InfoTabS: Inference on Tables as Semi-structured data

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz
Vivek Srikumar

School of Computing, University of Utah

Reasoning in NLP task

Looking Beyond the Surface:
A Challenge Set for Reading Comprehension over Multiple Sentences

Daniel Khashabi[1], Snigdha Chaturvedi[2], Michael Roth[3], Shyam Upadhyay[1], Dan Roth[1]
[1]University of Pennsylvania, [2]University of California, Santa Cruz, [3]Saarland University
{danielkh,shyamupa,danroth}@cis.upenn.edu, snigdha@ucsc.edu, mroth@coli.uni-sb.de

On the Capabilities and Limitations of Reasoning
for Natural Language Understanding

Daniel Khashabi[1], Erfan Sadeqi Azer[2], Tushar Khot[3], Ashish Sabharwal[3], Dan Roth[1]
[1]University of Pennsylvania, [2]Indiana University, [3]Allen Institute for Artificial Intelligence
{danielkh,danroth}@cis.upenn.edu, esadeqia@indiana.edu, {tushark,ashishs}@allenai.org

QUAREL: A Dataset and Models for
Answering Questions about Qualitative Relationships

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, Ashish Sabharwal
Allen Institute for AI, Seattle, WA
{oyvindt,peterc,mattg,scottyih,ashishs}@allenai.org

Understanding Context (BERT)

Acquiring Background Knowledge (NELL)

Understandable Reasoning

Question Answering via Integer Programming over Semi-Structured Knowledge

Daniel Khashabi†, Tushar Khot‡, Ashish Sabharwal‡, Peter Clark‡, Oren Etzioni‡, Dan Roth†
†University of Illinois at Urbana-Champaign, IL, U.S.A.
{khashab2,danr}@illinois.edu
‡Allen Institute for Artificial Intelligence (AI2), Seattle, WA, U.S.A.
{tushark,ashishs,peterc,orene}@allenai.org

SciTail: A Textual Entailment Dataset from Science Question Answering

Tushar Khot, Ashish Sabharwal, Peter Clark
Allen Institute for Artificial Intelligence, Seattle, WA, U.S.A.
{tushark,ashishs,peterc}@allenai.org

Breaking NLI Systems with Sentences that Require Simple Lexical Inferences

Max Glockner[1], Vered Shwartz[2] and Yoav Goldberg[2]
[1]Computer Science Department, TU Darmstadt, Germany
[2]Computer Science Department, Bar-Ilan University, Ramat-Gan, Israel
{maxg216,vered1986,yoav.goldberg}@gmail.com

"Ask not what Textual Entailment can do for You..."

Mark Sammons V.G.Vinod Vydiswaran Dan Roth
University of Illinois at Urbana-Champaign
{mssammon|vgvinodv|danr}@illinois.edu

AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples

Dongyeop Kang[1] Tushar Khot[2] Ashish Sabharwal[2] Eduard Hovy[1]
[1]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Allen Institute for Artificial Intelligence, Seattle, WA, USA
{dongyeok,hovy}@cs.cmu.edu {tushark,ashishs}@allenai.org

Neural Semantic Parsing with Type Constraints for Semi-Structured Tables

Jayant Krishnamurthy,[1] Pradeep Dasigi,[2] and Matt Gardner[1]
[1]Allen Institute for Artificial Intelligence
[2]Carnegie Mellon University
{jayantk, mattg}@allenai.org, pdasigi@cs.cmu.edu

Compositional Semantic Parsing on Semi-Structured Tables

Panupong Pasupat
Computer Science Department
Stanford University
ppasupat@cs.stanford.edu

Percy Liang
Computer Science Department
Stanford University
pliang@cs.stanford.edu

Analyzing Compositionality-Sensitivity of NLI Models

Yixin Nie* **Yicheng Wang*** **Mohit Bansal**
Department of Computer Science
University of North Carolina at Chapel Hill
{yixin1, yicheng, mbansal}@cs.unc.edu

Be Consistent! Improving Procedural Text Comprehension using Label Consistency

Xinya Du[1*] **Bhavana Dalvi Mishra**[2] **Niket Tandon**[2] **Antoine Bosselut**[2]
Wen-tau Yih[2] **Peter Clark**[2] **Claire Cardie**[1]
[1]Department of Computer Science, Cornell University, Ithaca, NY
{xdu, cardie}@cs.cornell.edu
[2]Allen Institute for Artificial Intelligence, Seattle, WA

Reasoning about Actions and State Changes by Injecting Commonsense Knowledge

Niket Tandon*, **Bhavana Dalvi Mishra***, **Joel Grus, Wen-tau Yih, Antoine Bosselut, Peter Clark**
Allen Institute for AI, Seattle, WA
{nikett,bhavanad,joelg,scottyih,antoineb,peterc}@allenai.org

Many Many more

1. Allen Ai2
   a. Arsito
   b. Mosaic
2. UW
3. CMU
4. UNC
5. Stanford
6. Related Group

Tabular Natural Language Inference

NLI is the process of reasoning about inferential relationships, meaning to establish whether a hypothesis is a true (entailment), false (contradiction), or undetermined (neutral) given a premise.

We propose a new natural language inference dataset, **InfoTabS**, to study the problem of reasoning about semi-structured data.

Thus, reasoning over **semi-structured, multi-domain, and heterogeneous data**, where premises are **Wiki InfoBox**, and hypotheses are **human written sentences.**

Hypothesis

H1: Dressage was introduced in the Olympic games in 1912.

H2: Both men and women compete in the sport of Dressage.

H3: A dressage athlete can participate in both individual and team events.

H4: FEI governs dressage only in the U.S.

Dressage

Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Contact	No
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

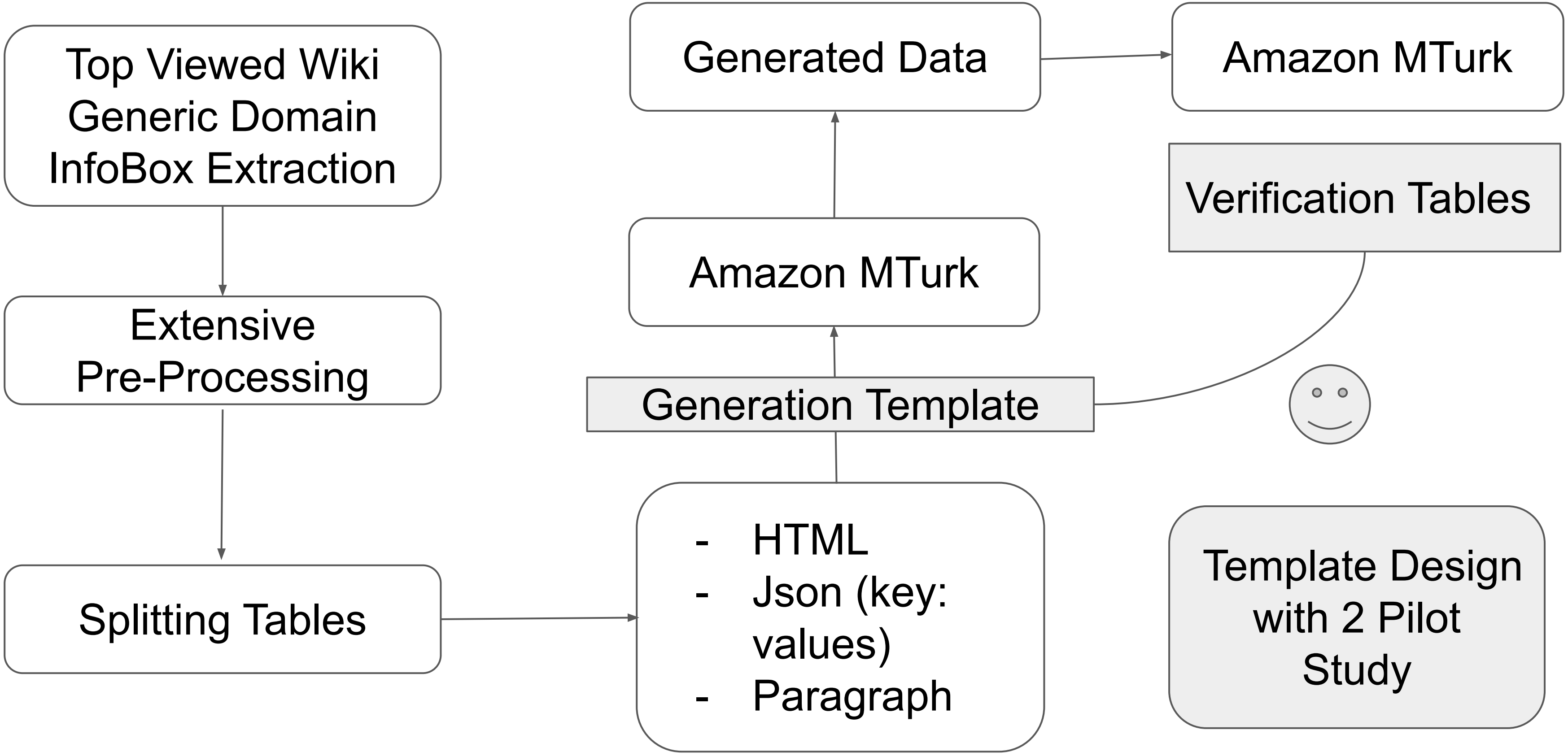
InfoTabS

Why a new dataset? -: SNLI (Caption as premise), MNLI (Diverse but stull single sentence) - Limited complex reasoning (few multi-hop & multi-row)

Why Tables? Tables are semi-structured and hence **encourage complex reasoning** which require composition of multiple types of inferences that combine multiple rows from the tables with knowledge about the world.

To determine that the **hypothesis H2** entails the premise table (Dressage), we need to look at multiple rows of the table, **understand the meaning of the row labeled as Mixed gender**, and also conclude that **Dressage is a sport**.

Construction of InfoTabS



Annotation Artifacts

Models trained on NLI datasets are prone to **learning spurious patterns** (e.g. Poliak et al., 2018)

Models can easily predict **correct labels** even with **incomplete or noisy inputs** i.e. no reasoning.

For instance, ***'not'*** and ***'no'*** in a hypothesis are **correlated with contradictions** (Niven and Kao, 2019)

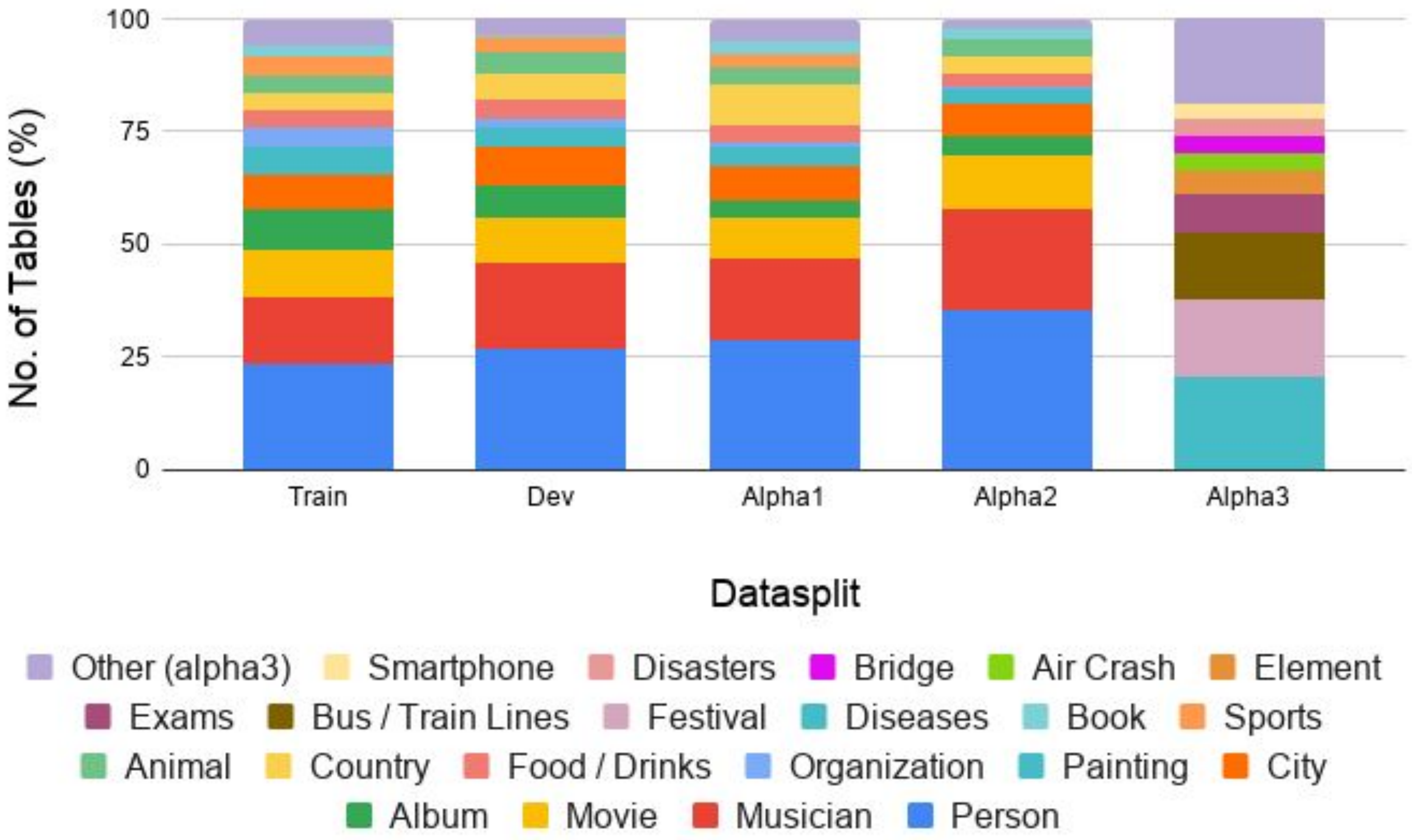
Classifiers trained on the hypotheses (ignoring the premises completely) report high accuracy; they **exhibit hypothesis bias**

The Case for multiple Test Splits

Single fix test set is not enough

We need multiple test sets (of similar sizes) with controlled differences from each other.

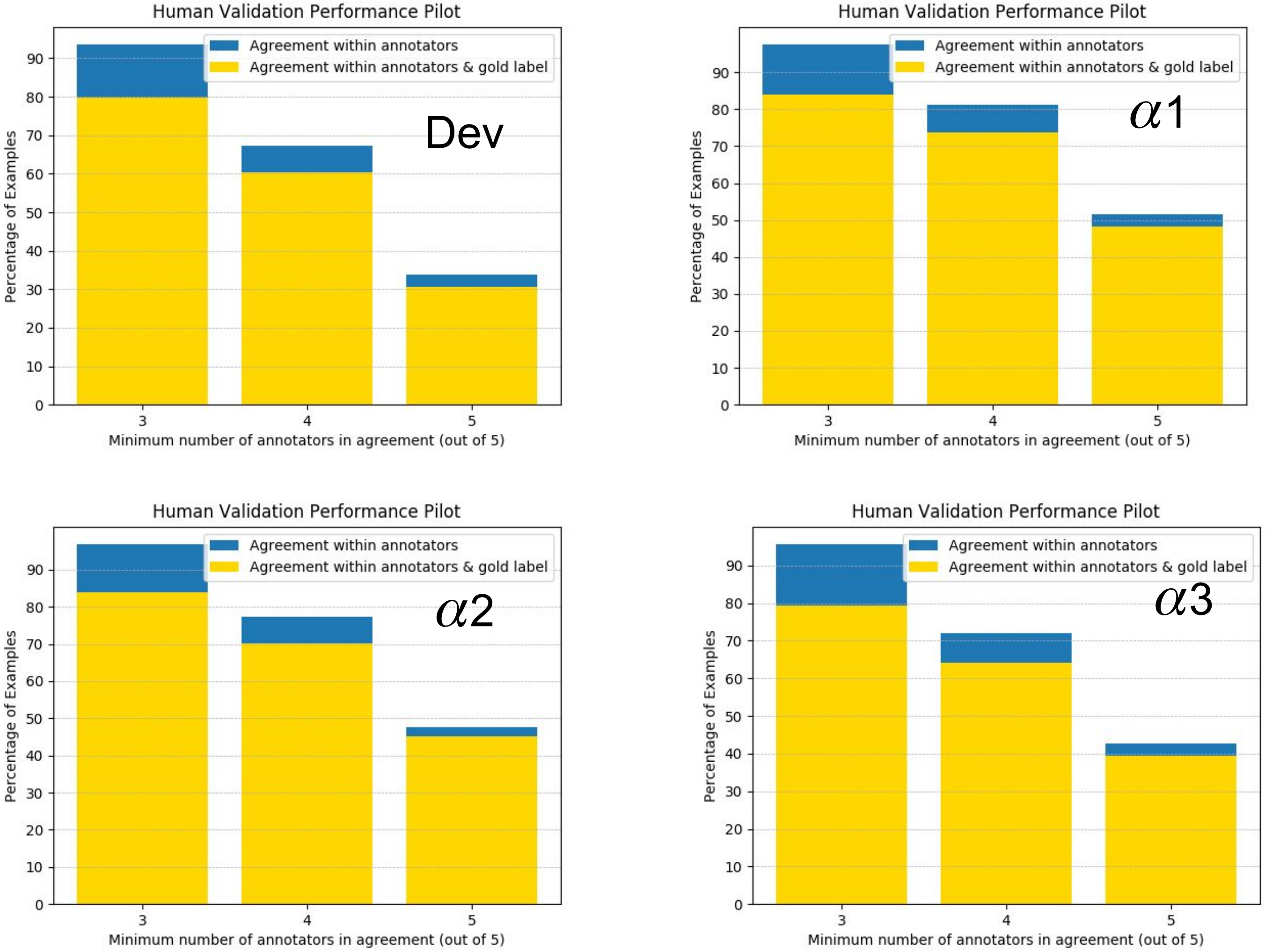
$\alpha 1$	**similar in distribution** to the **training data** in terms of **lexical makeup** of the hypotheses and the **domains of the premises**.
$\alpha 2$	New pairs where **experts change the label** of the hypothesis by **change in minimum number of keywords in the hypothesis**. Entail becomes contradict and vice-versa. Neutral remains unchanged.
$\alpha 3$	uses **premises from domains not in the training split**, but necessitate, **similar types of reasoning** to arrive at the decision.



No. of Tables (%)
100
75
50
25
0
Train
Dev
Alpha1
Alpha2
Alpha3
Datasplit
Other (alpha3)
Smartphone
Disasters
Bridge
Air Crash
Element
Exams
Bus / Train Lines
Festival
Diseases
Book
Sports
Animal
Country
Food / Drinks
Organization
Painting
City
Album
Movie
Musician
Person

Dataset Statistics

Data Split	#Tables	#Pairs
Train	1955	16538
Dev	200	1800
$\alpha 1$	200	1800
$\alpha 2$	200	1800
$\alpha 3$	200	1800


Human Validation Performance Pilot
Agreement within annotators
Agreement within annotators & gold label
Dev
Percentage of Examples
Minimum number of annotators in agreement (out of 5)
Human Validation Performance Pilot
Agreement within annotators
Agreement within annotators & gold label
$\alpha 1$
Percentage of Examples
Minimum number of annotators in agreement (out of 5)
Human Validation Performance Pilot
Agreement within annotators
Agreement within annotators & gold label
$\alpha 2$
Percentage of Examples
Minimum number of annotators in agreement (out of 5)
Human Validation Performance Pilot
Agreement within annotators
Agreement within annotators & gold label
$\alpha 3$
Percentage of Examples
Minimum number of annotators in agreement (out of 5)

Inter-annotator Agreement Statistics

Data Split	Cohen's Kappa	Human Accuracy	Majority Agreement
Dev	0.78	79.8	93.5
$\alpha 1$	0.80	84.0	97.5
$\alpha 2$	0.80	84.0	96.8
$\alpha 3$	0.74	79.3	95.6

Reasoning Analysis

We adapted the set of reasoning categories from **GLUE benchmark** for Table premises. We also define some new reasonings not in GLUE.

***Simple lookup*:** hypothesis is formed by literally restating the fact from table

Multi-row reasoning: requires multiple rows to make an inference

Subjective/out-of-table: involves value judgments about a proposition or reference to information out of the table that is neither well known/common sense

Finally, **authors independently annotated 160 pairs** from the **dev and α3 test sets each**, and edge cases were discussed to arrive at consensus labels.

Example from Pilot Study

Amsterdam	
• Municipality	219.32 km^2 (84.68 sq mi)
• Land	165.76 km^2 (64.00 sq mi)
• Water	53.56 km^2 (20.68 sq mi)
• Randstad	3,043 km^2 (1,175 sq mi)
Elevation	−2 m (−7 ft)

E : Amsterdam has a municipality less than 250 km2. **(numerical)**

N: Amsterdam has the largest land area in Netherlands. **(world knowledge)**

C : Amsterdam has over 3500 km2 Randstad. **(numerical)**

E : Parts of Amsterdam are below sea level. **(common - sense, world-knowledge)**

N : Amsterdam is the largest city in the Randstad **(world knowledge)**

C : There are fewer square kilometers of land than water in Amsterdam. **(common - sense, logical)**

Angelina Jolie DCMG	
Born	Angelina Jolie Voight (1975-06-04) June 4, 1975 (age 43) Los Angeles, California, U.S.
Citizenship	• United States • Cambodia
Occupation	• Actress • filmmaker • activist
Years active	1982–present
Spouse(s)	• Jonny Lee Miller (m. 1996; div. 2000) • Billy Bob Thornton (m. 2000; div. 2003) • Brad Pitt (m. 2014; sep. 2016)
Children	6
Parent(s)	• Jon Voight • Marcheline Bertrand
Relatives	• James Haven (brother) • Barry Voight (uncle) • Chip Taylor (uncle)

E: Angelina Jolie was born in the summer of 1975. **(common sense, world-knowledge)**

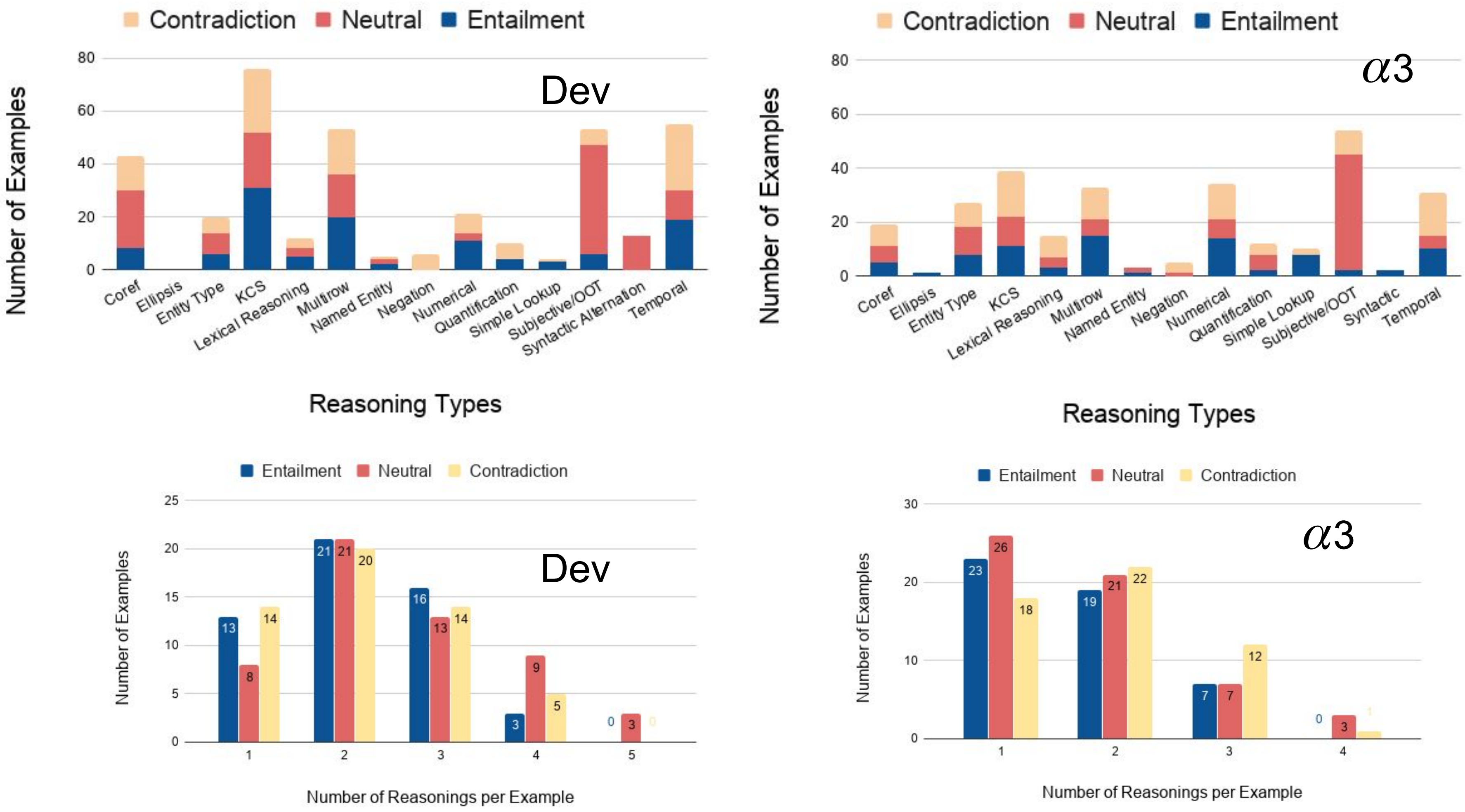
N: Angelina Jolie has 6 sons. **(common-sense, world knowledge)**

C: Angelina Jolie has been married four times. **(numerical)**

E: Angelina Jolie is 43 years old. **(lexical)**

N: Angelina Jolie is currently married. **(lexical)**

C: Angelina Jolie has 2 brothers. **(numerical)**



Contradiction
Neutral
Entailment
Dev
Number of Examples
Reasoning Types
Coref
Ellipsis
Entity Type
KCS
Lexical Reasoning
Multirow
Named Entity
Negation
Numerical
Quantification
Simple Lookup
Subjective/OOT
Syntactic Alternation
Temporal
$\alpha3$
Syntactic
Number of Reasonings per Example

Reasoning Properties

Semi-structured premises force the annotators to call upon **knowledge & common sense** about the world.

- because information about the entities and their types is not explicitly stated in tables. E.g. *“X was born in the summer”* for a person whose birth is in May in New York

Neutrals are more inclined to being **subjective/out-of-table** since anything which is not mentioned in the table or is subjective is a neutral statement.

Tables for α3 are from different domains, hence **not of the same distribution** as the previous splits.

- Expected as we cannot expect temporal reasoning from tables in a domain that does not contain temporal quantities.

Overall Reasoning Properties

1. **Multi-row reasoning** - multi-sentience document reasoning
   - Combining information from multiple sources for inference

2. **Multi-hop reasoning** - multiple levels of reasoning
   - Involved organised multiple reasoning for final inference

3. **Multi-Domain reasoning** - multiple sources of reasoning
   - Variability in the data, i.e diversity in the dataset (multi-domain)

4. **Open-endedness** - generated sentences not simple verbatim
   - represents information that is not explicitly stated (only inferred)

Premise Table Representation

Premise as Paragraph (Para): For a table titled *t*, a row with key *k* and value *v* will be written as the sentence The *k* of t are *v*.

- E.g. for Dressage table, the row with key Equipment will be converted into the sentence *“The equipment of Dressage are horse, appropriate horse tack”*.

Premise as Sentence (Sent): hypotheses are typically short, they may be derived from a small subset of rows. Use word mover distance (Kusner et al., 2015) to find top1 (wmd1) and top3 (wmd3)

Premise as Structure 1 (TabFact): Follow Chen et al. (2019) and represent tables by a sequence of <key> : <value> tokens, concatenated by “ ; ”.

Hypothesis Bias

Adversarial Baseline to check Hypothesis Bias

Training a classifier by **ignoring the premise** (only hypothesis baseline)

Training a classifier with a **dummy premise** (*"to be or not to be"*)

Training a classifier with a **swapped premise** (random premise taken)

Does our dataset exhibit hypothesis bias?

Classifier train by ignoring the premise (hypothesis only model)

Model	Dev	$\alpha 1$	$\alpha 2$	$\alpha 3$
SVM	59.00	60.61	45.89	45.90
RoBERT (L)	60.5	60.48	48.26	48.89

Classifier train with a dummy premise (*"to be or not to be"*)
Or swapped premise (random premise taken)

Premise	Dev	$\alpha 1$	$\alpha 2$	$\alpha 3$
Dummy	60.2	59.78	48.91	46.37
Swapped	63.81	63.15	50.3	51.31

Analysis

All the BERT-class models discover annotation artifacts equally well.

However, performance on α**2** and α**3** data splits is worse (~ 12% gap) compared to dev and α**1** since the artifacts in the training data do not occur in these splits.

How do pre-trained NLI systems perform on our dataset?

Premise	**Dev**	$\boldsymbol{\alpha 1}$	$\boldsymbol{\alpha 2}$	$\boldsymbol{\alpha 3}$
Train on SNLI (SNLI Test Accuracy 92.5 %)				
WMD-1	49.33	47.61	49.44	46.50
Para	52.94	52.11	52.78	46.28
Train on MNLI (MNLI test accuracy matched 89.0 %, mis-matched 88.9%)				
WMD-1	44.23	44.72	46.94	43.94
Para	**53.11**	**51.33**	**53.06**	**47.39**

Analysis

All the BERT-class models discover annotation artifacts equally well.

However, performance on α**2** and α**3** data splits is worse (~ 12% gap) compared to dev and α**1** since the artifacts in the training data do not occur in these splits.

Pre-trained NLI systems trained on SNLI & MNLI do not perform well.

Full premise is better than single sentence a) ineffectiveness of *wmd* to get correct top sentence or b) sentences require multi-row reasoning.

Does Training on Paragraph/Sentence Premise help?

Model	**Premise**	**Dev**	$\alpha 1$	$\alpha 2$	$\alpha 3$
SVM	Para	59.11	59.17	46.44	41.28
BERT (base)	Para	63.0	63.54	52.57	48.17
RoBERT (base)	Para	67.2	66.98	56.87	55.36
	WMD-1	67.26	66.15	56.24	53.48
RoBERT (large)	WMD-3	70.09	69.69	59.8	57.13
	Para	**76.04**	**74.28**	**66.8**	**64.37**

Analysis

All the BERT-class models discover annotation artifacts equally well.

However, performance on α**2** and α**3** data splits is worse (~ 12% gap) compared to dev and α**1** since the artifacts in the training data do not occur in these splits.

Pre-trained NLI systems trained on SNLI & MNLI do not perform well.

Full premise is better than single sentence a) ineffectiveness of *wmd* to get correct top sentence or b) sentences require multi-row reasoning.

Training on full/sentence premise help BERT-class model significantly (10-14%).

Does Training on Structured Premise (TabFact) help?

Model	Dev	$\alpha 1$	$\alpha 2$	$\alpha 3$
BERT (base)	63.67	64.04	53.59	49.05
RoBERT (base)	68.06	66.7	56.87	55.26
RoBERT (large)	**77.31**	**76.7**	**67.22**	**65.67**

Analysis

All the BERT-class models discover annotation artifacts equally well.

However, performance on α**2** and α**3** data splits is worse (~ 12% gap) compared to dev and α**1** since the artifacts in the training data do not occur in these splits.

Pre-trained NLI systems trained on SNLI & MNLI do not perform well.

Full premise is better than single sentence a) ineffectiveness of *wmd* to get correct top sentence or b) sentences require multi-row reasoning.

Training on full/sentence premise help BERT-class model significantly (10-14%).

Providing premise structure help BERT-class model, ~1.3% improvement

Conclusion

Introduced a **new task/dataset InfoTabS**, with **heterogeneous semi-structured premises and natural language hypotheses**.

InfoTabS has **multiple test sets** which **poses difficulties to models** that **learn superficial correlations** between inputs and the label rather than reasoning about the information.

InfoTabS poses **several inference challenge** for **state-of-the-art BERT-Class models**, as evident from gap in human and model performance (esp. $\alpha 2$ & $\alpha 3$)

Our task **encourage new kinds of models and representations** that can handle **semi-structured information** as first class citizens.

University of Utah Data Science Faculty

SCHOOL OF COMPUTING
UNIVERSITY OF UTAH

Algorithms, "foundations" of data science

Aditya Bhaskara (Algorithms, ML)
Jeff Phillips (Geometry, Learning)
Blair Sullivan (Graphs Theory)
Suresh Venkatasubramanian (Algorithms, Fairness, ML)
Bei Wang (Comp Topology, Visualization, ML)

Databases, "applied" ML/NLP

Feifei Li (Databases, Data mining)
Vivek Srikumar (ML, NLP)
Shandian Zhe (ML, Prob Modeling)
Ellen Riloff (NLP)
Qingyao Ai (ML, IR)
Marina Kogan (Social Computing)

Several other faculty in Imaging, Visualization, HPC, ….

Thanks for Listening

Questions