Best Paper and Honourable Mention

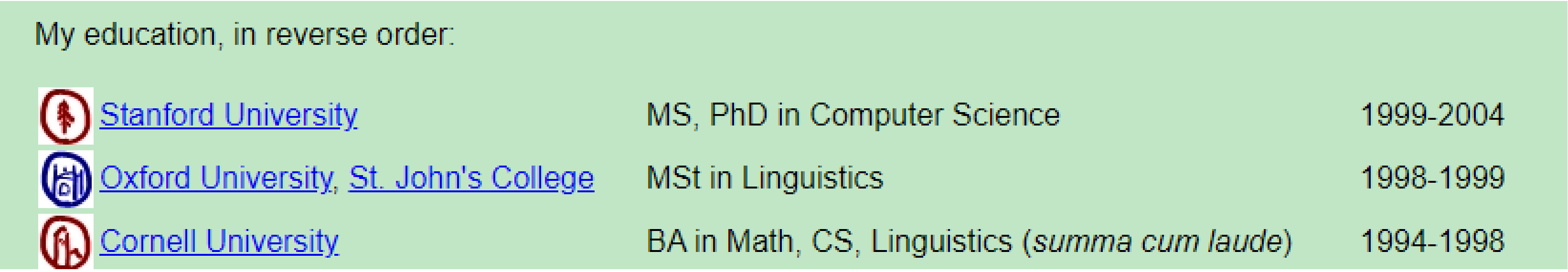
Wei Wang

Overview

- Digital Voicing of Silent Speech
- Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems
- GLUCOSE: GeneraLized and COntextualized Story Explanations

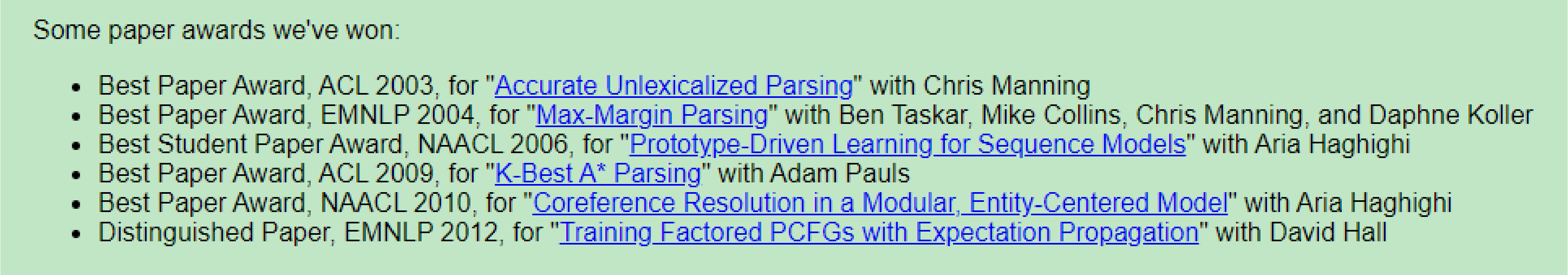
Digital Voicing of Silent Speech

David Gaddy and **Dan Klein**
University of California, Berkeley
{dgaddy,klein}@berkeley.edu



My education, in reverse order:

Stanford University	MS, PhD in Computer Science	1999-2004
Oxford University, St. John's College	MSt in Linguistics	1998-1999
Cornell University	BA in Math, CS, Linguistics (*summa cum laude*)	1994-1998



Some paper awards we've won:

- Best Paper Award, ACL 2003, for "Accurate Unlexicalized Parsing" with Chris Manning
- Best Paper Award, EMNLP 2004, for "Max-Margin Parsing" with Ben Taskar, Mike Collins, Chris Manning, and Daphne Koller
- Best Student Paper Award, NAACL 2006, for "Prototype-Driven Learning for Sequence Models" with Aria Haghighi
- Best Paper Award, ACL 2009, for "K-Best A* Parsing" with Adam Pauls
- Best Paper Award, NAACL 2010, for "Coreference Resolution in a Modular, Entity-Centered Model" with Aria Haghighi
- Distinguished Paper, EMNLP 2012, for "Training Factored PCFGs with Expectation Propagation" with David Hall

Introduction

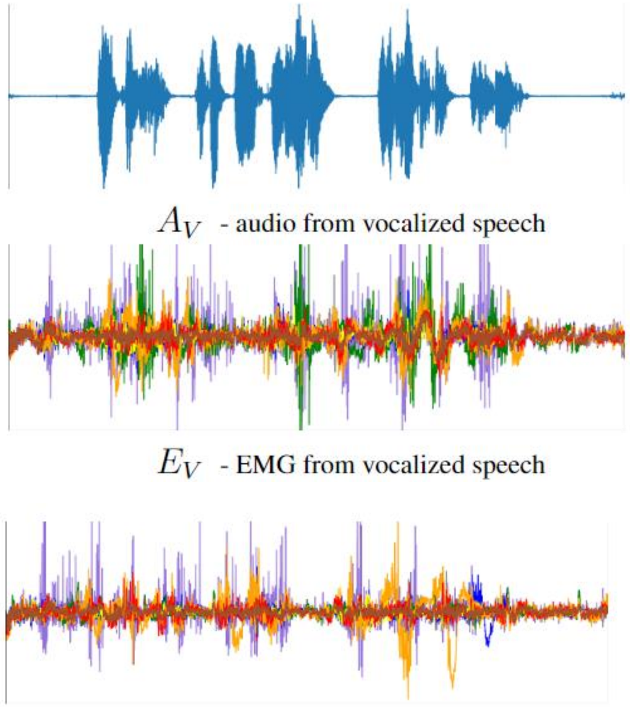
Digital Voicing of Silent Speech

- By using muscular sensor measurements of speech articulator movement, the paper aims to capture **silent speech** - utterances that have been articulated without producing sound.
- **Digital voicing**, or generating synthetic speech to be transmitted or played back.
- It could be used to create a device analogous to a Bluetooth headset that allows people to carry on phone conversations **without disrupting those around them**.
- It could be used by some people who are **no longer able to produce audible speech.**
- It could make silent speech accessible to our devices and digital assistants by leveraging existing high-quality audio-based speech-to-text systems.

Motivation

Digital Voicing of Silent Speech

- Several initial attempts have been made to convert surface electromyography (EMG) signals to speech, similar to the task we approach in this paper.
- However, these works have focused on the artificial task of **recovering audio from EMG that was recorded during vocalized speech**, rather than the end-goal task of generating from silent speech.
- The authors extend digital voicing to train on **silent EMG** ES rather than only vocalized EMG EV.
- The **challenge** is that when training on vocalized EMG data we have both EMG inputs and time-aligned speech targets, but for **silent EMG any recorded audio will be silent**. (How to get vocalized speech for silent EMG)



A_V - audio from vocalized speech

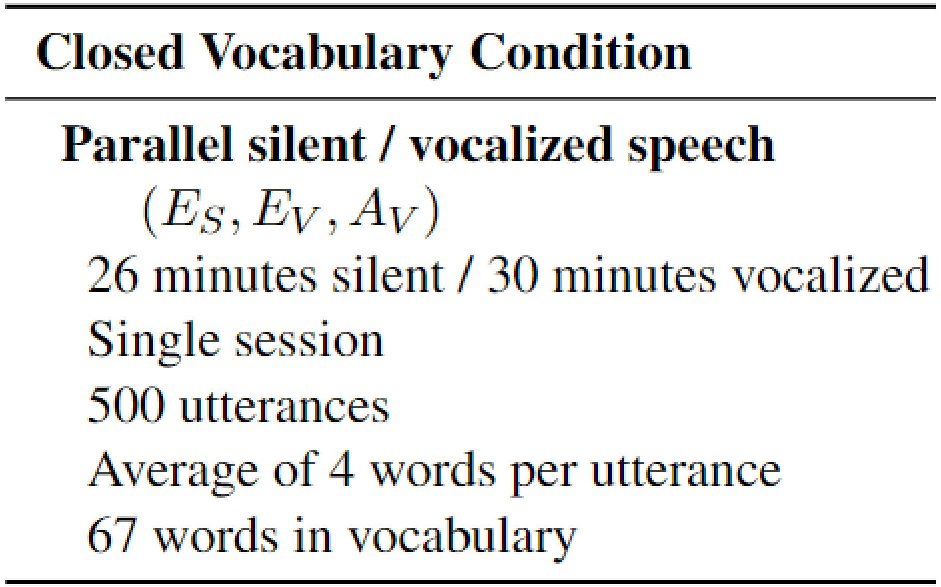
E_V - EMG from vocalized speech

E_S - EMG from silent speech

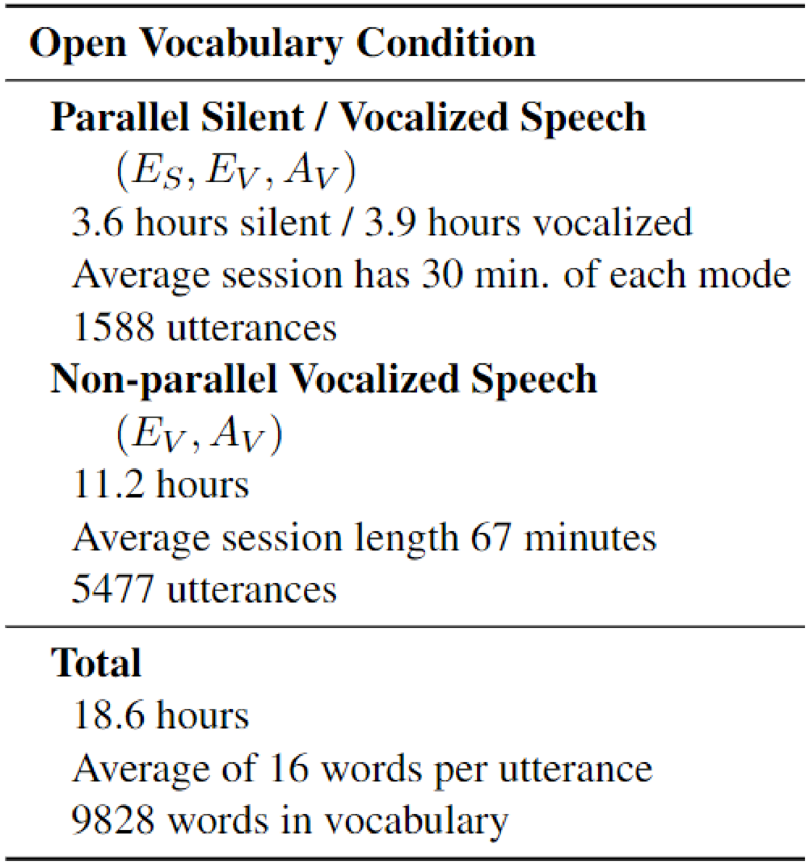
Data Collection

Digital Voicing of Silent Speech

- Closed Vocabulary: These expressions come from a small set of templates such as “<weekday> <month> <year>”
- Open Vocabulary: Sentences from books.
- 30 utterances for validation and 100 for test



Closed Vocabulary Condition
Parallel silent / vocalized speech
(E_S, E_V, A_V)
26 minutes silent / 30 minutes vocalized
Single session
500 utterances
Average of 4 words per utterance
67 words in vocabulary



Open Vocabulary Condition
Parallel Silent / Vocalized Speech
(E_S, E_V, A_V)
3.6 hours silent / 3.9 hours vocalized
Average session has 30 min. of each mode
1588 utterances
Non-parallel Vocalized Speech
(E_V, A_V)
11.2 hours
Average session length 67 minutes
5477 utterances
Total
18.6 hours
Average of 16 words per utterance
9828 words in vocabulary

Method

Digital Voicing of Silent Speech

Feature Representation

raw features of EMG and speech are converted into vectors using common methods in previous works.

EMG to Speech Feature Transducer

a bidirectional LSTM is used to convert between featurized versions of the signals, EMG and Audio.

Audio Target Transfer

- dynamic time warping (DTW)
- canonical correlation analysis (CCA)
- Refinement with Predicted Audio

WaveNet Synthesis:

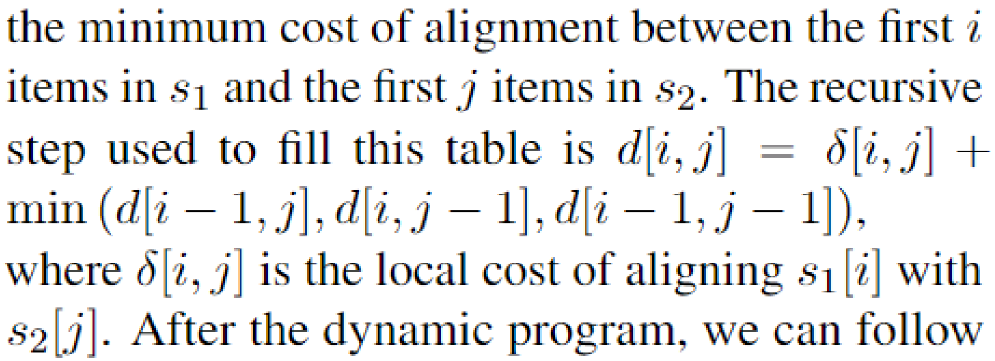
a WaveNet decoder generates the audio sample by sample conditioned on speech features.

Method

Digital Voicing of Silent Speech

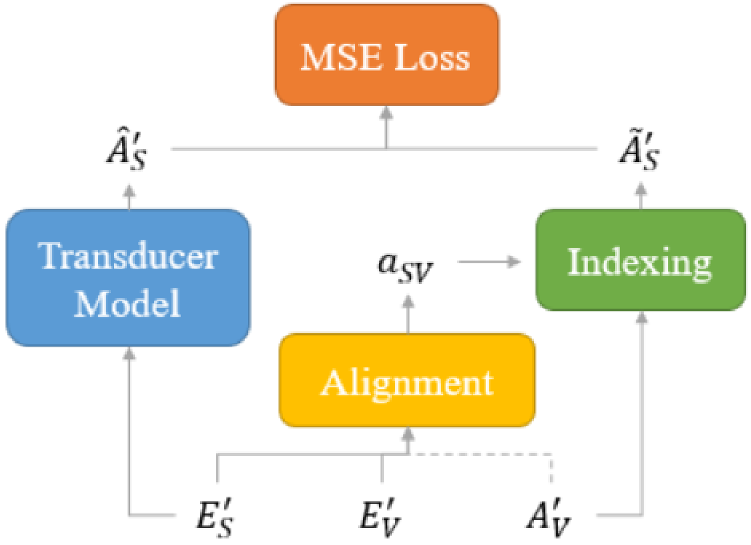
Audio Target Transfer

- dynamic time warping (DTW)



the minimum cost of alignment between the first i items in s_1 and the first j items in s_2. The recursive step used to fill this table is $d[i, j] = \delta[i, j] + \min(d[i-1, j], d[i, j-1], d[i-1, j-1])$, where $\delta[i, j]$ is the local cost of aligning $s_1[i]$ with $s_2[j]$. After the dynamic program, we can follow

- canonical correlation analysis (CCA)
- Refinement with Predicted Audio



$$\delta_{\text{EMG}}[i, j] = \|E'_S[i] - E'_V[j]\|$$

$$\delta_{\text{CCA}}[i, j] = \|P_S E'_S[i] - P_V E'_V[j]\|$$

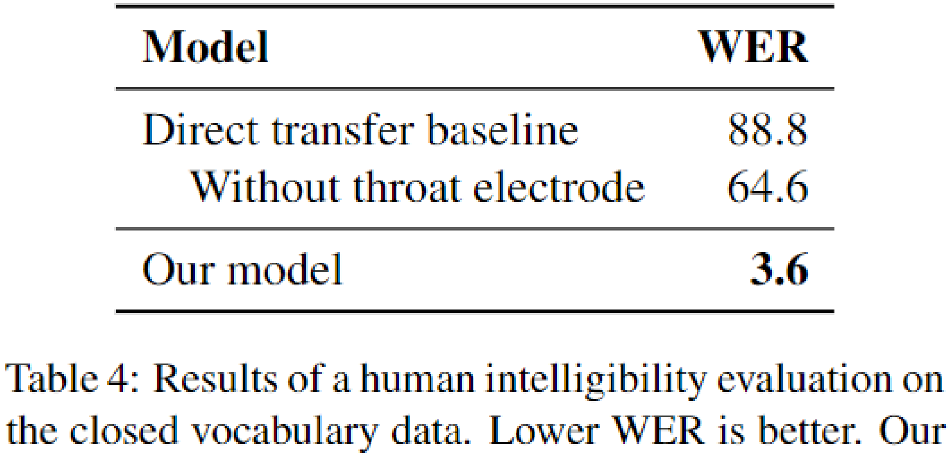
$$\delta_{\text{full}}[i, j] = \delta_{\text{CCA}}[i, j] + \lambda \left\| \hat{A}'_S[i] - A'_V[j] \right\|$$

Experiment

Digital Voicing of Silent Speech

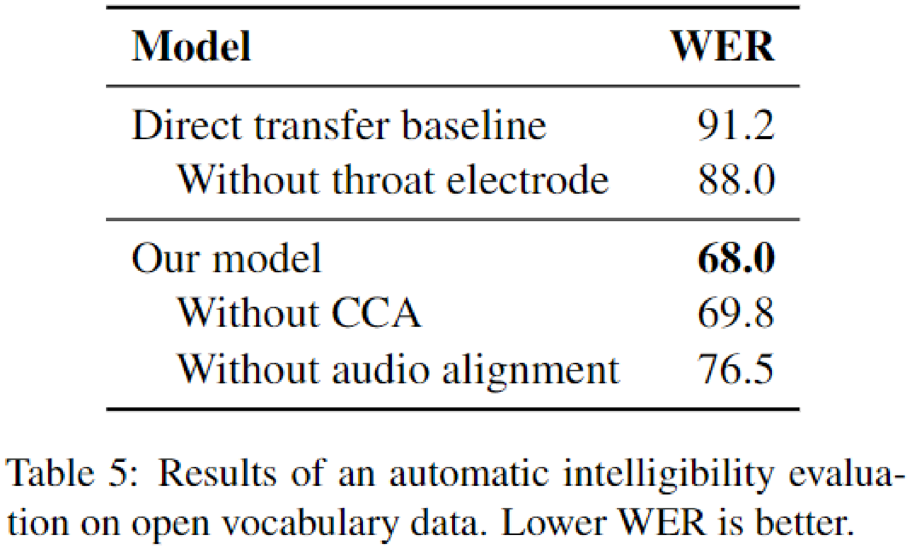
- Closed vocabulary
- Open vocabulary

$$\text{WER} = \frac{\text{substitutions} + \text{insertions} + \text{deletions}}{\text{reference length}}$$



Model	WER
Direct transfer baseline	88.8
Without throat electrode	64.6
Our model	**3.6**

Table 4: Results of a human intelligibility evaluation on the closed vocabulary data. Lower WER is better. Our



Model	WER
Direct transfer baseline	91.2
Without throat electrode	88.0
Our model	**68.0**
Without CCA	69.8
Without audio alignment	76.5

Table 5: Results of an automatic intelligibility evaluation on open vocabulary data. Lower WER is better.

Summary

Digital Voicing of Silent Speech

- The results show that digital voicing of silent speech, while still challenging in open domain settings, shows promise as an achievable technology.
- The proposed method also significantly improve intelligibility in an open vocabulary condition, with a relative error reduction over 20%.
- The authors release a new dataset of EMG signals collected during both silent and vocalized speech.

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems

Jan Deriu[1], Don Tuggener[1], Pius von Däniken[1], Jon Ander Campos[3],
Alvaro Rodrigo[2], Thiziri Belkacem[4], Aitor Soroa[3], Eneko Agirre[3], and Mark Cieliebak[1]

[1]Zurich University of Applied Sciences (ZHAW), Winterthur, Switzerland, {*deri, tuge, vode, ciel*}@*zhaw.ch*

[2]National Distance Education University (UNED), Madrid, Spain, *alvarory@lsi.uned.es*

[3]University of the Basque Country (UPV/EHU), Donostia, Spain, {*jonander.campos, a.soroa, e.agirre*}@*ehu.eus*

[4]Synapse Développement, Toulouse, France, *belkacemthiziri@gmail.com*

Introduction

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems

- Evaluation is a long-standing issue in developing conversational dialogue systems (i.e., chatbots).
- Chatbots do not solve a clearly-defined task whose success can be **measured in relation to an a priori defined ground truth.**
- Automatic metrics have so far **failed to show high correlation with human evaluations**.
- Human evaluation is necessary. However, single-turn ratings disregard the multi-turn nature of a dialogue. Most of multi-turn evaluations are based on human-bot conversations, which are **costly to obtain and tend to suffer from low quality.**

Motivation

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems

Spot The Bot, a cost-efficient evaluation methodology that can be used to rank several bots with regard to their ability to disguise as humans, is based on two observations:

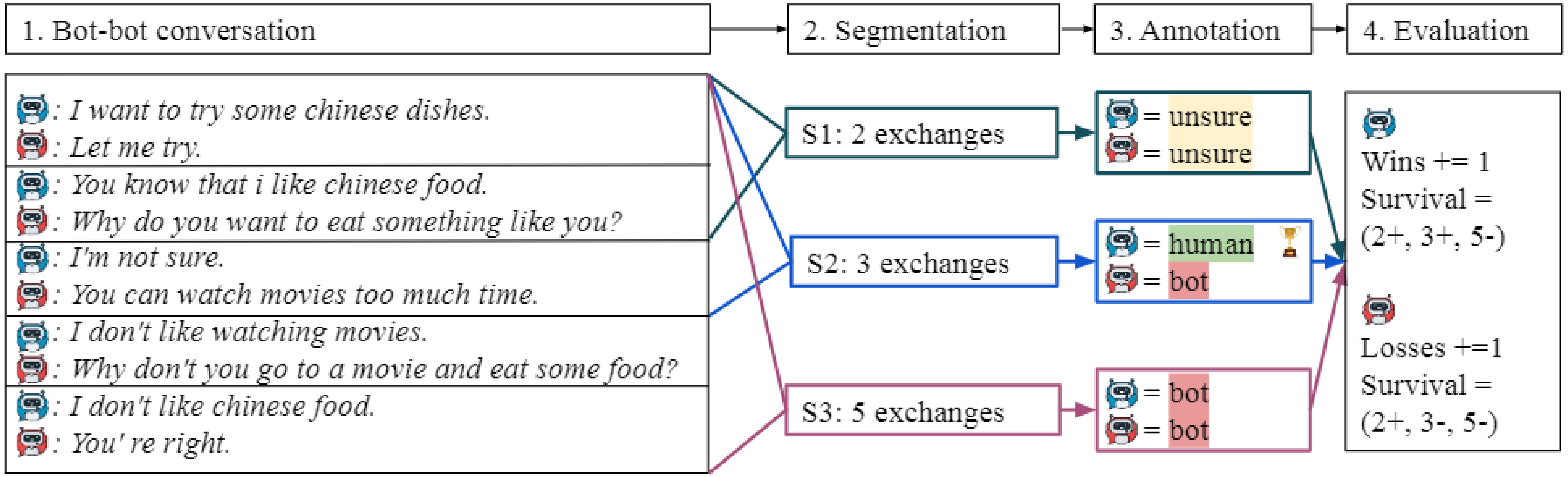
- First, chatbots are trained on conversations between humans, and thus, they should be evaluated regarding their ability to mimic human behavior.
- Second, the longer a conversation is, the more likely it is that a bot exhibits non-human-like behavior.

Spot The Bot works by generating conversations between bots, then mixing these bot-bot conversations with human-human conversations and letting human judges decide for each entity in the conversations if it is a human or a bot.

It does not rely on human-bot conversations and generally requires fewer annotations.

Method

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems



: I want to try some chinese dishes.
: Let me try.
: You know that i like chinese food.
: Why do you want to eat something like you?
: I'm not sure.
: You can watch movies too much time.
: I don't like watching movies.
: Why don't you go to a movie and eat some food?
: I don't like chinese food.
: You' re right.

The crowdworkers' task is to determine for each entity in a conversation whether it is a human or a bot (or whether the crowdworker is unsure).

The bot that is most frequently annotated as being human wins the tournament.

Method

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems

Segmentation

The more exchanges there are in a conversation, the more likely it is that a bot gets recognized as such.

Thus, different segments of the conversation are shown to the crowdworkers.

Annotation

First, the annotators have to decide for each entity in a conversation segment if it is a bot or a human.

Second, to correlate the outcome to various characteristics of a bot, the framework allows rating specific features.

The authors choose three features: sensibleness, specificity and fluency.

Method

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems

Ranking

$$\frac{\textsc{wins}(B_i, B_j)}{\textsc{wins}(B_i, B_j) + \textsc{wins}(B_j, B_i)}$$

Ranking is generated by the TrueSkill algorithm based on the win rate, and significant differences in performance are determined by bootstrap sampling.

The result is a ranked set of clusters, where each cluster is composed of entities that do not have a significant difference in performance.

Survival Analysis

We interpret the annotation data as such: the spotted event occurred if the system was annotated as “bot” and it survived if it was annotated as “unsure” or “human”.

If the dialog system was not spotted, we know it survived for at least k exchanges. If the dialogue system was spotted as such, we cannot tell the exact number of exchanges it took for an annotator to spot it, meaning it could have taken less than k exchanges.

Experiment

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems

Domains.

Dailydialog (Li et al., 2017), segments of 2, 3, and 5 exchanges

Empathetic Dialogues (Rashkin et al., 2019), 1, 2, and 3 exchanges

PersonaChat (Zhang et al., 2018), 2, 3, and 5 exchanges

Dialogue Systems.

small sequence-to-sequence model (DR)

sequence-to-sequence model (S2) with attention

GPT-2 (GPT) model

BERT-Rank (BR) model

Blender model (BL)

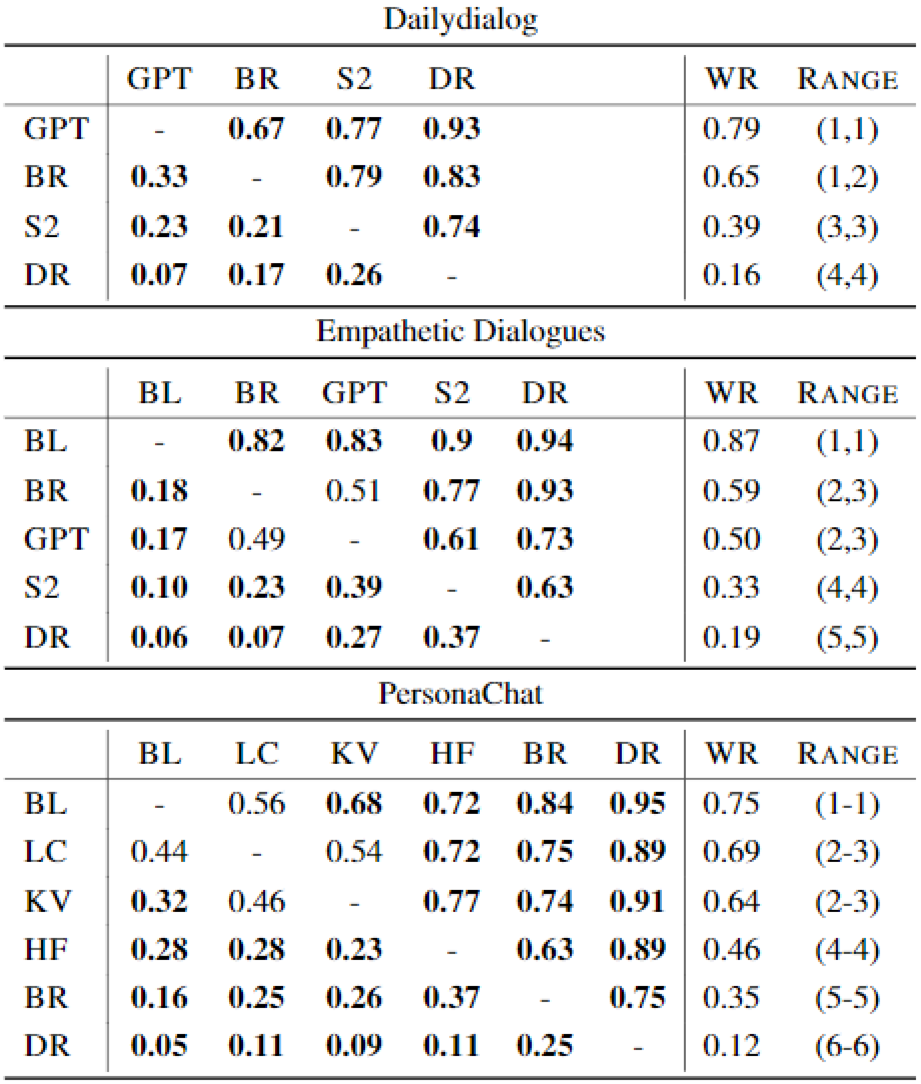
Lost in Conversation7 (LC), Huggingface (HF) and KVMemNN (KV) (for PersonaChat)

Experiment

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems

Ranking Results

- As expected, DR performs worst in all three domains, which is due to its repetitive nature
- In the Dailydialog and the Empathetic Dialogues domains, the GPT2 and the BR models perform equally, i.e., they end up in the same cluster.
- In both domains, systems using pre-trained language models outperform the S2 model, which aligns with the expectation of related findings.
- The BL model outperforms all other models in both the PersonaChat and Empathetic Dialogues domains, which is in line with related findings.



Dailydialog

	GPT	BR	S2	DR	WR	RANGE
GPT	-	**0.67**	**0.77**	**0.93**	0.79	(1,1)
BR	**0.33**	-	**0.79**	**0.83**	0.65	(1,2)
S2	**0.23**	**0.21**	-	**0.74**	0.39	(3,3)
DR	**0.07**	**0.17**	**0.26**	-	0.16	(4,4)

Empathetic Dialogues

	BL	BR	GPT	S2	DR	WR	RANGE
BL	-	**0.82**	**0.83**	**0.9**	**0.94**	0.87	(1,1)
BR	**0.18**	-	0.51	**0.77**	**0.93**	0.59	(2,3)
GPT	**0.17**	0.49	-	**0.61**	**0.73**	0.50	(2,3)
S2	**0.10**	**0.23**	**0.39**	-	**0.63**	0.33	(4,4)
DR	**0.06**	**0.07**	**0.27**	**0.37**	-	0.19	(5,5)

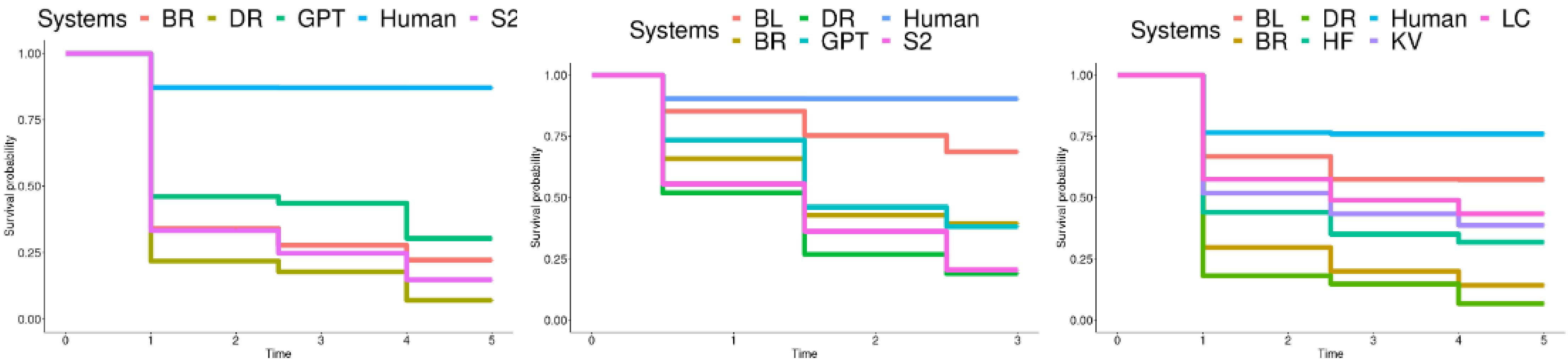
PersonaChat

	BL	LC	KV	HF	BR	DR	WR	RANGE
BL	-	0.56	**0.68**	**0.72**	**0.84**	**0.95**	0.75	(1-1)
LC	0.44	-	0.54	**0.72**	**0.75**	**0.89**	0.69	(2-3)
KV	**0.32**	0.46	-	**0.77**	**0.74**	**0.91**	0.64	(2-3)
HF	**0.28**	**0.28**	**0.23**	-	**0.63**	**0.89**	0.46	(4-4)
BR	**0.16**	**0.25**	**0.26**	**0.37**	-	**0.75**	0.35	(5-5)
DR	**0.05**	**0.11**	**0.09**	**0.11**	**0.25**	-	0.12	(6-6)

Experiment

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems

Survival Analysis



(a) Dailydialog

(b) Empathetic Dialogues

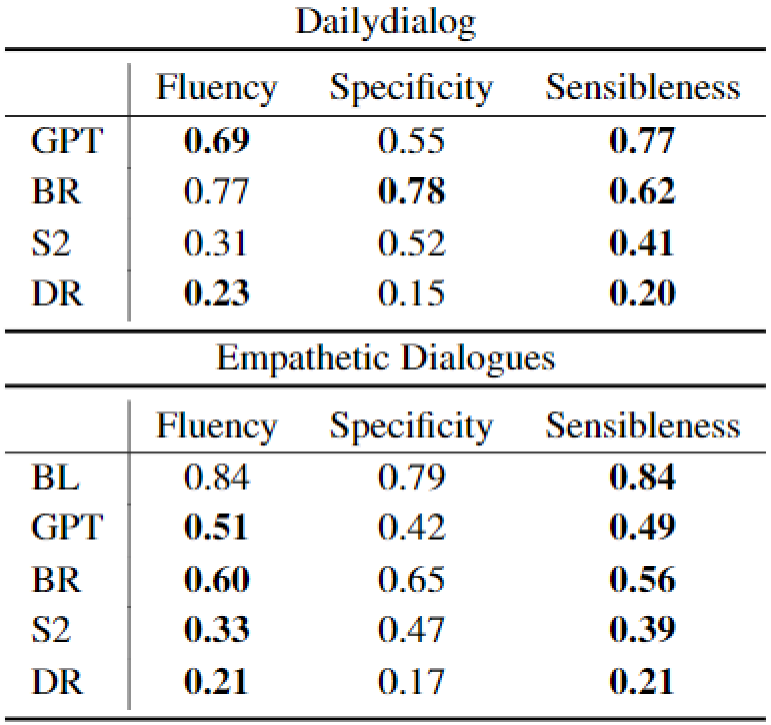
(c) PersonaChat

Further non-significant differences within the Survival Analysis are S2 and DR in the Empathetic Dialogues domain, BR and S2 in the Dailydialog domain, and LC and KV in the PersonaChat domain.

Experiment

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems

Survival Analysis

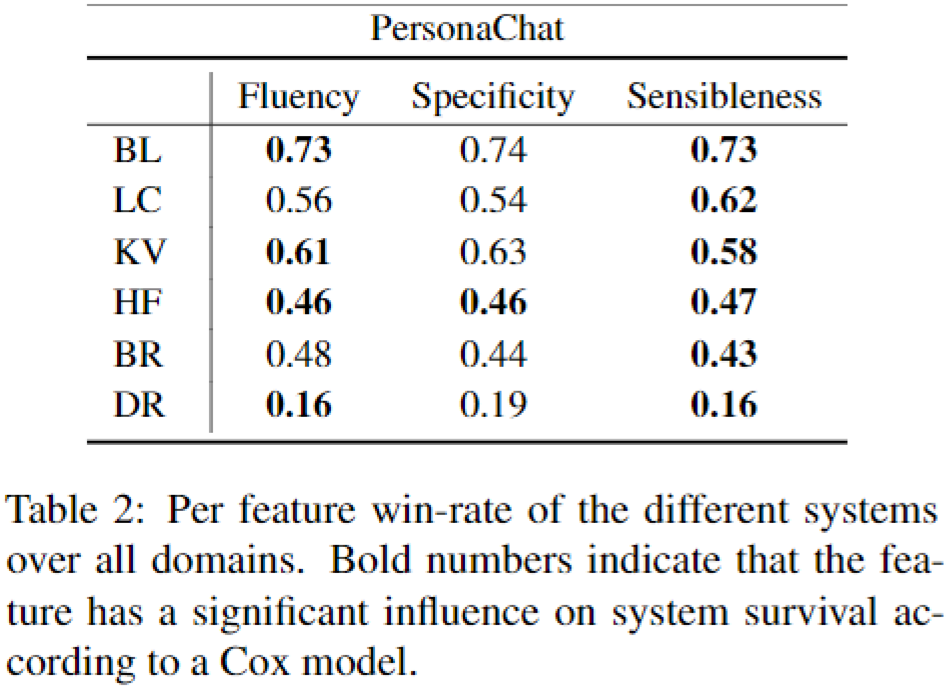


Dailydialog

	Fluency	Specificity	Sensibleness
GPT	**0.69**	0.55	**0.77**
BR	0.77	**0.78**	**0.62**
S2	0.31	0.52	**0.41**
DR	**0.23**	0.15	**0.20**

Empathetic Dialogues

	Fluency	Specificity	Sensibleness
BL	0.84	0.79	**0.84**
GPT	**0.51**	0.42	**0.49**
BR	**0.60**	0.65	**0.56**
S2	**0.33**	0.47	**0.39**
DR	**0.21**	0.17	**0.21**



PersonaChat

	Fluency	Specificity	Sensibleness
BL	**0.73**	0.74	**0.73**
LC	0.56	0.54	**0.62**
KV	**0.61**	0.63	**0.58**
HF	**0.46**	**0.46**	**0.47**
BR	0.48	0.44	**0.43**
DR	**0.16**	0.19	**0.16**

Table 2: Per feature win-rate of the different systems over all domains. Bold numbers indicate that the feature has a significant influence on system survival according to a Cox model.

For example, for the DR model, the fluency feature is significant across all three domains, and together with its low fluency win rate, we can deduce that it is often spotted due to its low fluency.

Experiment

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems

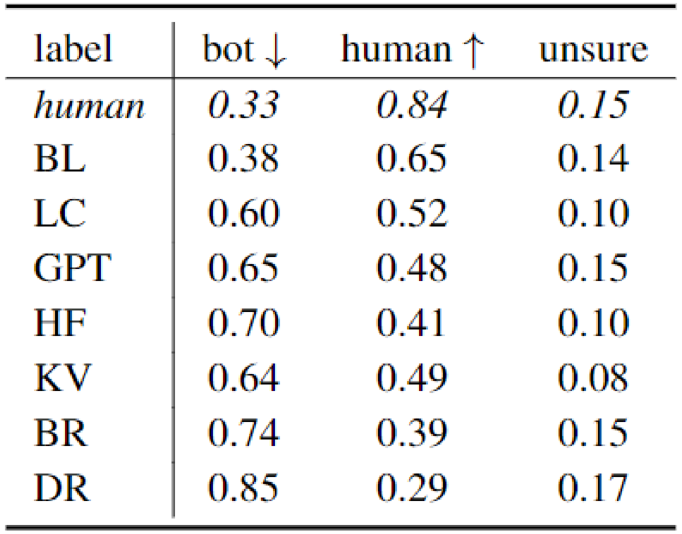
On Inter-Annotator Agreement

In our setting, annotator disagreement on a bot's human-like behavior can be interpreted as a feature of a bot's performance.

we calculate per bot and label the percentage of cases where both annotators annotate the label if one of them does.

the DR system obtains the highest agreement when being identified as a bot, and lowest when it is perceived as a human.

rank high based on win rates and in the survival analysis (BL, GPT, LC) obtain the highest agreement on the human label and lowest agreement on the bot label.



label	bot ↓	human ↑	unsure
human	*0.33*	*0.84*	*0.15*
BL	0.38	0.65	0.14
LC	0.60	0.52	0.10
GPT	0.65	0.48	0.15
HF	0.70	0.41	0.10
KV	0.64	0.49	0.08
BR	0.74	0.39	0.15
DR	0.85	0.29	0.17

Table 3: Annotator agreement on labels.

Experiment

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems

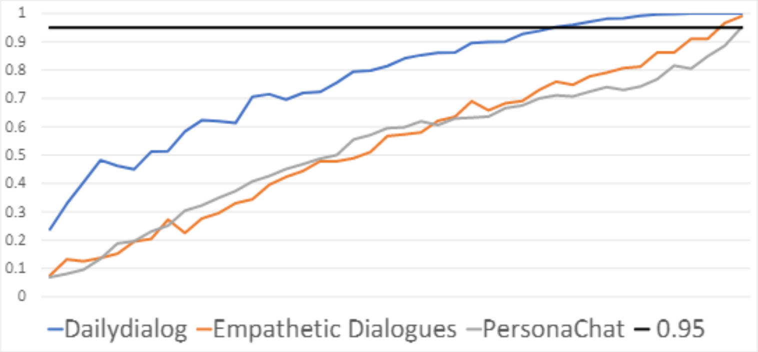
On Reliability

We measure how many pairwise conversations between two bots are needed to guarantee a stable ranking.

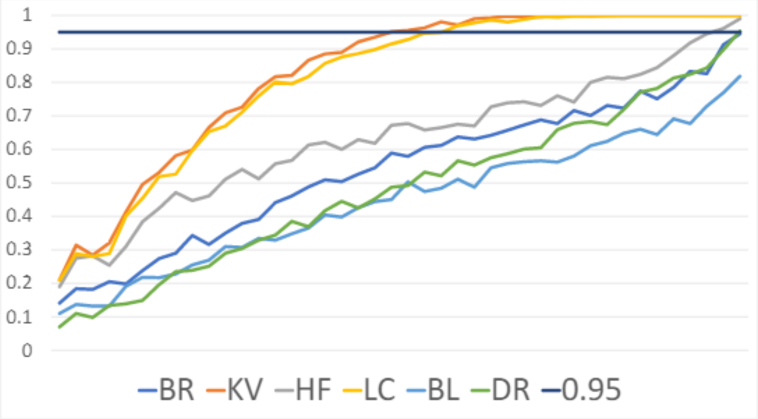
For the Dailydialog domain, 33 pairwise conversations are enough to guarantee a stable ranking. In the other two domains, this value is reached with over 40 pairwise dialogues

A more in-depth analysis reveals that ranking stability depends on the significance of pairwise comparisons.

Figure 3b shows the result of the leave-one-out stability analysis. When leaving one between LC or KV out, the stability is achieved with 25 pairwise dialogues.



(a) Stability Experiment.



(b) Leave-one-out Experiment.

Experiment

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems

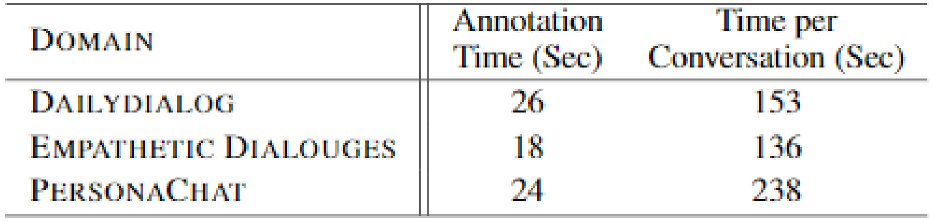
On Time Efficiency

For the Dailydialog and PersonaChat domain, the average annotation time is at around 25 seconds. For the Empathetic Dialogues, it is at 18 seconds, which is due to the shorter dialogues.

We compare this to the time to create conversations between humans and bots.

For the Dailydialog and Empathetic Dialogues domains, it takes over 2 Minutes per conversation. For PersonaChat, the time increased to almost 4 minutes.

Thus, Spot The Bot increases the annotation speed while reducing the human raters' mental strain



DOMAIN	Annotation Time (Sec)	Time per Conversation (Sec)
DAILYDIALOG	26	153
EMPATHETIC DIALOUGES	18	136
PERSONACHAT	24	238

Table 4: Overview of time efficiency in Seconds. Spot The Bot annotation versus creating human-bot conversations.

Summary

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems

- The authors propose Spot The Bot, a cost-efficient and robust evaluation framework that replaces human-bot conversations with conversations between bots.
- Human judges then only annotate for each entity in a conversation whether they think it is human or not.
- They apply the framework to three well-known domains and common baselines and state-of-the-art systems to produce a stable ranking among them. They release the framework as a ready-to-use tool for evaluating dialogue systems.

GLUCOSE: GeneraLized and COntextualized Story Explanations

Nasrin Mostafazadeh[*] **Aditya Kalyanpur** **Lori Moon** **David Buchanan**[†]
Lauren Berkowitz **Or Biran** **Jennifer Chu-Carroll**
Elemental Cognition
New York, NY, USA
nasrin@verneek.com
{adityak, lorim, orb, jenniferc}@elementalcognition.com
david.buchanan@quillbot.com

Introduction

GLUCOSE: GeneraLized and COntextualized Story Explanations

- When humans read or listen, they make implicit commonsense inferences that frame their understanding of what happened and why.

- AI systems for tasks such as reading comprehension and dialogue remain far from exhibiting similar commonsense reasoning capabilities.

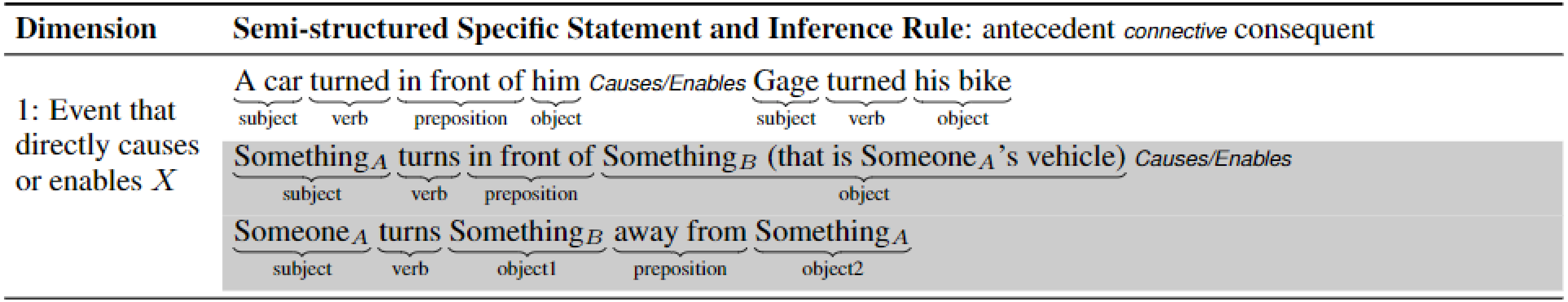
- Two major bottlenecks have been **acquiring commonsense knowledge** and successfully **incorporating it into state-of-the-art AI systems**.

Motivation

GLUCOSE: GeneraLized and COntextualized Story Explanations

- To address the first bottleneck, the authors have built an effective platform to acquire causal commonsense knowledge at scale.
- To address the second, the authors show that pre-trained neural models can start making similar inferences when trained on such rich curated data.
- The GLUCOSE (GeneraLized and COntextualized Story Explanations) dataset. Given a short story and a sentence X in the story, GLUCOSE captures ten dimensions of causal explanation related to X.

Context: Gage was riding his bike. A car turned in front of him. **Gage turned his bike sharply**. He fell off of his bike. Gage skinned his knee.



Dimension	**Semi-structured Specific Statement and Inference Rule:** antecedent *connective* consequent
1: Event that directly causes or enables X	$\underbrace{\text{A car}}_{\text{subject}}\ \underbrace{\text{turned}}_{\text{verb}}\ \underbrace{\text{in front of}}_{\text{preposition}}\ \underbrace{\text{him}}_{\text{object}}$ *Causes/Enables* $\underbrace{\text{Gage}}_{\text{subject}}\ \underbrace{\text{turned}}_{\text{verb}}\ \underbrace{\text{his bike}}_{\text{object}}$
	$\underbrace{\text{Something}_A}_{\text{subject}}\ \underbrace{\text{turns}}_{\text{verb}}\ \underbrace{\text{in front of}}_{\text{preposition}}\ \underbrace{\text{Something}_B\ \text{(that is Someone}_A\text{'s vehicle)}}_{\text{object}}$ *Causes/Enables* $\underbrace{\text{Someone}_A}_{\text{subject}}\ \underbrace{\text{turns}}_{\text{verb}}\ \underbrace{\text{Something}_B}_{\text{object1}}\ \underbrace{\text{away from}}_{\text{preposition}}\ \underbrace{\text{Something}_A}_{\text{object2}}$

Method

GLUCOSE: GeneraLized and COntextualized Story Explanations

The Knowledge Model of GLUCOSE

- Each story is explained through ten causal dimensions. The semi-structured explanation for each dimension includes both a specific statement and a general rule.
- **Causal Dimensions**: For an event or state X stated in a sentence, we categorize the dimensions of causality into events and states happening before X and those occurring after X. Each category includes five dimensions.
- **Semi-structured Inference Rules**: Each rule takes the form "antecedent connective consequent," where the antecedent and consequent are composed by filling in syntactic slots for subject, verb, object(s), and preposition(s).

Method

GLUCOSE: GeneraLized and COntextualized Story Explanations

The GLUCOSE Dataset

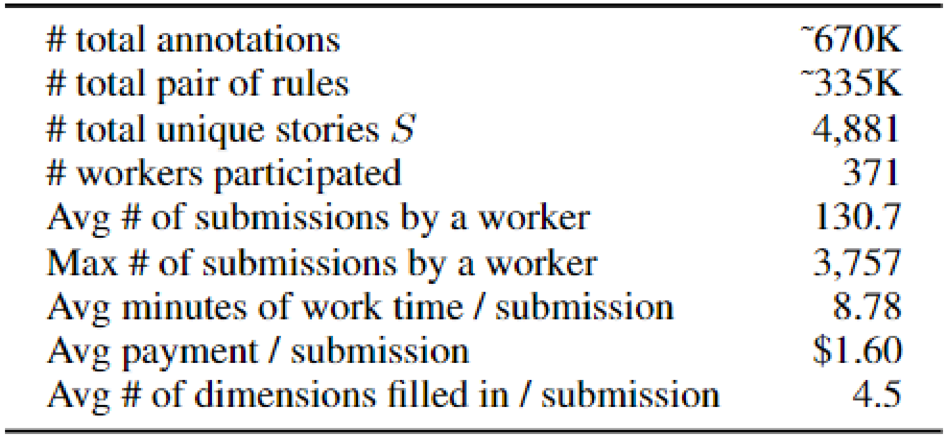
- Data Acquisition Platform. a three-stage knowledge acquisition pipeline.
- The workers first go through a qualification test where they must score at least 90% on 10 multiple-choice questions on select GLUCOSE dimensions.
- Next, qualified workers can work on the main GLUCOSE data collection task: given a story S and a story sentence X, they are asked to fill in (allowing for non-applicable) all ten GLUCOSE dimensions.
- Finally, the submissions are reviewed by an expert who rates each worker on a scale from 0 to 3, and provides feedback on how to improve.

Method

GLUCOSE: GeneraLized and COntextualized Story Explanations

Source of stories for the GLUCOSE dataset is ROCStories.

The authors compared its coverage against that of the two most relevant commonsense resources: ConceptNet and ATOMIC. They performed a best-effort mapping from GLUCOSE dimensions to relations in ConceptNet and ATOMIC.



# total annotations	˜670K
# total pair of rules	˜335K
# total unique stories S	4,881
# workers participated	371
Avg # of submissions by a worker	130.7
Max # of submissions by a worker	3,757
Avg minutes of work time / submission	8.78
Avg payment / submission	\$1.60
Avg # of dimensions filled in / submission	4.5

Table 2: Statistics about the GLUCOSE dataset.

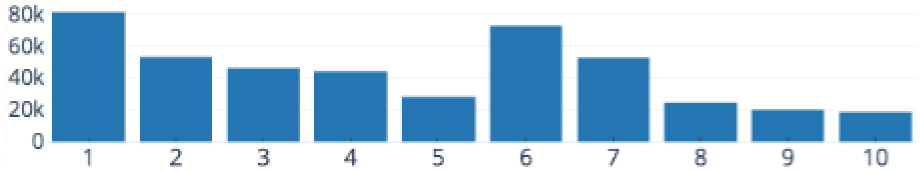
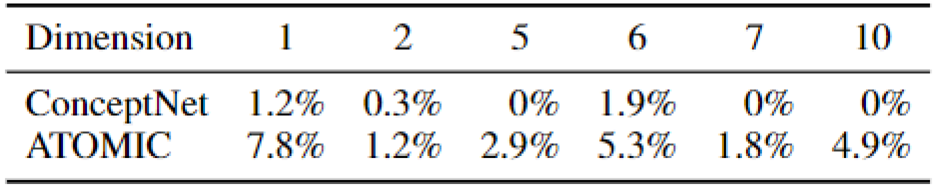


Figure 1: Number of rules collected for each dimension. Dimensions 1 and 6 have the most representation, while dimensions 9 and 10 are most often marked as not applicable.



Dimension	1	2	5	6	7	10
ConceptNet	1.2%	0.3%	0%	1.9%	0%	0%
ATOMIC	7.8%	1.2%	2.9%	5.3%	1.8%	4.9%

Table 3: Ceiling overlap between GLUCOSE and other resources. Omitted dimensions had no overlap.

Experiment

GLUCOSE: GeneraLized and COntextualized Story Explanations

Human and Automatic Evaluation

They are shown a randomly-shuffled list of candidate answers, each produced by a different system and rates each candidate answer on a four-point Likert scale.

SacreBLEU with equal weights up to 4-grams at corpus-level on the three-reference test set.

Models Trained on GLUCOSE

Pretrained Language Model (PT-LM)

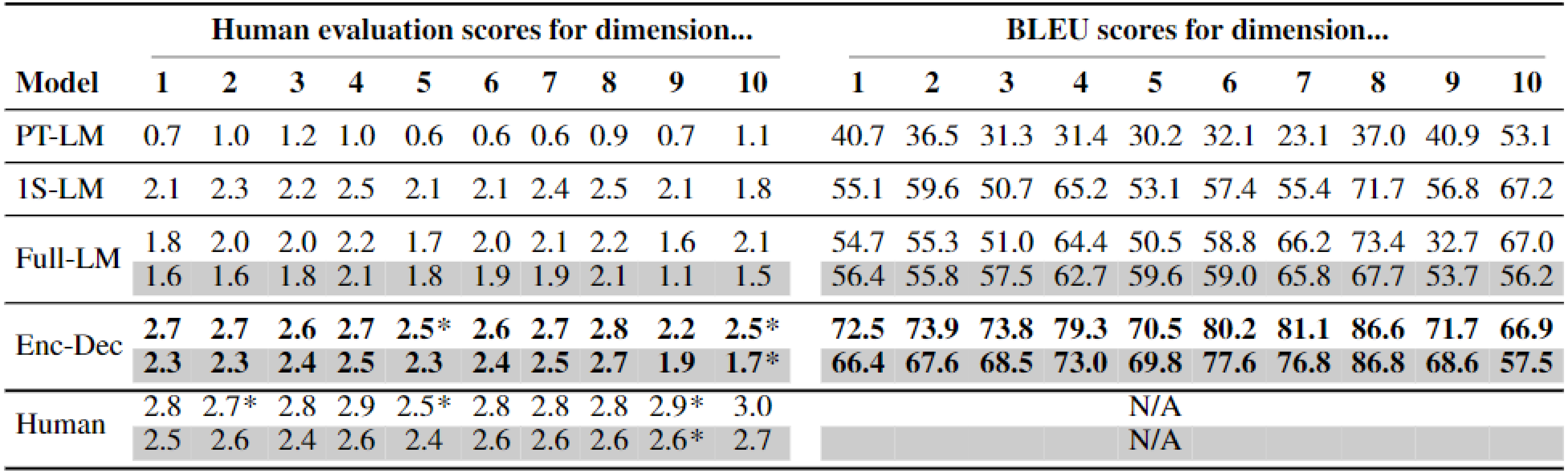
One-sided Generation (1S-LM)

Full Rule Generation (Full-LM)

Encoder-Decoder Model (Enc-Dec)

Experiment

GLUCOSE: GeneraLized and COntextualized Story Explanations

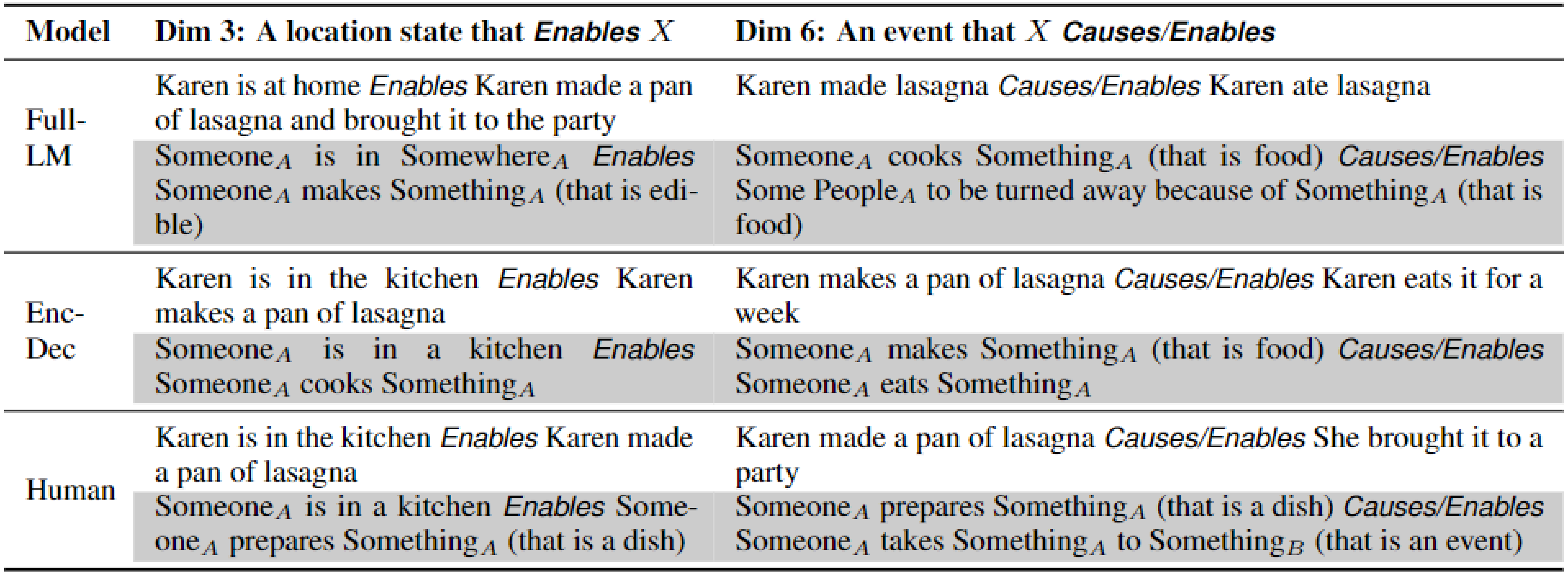


	Human evaluation scores for dimension...										BLEU scores for dimension...									
Model	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
PT-LM	0.7	1.0	1.2	1.0	0.6	0.6	0.6	0.9	0.7	1.1	40.7	36.5	31.3	31.4	30.2	32.1	23.1	37.0	40.9	53.1
1S-LM	2.1	2.3	2.2	2.5	2.1	2.1	2.4	2.5	2.1	1.8	55.1	59.6	50.7	65.2	53.1	57.4	55.4	71.7	56.8	67.2
Full-LM	1.8	2.0	2.0	2.2	1.7	2.0	2.1	2.2	1.6	2.1	54.7	55.3	51.0	64.4	50.5	58.8	66.2	73.4	32.7	67.0
	1.6	1.6	1.8	2.1	1.8	1.9	1.9	2.1	1.1	1.5	56.4	55.8	57.5	62.7	59.6	59.0	65.8	67.7	53.7	56.2
Enc-Dec	**2.7**	**2.7**	**2.6**	**2.7**	**2.5***	**2.6**	**2.7**	**2.8**	**2.2**	**2.5***	**72.5**	**73.9**	**73.8**	**79.3**	**70.5**	**80.2**	**81.1**	**86.6**	**71.7**	**66.9**
	2.3	**2.3**	**2.4**	**2.5**	**2.3**	**2.4**	**2.5**	**2.7**	**1.9**	**1.7***	**66.4**	**67.6**	**68.5**	**73.0**	**69.8**	**77.6**	**76.8**	**86.8**	**68.6**	**57.5**
Human	2.8	2.7*	2.8	2.9	2.5*	2.8	2.8	2.8	2.9*	3.0					N/A					
	2.5	2.6	2.4	2.6	2.4	2.6	2.6	2.6	2.6*	2.7					N/A					

Enc-Dec uniformly outperforms all other models, confirming that full visibility into context helps an architecture better learn the intricacies of GLUCOSE rules. Its worst performance is on general rules for dimensions 5 and 10, which have the lowest number of training points and are the most diverse in content.

Experiment

GLUCOSE: GeneraLized and COntextualized Story Explanations

- 

Model	Dim 3: A location state that *Enables* X	Dim 6: An event that X *Causes/Enables*
Full-LM	Karen is at home *Enables* Karen made a pan of lasagna and brought it to the party	Karen made lasagna *Causes/Enables* Karen ate lasagna
	Someone$_A$ is in Somewhere$_A$ *Enables* Someone$_A$ makes Something$_A$ (that is edible)	Someone$_A$ cooks Something$_A$ (that is food) *Causes/Enables* Some People$_A$ to be turned away because of Something$_A$ (that is food)
Enc-Dec	Karen is in the kitchen *Enables* Karen makes a pan of lasagna	Karen makes a pan of lasagna *Causes/Enables* Karen eats it for a week
	Someone$_A$ is in a kitchen *Enables* Someone$_A$ cooks Something$_A$	Someone$_A$ makes Something$_A$ (that is food) *Causes/Enables* Someone$_A$ eats Something$_A$
Human	Karen is in the kitchen *Enables* Karen made a pan of lasagna	Karen made a pan of lasagna *Causes/Enables* She brought it to a party
	Someone$_A$ is in a kitchen *Enables* Someone$_A$ prepares Something$_A$ (that is a dish)	Someone$_A$ prepares Something$_A$ (that is a dish) *Causes/Enables* Someone$_A$ takes Something$_A$ to Something$_B$ (that is an event)

Table 5: Example model generations for the input story: *Karen made a pan of lasagna. She brought it to the party. Nobody wanted to eat lasagna. Karen ate it for a week. She became tired of lasagna.* (Sentence X is underlined.) Note that all test stories are unseen in the train or validation set.

Summary

GLUCOSE: GeneraLized and COntextualized Story Explanations

- The authors introduce GLUCOSE, a large-scale dataset of implicit commonsense causal knowledge, encoded as causal mini-theories about the world, each grounded in a narrative context.
- They show that existing knowledge resources and pretrained language models do not include or readily predict GLUCOSE’s rich inferential content.
- when state-of-the-art neural models are trained on this knowledge, they can start to make commonsense inferences on unseen stories that match humans’ mental models.

Thanks