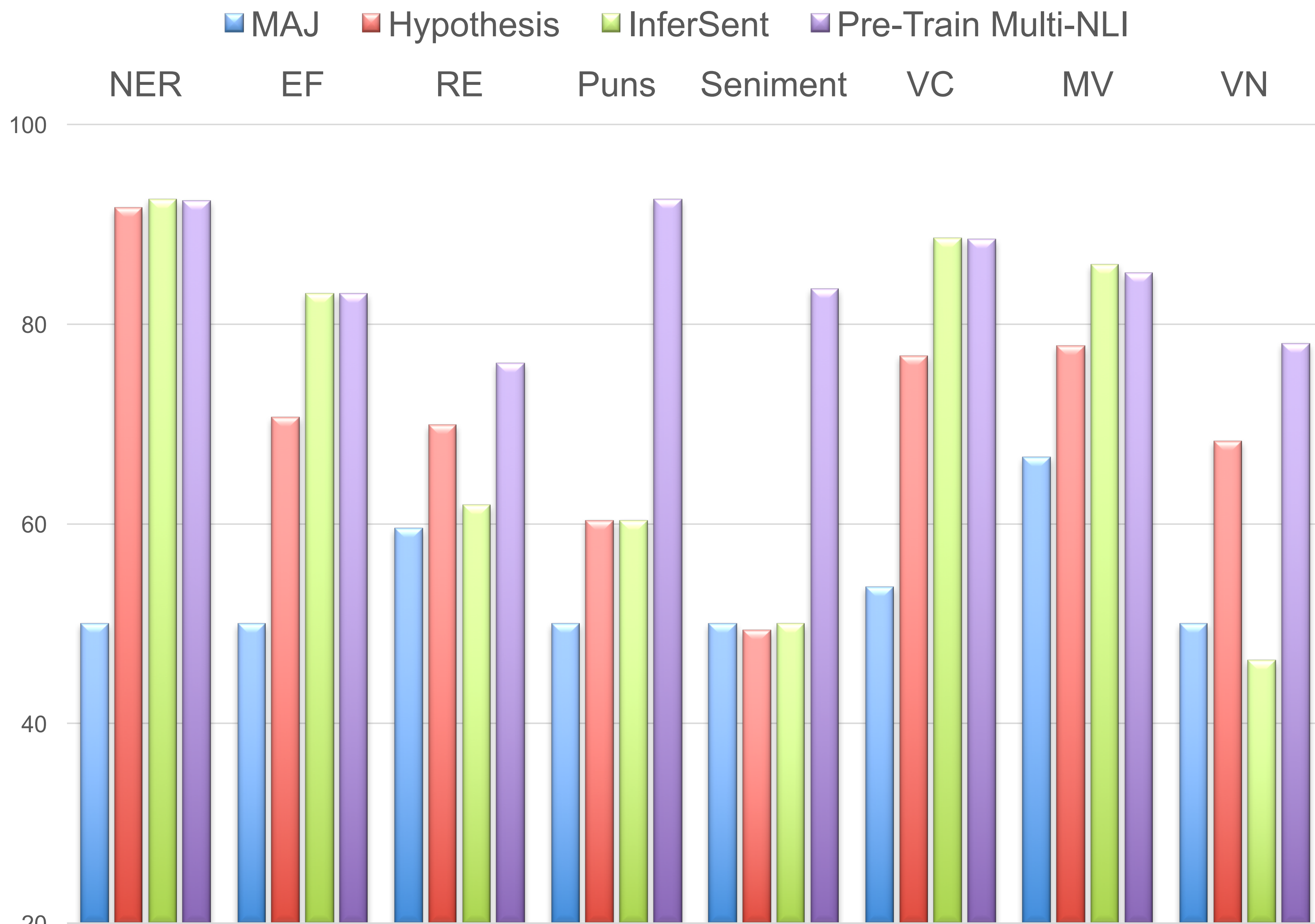
Experimental Setup

Train models on each DNC dataset

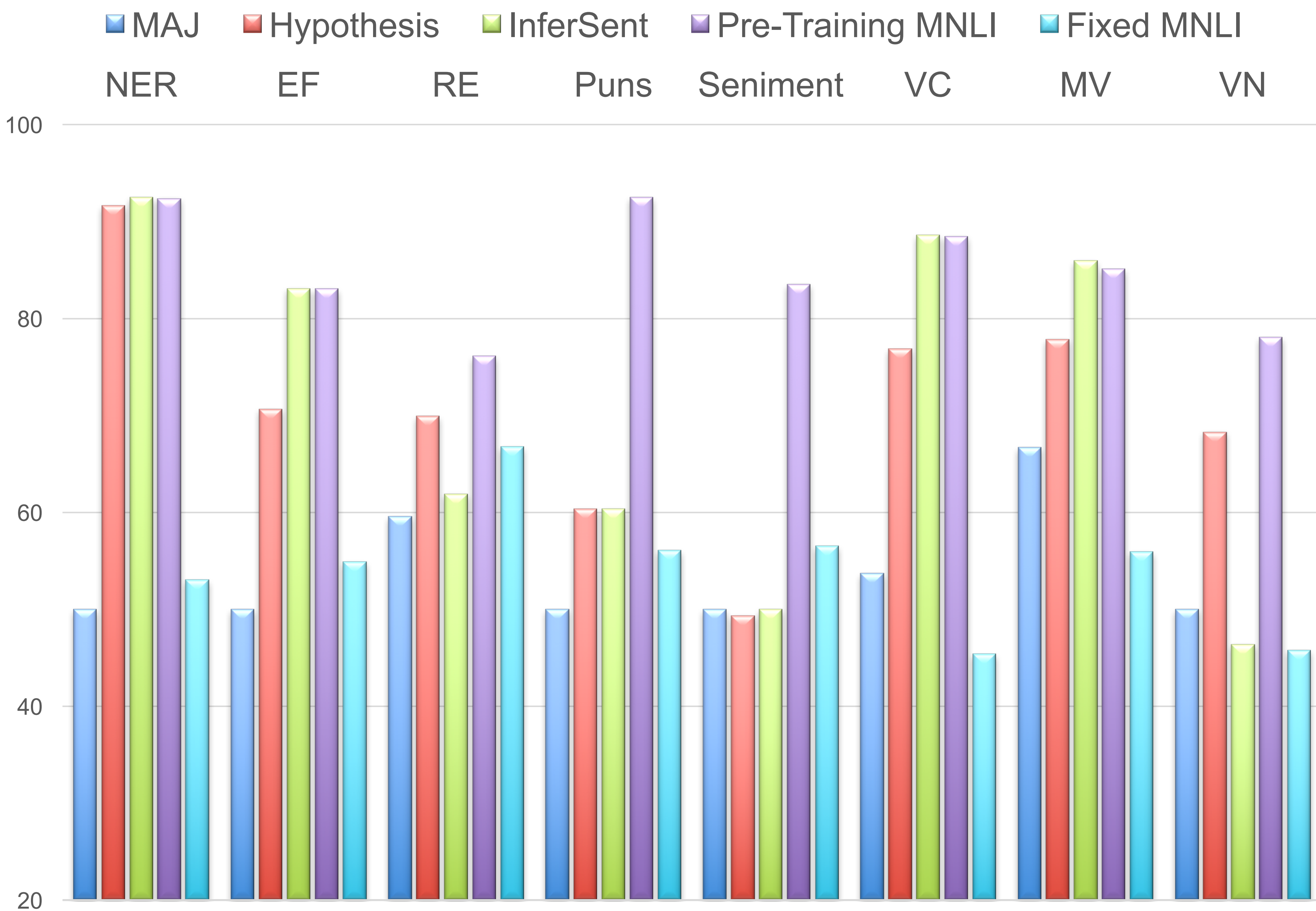
Pre-train models on all of DNC or Multi-NLI

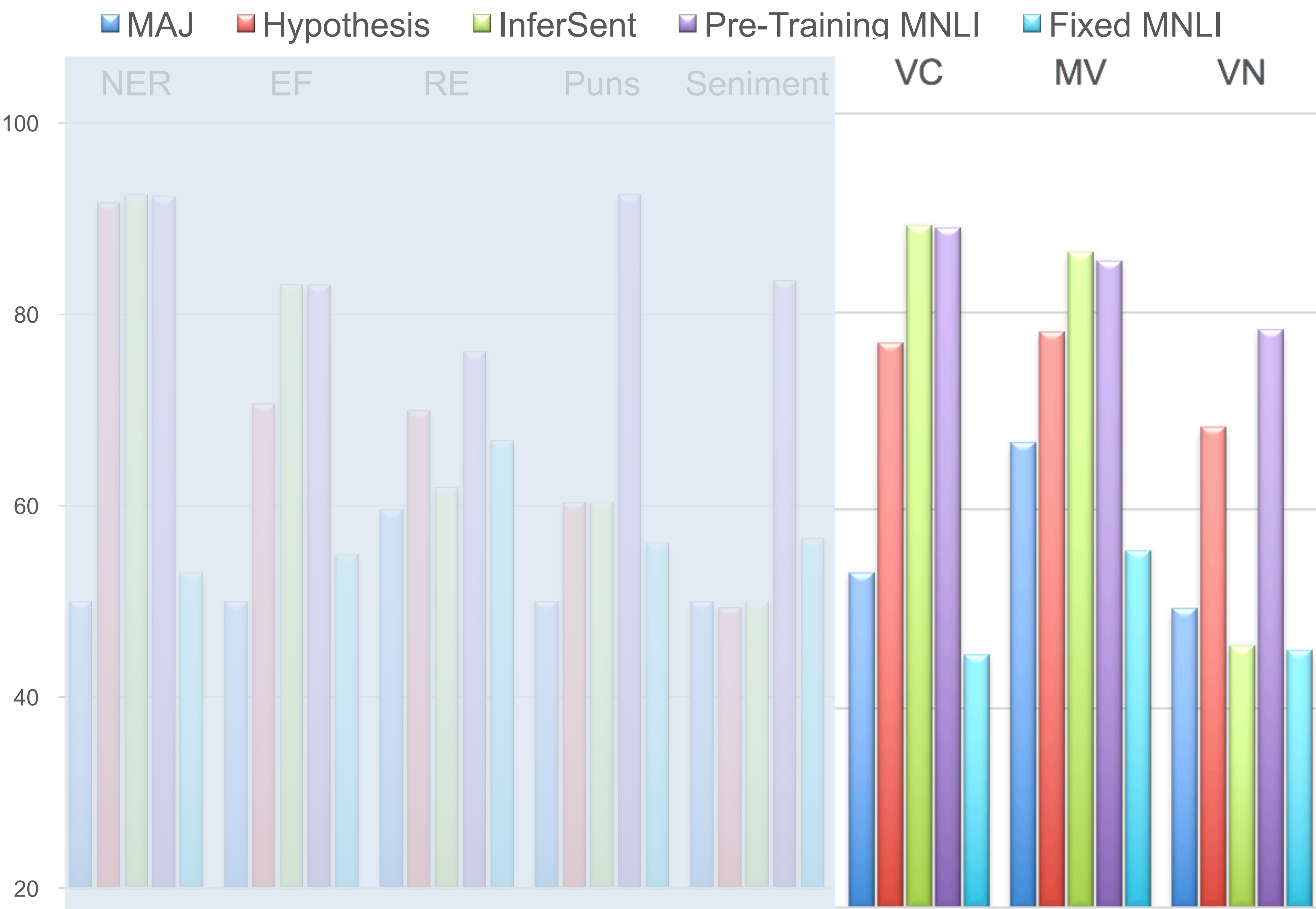
Evaluate fixed models trained on all of DNC or Multi-NLI

Johns Hopkins University



JOHNS HOPKINS
UNIVERSITY


MAJ
Hypothesis
InferSent
Pre-Training MNLI
Fixed MNLI
NER
EF
RE
Puns
Seniment
VC
MV
VN
100
80
60
40
20



JOHNS HOPKINS
UNIVERSITY

Summary

Summary

The ***DNC: Diverse NLI Collection***

Summary

The ***DNC: Diverse NLI Collection***

Convert 13 existing datasets into NLI covering 7 semantic phenomena

JOHNS HOPKINS UNIVERSITY

Summary

The ***DNC: Diverse NLI Collection***

Convert 13 existing datasets into NLI covering 7 semantic phenomena

Over half a million examples

JOHNS HOPKINS UNIVERSITY

Summary

The ***DNC: Diverse NLI Collection***

Convert 13 existing datasets into NLI covering 7 semantic phenomena

Over half a million examples

Presented use case of DNC

Call to the Community

Call to the Community

Dataset creators:

Call to the Community

Dataset creators:

convert your data into NLI

Call to the Community

Dataset creators:

- convert your data into NLI
- included in future DNC releases

Call to the Community

Dataset creators:

convert your data into NLI

included in future DNC releases

Model creators:

Call to the Community

Dataset creators:

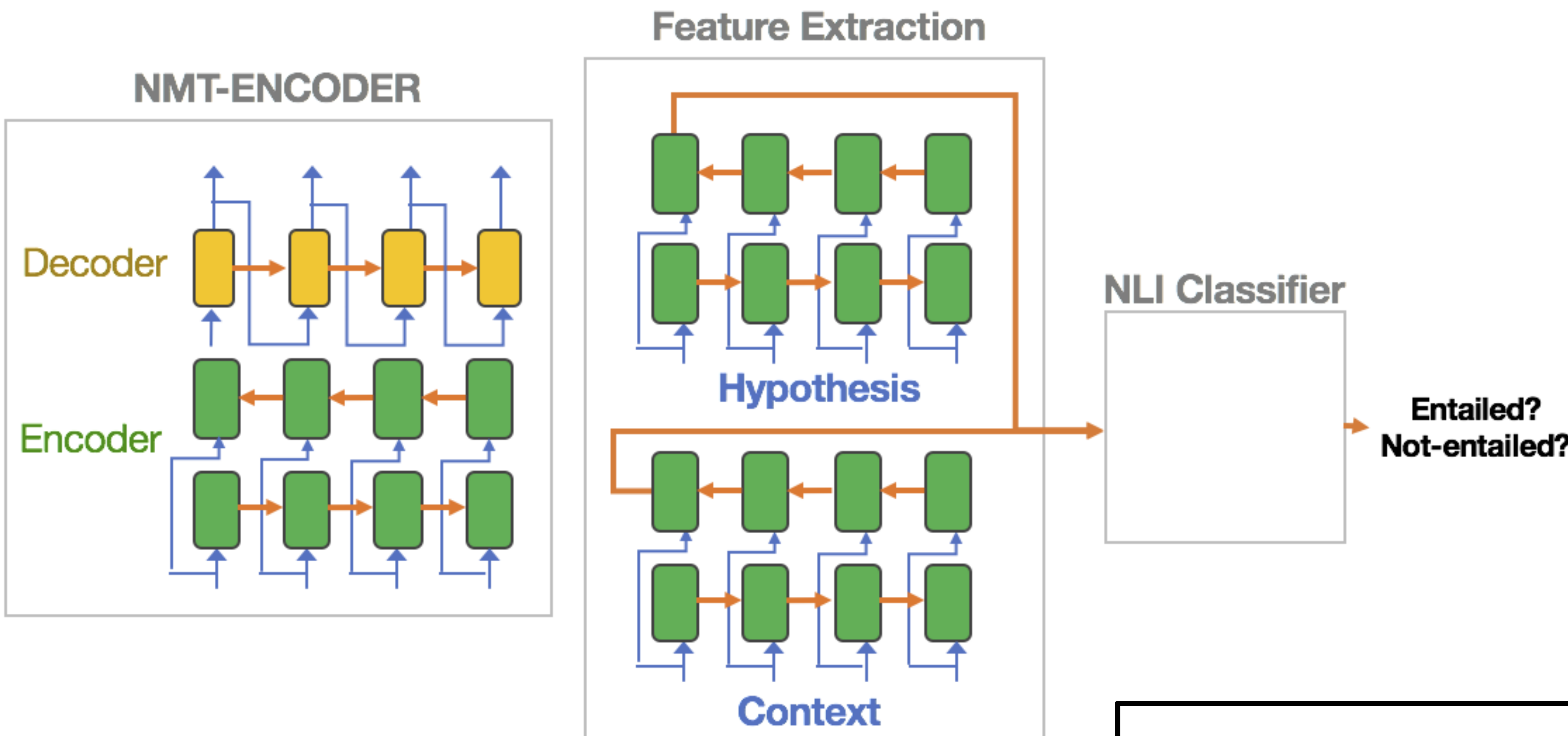
convert your data into NLI

included in future DNC releases

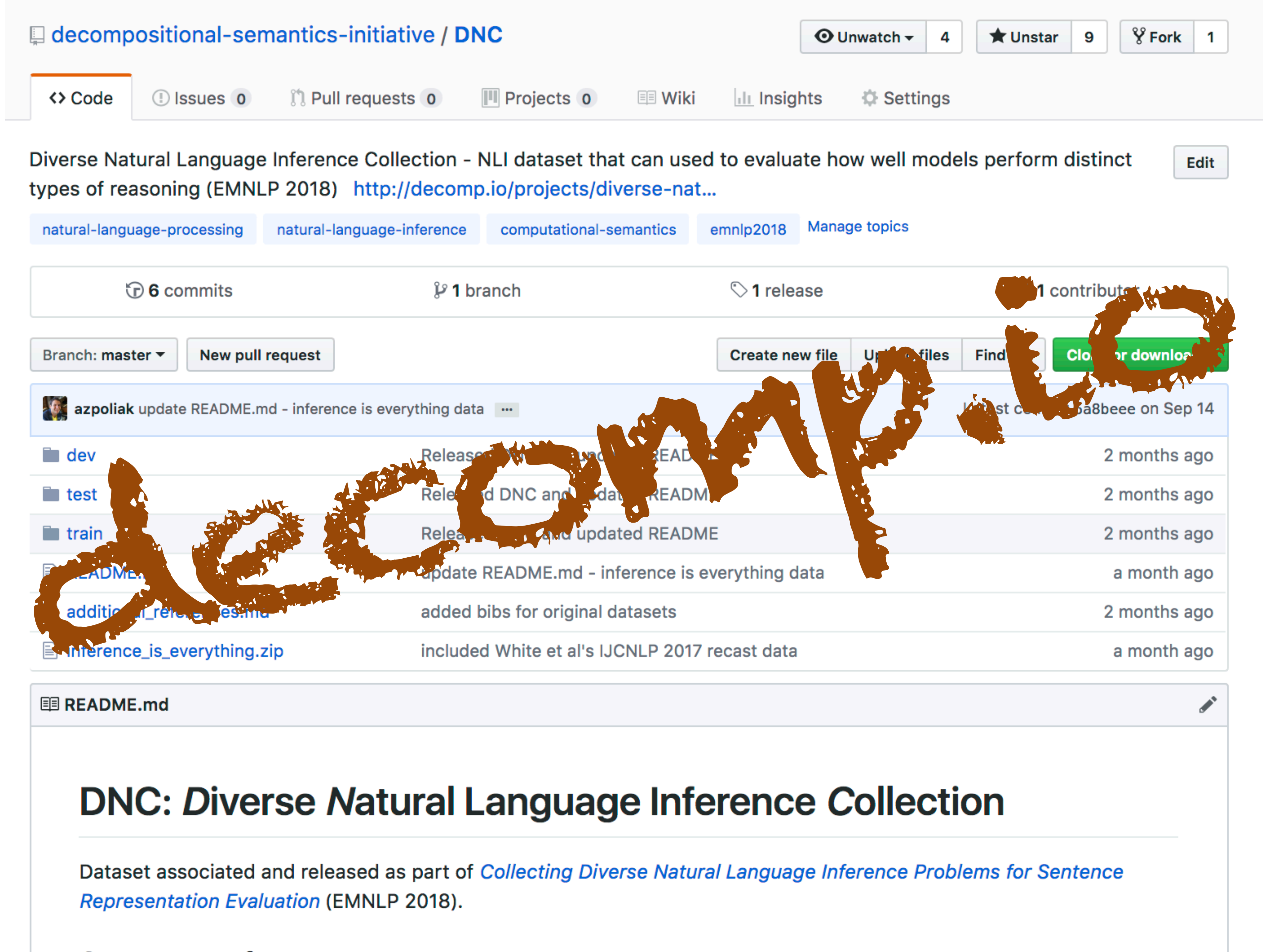
Model creators:

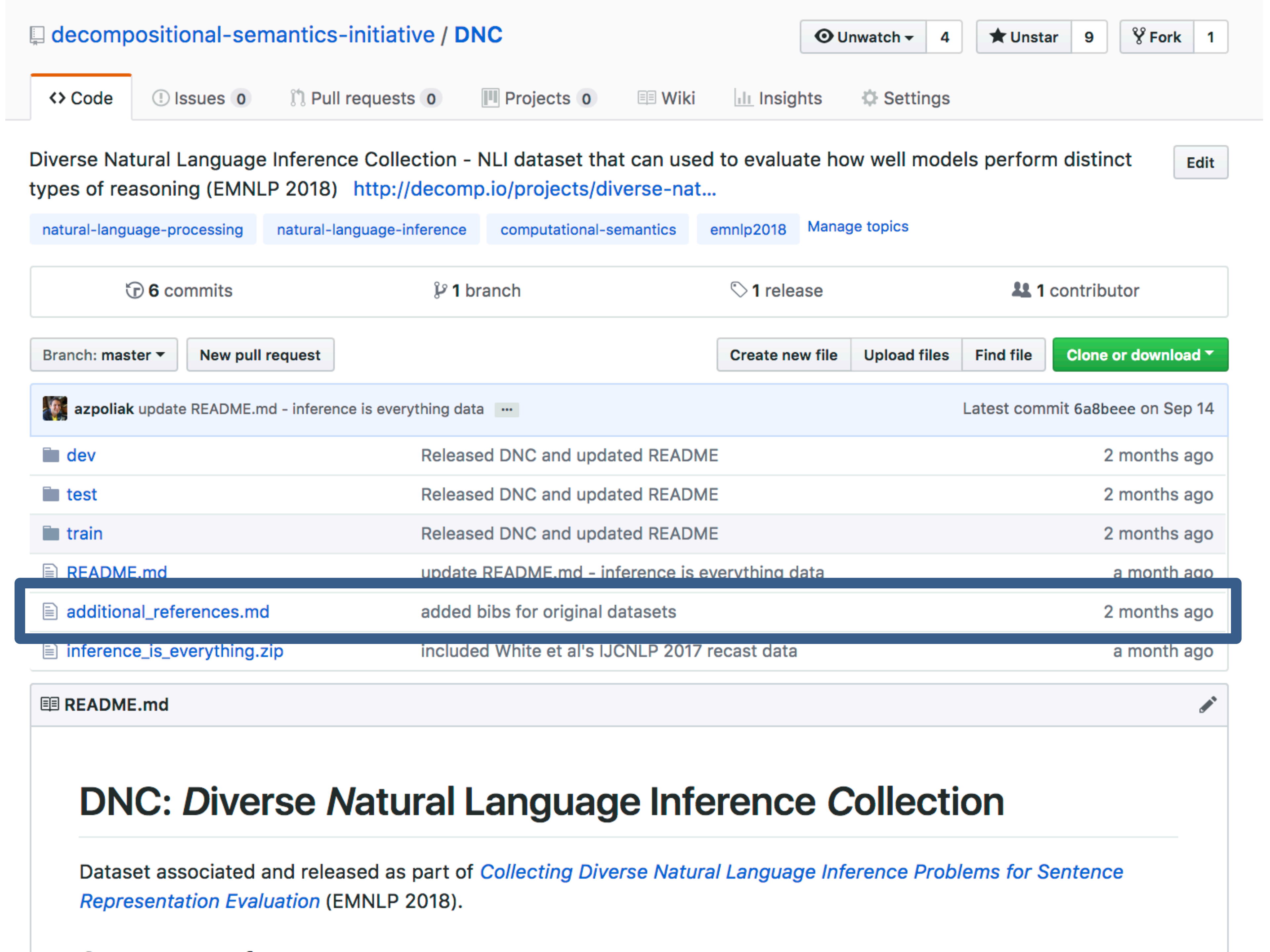
test your models ability to capture diverse types of reasoning

On the Evaluation of Semantic Phenomena in NMT Using NLI



(Poliak et. al. NAACL 2018)


decompositional-semantics-initiative / DNC
Unwatch 4
Unstar 9
Fork 1
Code
Issues 0
Pull requests 0
Projects 0
Wiki
Insights
Settings
Diverse Natural Language Inference Collection - NLI dataset that can used to evaluate how well models perform distinct types of reasoning (EMNLP 2018) http://decomp.io/projects/diverse-nat...
Edit
natural-language-processing
natural-language-inference
computational-semantics
emnlp2018
Manage topics
6 commits
1 branch
1 release
Branch: master
New pull request
Create new file
azpoliak update README.md - inference is everything data
dev
2 months ago
test
2 months ago
train
Released DNC and updated README
2 months ago
update README.md - inference is everything data
a month ago
added bibs for original datasets
2 months ago
inference_is_everything.zip
included White et al's IJCNLP 2017 recast data
a month ago
README.md
DNC: Diverse Natural Language Inference Collection
Dataset associated and released as part of Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation (EMNLP 2018).
decomp.io



decompositional-semantics-initiative / DNC
Unwatch 4
Unstar 9
Fork 1
Code
Issues 0
Pull requests 0
Projects 0
Wiki
Insights
Settings
Diverse Natural Language Inference Collection - NLI dataset that can used to evaluate how well models perform distinct types of reasoning (EMNLP 2018) http://decomp.io/projects/diverse-nat...
Edit
natural-language-processing
natural-language-inference
computational-semantics
emnlp2018
Manage topics
6 commits
1 branch
1 release
1 contributor
Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
azpoliak update README.md - inference is everything data
Latest commit 6a8beee on Sep 14
dev
Released DNC and updated README
2 months ago
test
Released DNC and updated README
2 months ago
train
Released DNC and updated README
2 months ago
README.md
update README.md - inference is everything data
a month ago
additional_references.md
added bibs for original datasets
2 months ago
inference_is_everything.zip
included White et al's IJCNLP 2017 recast data
a month ago
README.md
DNC: Diverse Natural Language Inference Collection
Dataset associated and released as part of Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation (EMNLP 2018).

Branch: master | **DNC / dev /** | Create new file | Upload files | Find file | History

azpoliak Released DNC and updated README — Latest commit 399c4f7 on Aug 31

..		
recast_factuality_data.json	Released DNC and updated README	2 months ago
recast_factuality_metadata.json	Released DNC and updated README	2 months ago
recast_kg_relations_data.json	Released DNC and updated README	2 months ago
recast_megaveridicality_data.json	Released DNC and updated README	2 months ago
recast_megaveridicality_metadata.json	Released DNC and updated README	2 months ago
recast_ner_data.json	Released DNC and updated README	2 months ago
recast_ner_metadata.json	Released DNC and updated README	2 months ago
recast_puns_data.json	Released DNC and updated README	2 months ago
recast_puns_metadata.json	Released DNC and updated README	2 months ago
recast_sentiment_data.json	Released DNC and updated README	2 months ago
recast_sentiment_metadata.json	Released DNC and updated README	2 months ago
recast_verbcorner_data.json	Released DNC and updated README	2 months ago
recast_verbcorner_metadata.json	Released DNC and updated README	2 months ago
recast_verbnet_data.json	Released DNC and updated README	2 months ago
recast_verbnet_metadata.json	Released DNC and updated README	2 months ago

Data Example


```
{
    "binary-label": false,
    "context": "The hygienist flossed my teeth .",
    "hypothesis": "My teeth took care of the hygienist .",
    "label": "not-entailed",
    "label-set": [
        "entailed",
        "not-entailed"
    ],
    "pair-id": 504820,
    "split": "dev",
    "type-of-inference": "Thematic Roles"
},
```


JOHNS HOPKINS
UNIVERSITY

MetaData Example


```
{
    "corpus": "VerbNet",
    "corpus-license": "http://verbs.colorado.edu/verbn
    "corpus-sent-id": "floss-41.2.1_NP V NP",
    "creation-approach": "automatic",
    "misc": {
        "descriptionNumber": "0.2",
        "secondary": "Transitive",
        "xtag": ""
    },
    "pair-id": 504820
},
```

Structure of json files:

Data files:

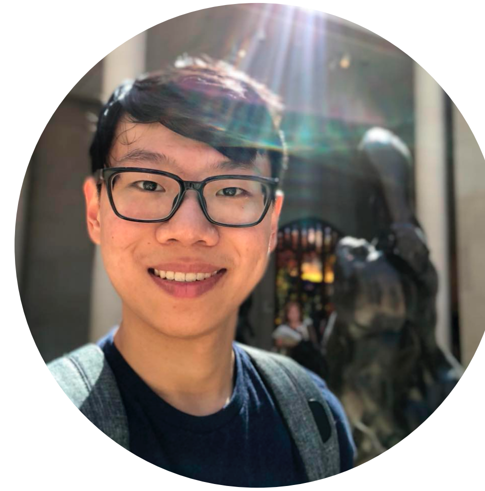
Each datafile has the following keys and values:

- `context` : The context sentence for the NLI pair. The context is already tokenized.
- `hypothesis` : The hypothesis sentence for the NLI pair. The hypothesis is already tokenized.
- `label` : The label for the NLI pair
- `label-set` : The set of possible labels for the specific NLI pair
- `binary-label` : A `True` or `False` label. See the paper for details on how we convert the `label` into a binary label.
- `split` : This can be `train` , `dev` , or `test` .
- `type-of-inference` : A string indicating what type of inference is tested in this example.
- `pair-id` : A unique integer id for the NLI pair. The `pair-id` is used to find the corresponding metadata for any given NLI pair

Metadata files:

- `pair-id` : A unique integer id for the NLI pair.
- `corpus` : The original corpus where this example came from.
- `corpus-sent-id` : The id of the sentence (or example) in the original dataset that we recast.
- `corpus-license` : The license for the data from the original dataset.
- `creation-approach` : Determines the method used to recast this example. Options are `automatic` , `manual` , or `human-labeled` .
- `misc` : A dictionary of other relevant information. This is an optional field.

Thank you!



JOHNS HOPKINS UNIVERSITY

Data and paper available

decomp.io