



Trustworthy Natural Language Generation with Communicative Goals

Dipanjan Das

AKBC 2022

Joint work with

Ankur Parikh



Chris Alberti



Daniel Andor



David Reitter



Fantine Huot



Joshua Maynez



Mirella Lapata



Elizabeth Clark



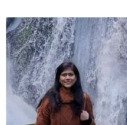
Gaurav Singh Tomar



Hannah Rashkin



Priyanka Agrawal



Reinald Kim



Shashi Narayan



Kellie Webster



Kuzman Ganchev



Livio Baldini Soares



Ran Tian



Sebastian Gehrmann



Matthew Lamm



Michael Collins



Tom Kwiatkowski



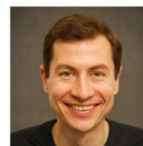
Thibault Sellam



Vitaly Nikolaev



Slav Petrov

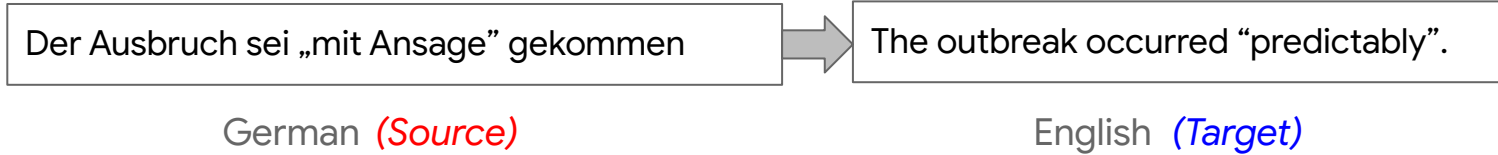


Research Goal

Develop trustworthy natural language generation models for communicative scenarios

Develop trustworthy natural language generation models for communicative scenarios

(not open ended
generation from
language models)



Develop trustworthy natural language generation models for communicative scenarios

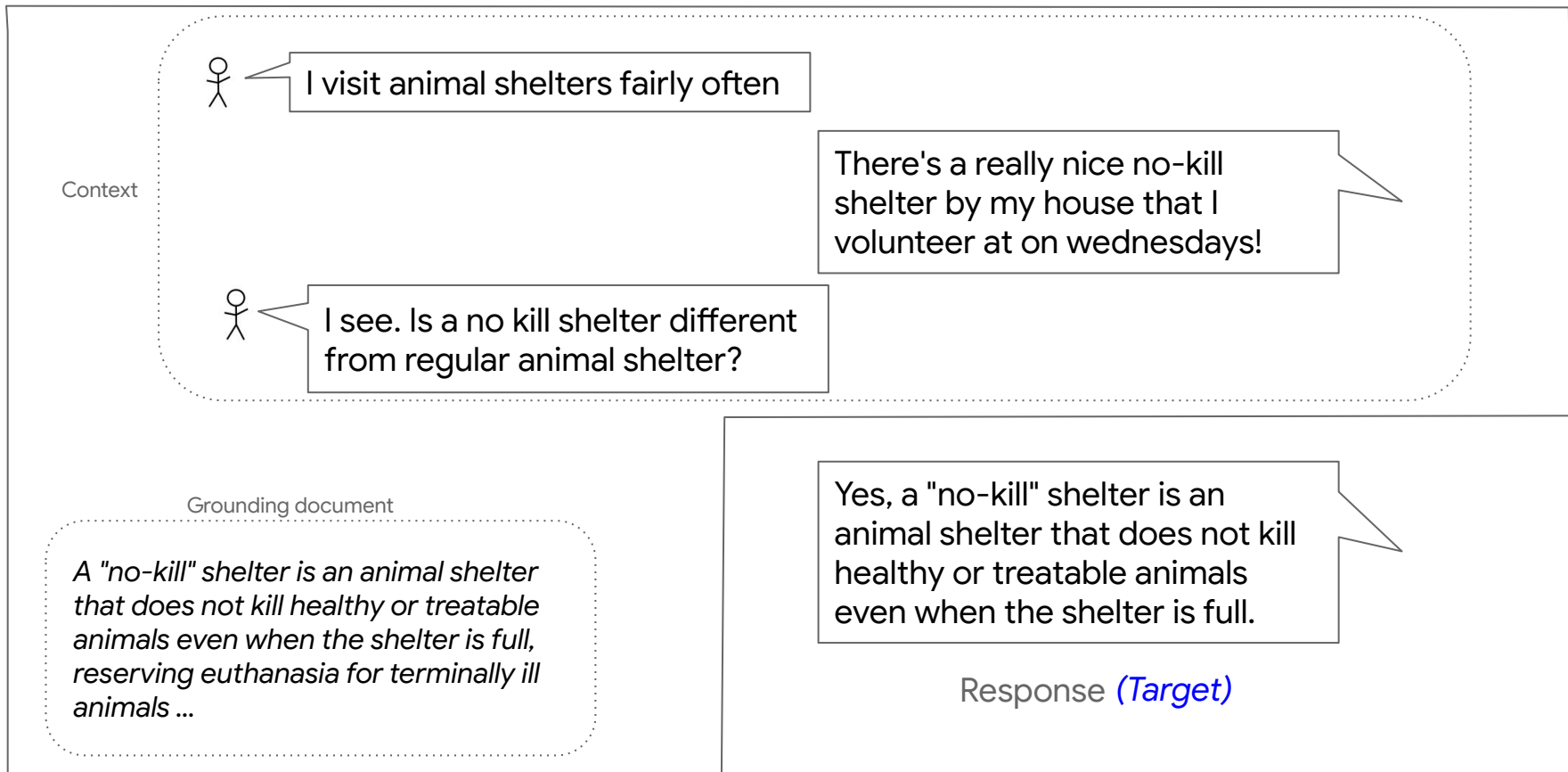
(not open ended generation from language models)

Chelsea's Eden Hazard and Arsenal's Santi Cazorla are set to reach a Premier League milestone this weekend when they each make their 100th appearance. Both players have been hugely influential since they moved to London in the summer of 2012, but who has been the most exciting import to watch? Here, Sportsmail's reporters choose the player they most enjoy seeing in action. Eden Hazard (L) and Santi Cazorla are both set to make their 100th Premier League appearance this weekend. Lee Clayton. Cazorla has wonderful balance. So does Hazard. Cazorla scores important goals. So does Hazard. Cazorla is two-footed. So is Hazard. Cazorla dances past opponents. So does Hazard. So, while there is not a lot to choose between them and Hazard is likely to get the most picks in this article, I am going for Cazorla. It's a personal choice. He is a wonderful footballer. I have paid to watch them both (and I will pay to watch them both again), but the little Spanish magician edges it for me. VERDICT: CAZORLA. Cazorla, pictured in action against Burnley, has been an influential part of Arsenal's midfield this season. Ian Ladyman. I remember when Manchester City balked at paying Hazard's wages when the Belgian was up for grabs in 2012. Back then City thought the young forward had a rather high opinion of his own worth for a player who was yet to play in a major European league. In the early days of his time at Chelsea, it looked as though City may have been right. He showed flashes of brilliance but also looked rather too easy to push off the ball. Roll forward to 2015, however, and the 24-year-old has developed in to one of the most important players in the Barclays Premier League. Brave, strong and ambitious, Hazard plays on the front foot and with only one thought in this mind. Rather like Cristiano Ronaldo, he has also developed in to the type of player ever defender hates, simply because he gets back up every time he is knocked to the ground. He would get in every team in the Premier League and is one of the reasons Chelsea will win the title this season. VERDICT: HAZARD. Hazard controls the ball under pressure from Stoke midfielder Stephen Ireland at Stamford Bridge. Dominic King. It has to be Hazard. I saw him play for Lille twice in the season before he joined Chelsea – once against St Etienne, the other was what proved to be his final appearance against Nancy. He scored two in the first match, a hat-trick the latter and played a different game to those around him. He hasn't disappointed since arriving here and I love the nonchalance with which he takes a penalty, his low centre of gravity and the way he can bamboozle defenders. If there is such a thing as £32million bargain, it is Hazard. VERDICT: HAZARD. Hazard celebrates after scoring a fine individual goal in Chelsea's 3-2 win against Hull in March. Nick Harris. Now this is a tricky one because while Eden Hazard will frequently embark on a dribble or dink in a pass that will make you nod in appreciation, he'll also miss a penalty and make you groan. Whereas the older Cazorla, less flashy but no less of a technical master, is to my mind more of a fulcrum, more important relatively to the sum of Arsenal's parts than Hazard is to Chelsea. You'll gasp at Hazard but Cazorla's wow factor is richer. That's not to dismiss either: both are brilliant footballers, contributing goals, assists and flair. Any neutral would bite your hand off to have either playing in your team. Forced to pick though, it's Cazorla, for his consistency and crucially doing it in the biggest games. Exhibit A would be Manchester City 0 Arsenal 2 in January; goal, assist, all-round brilliance, against a big team, at an important time. VERDICT: CAZORLA. Cazorla scores from the penalty spot in Arsenal's 2-0 away win at Manchester City in January. Riath Al-Samarrai. Eden Hazard for me. Cazorla is an utter delight, a little pinball of a man who is probably the most two-footed player I've seen. Put him in a tight space and then you see what makes him rare among the best. But Hazard is the top player in the Premier League, in my opinion. This is the sixth of his eight seasons as a professional where he has reached double figures and yet he offers so much more than goals (36 in 99 in the Premier League for Chelsea). He can beat a man and, better still, you sense he likes doing it. Technically, his passing and shooting are excellent and he also has a mind capable of sussing out the shapes and systems in front of him. That intelligence, more specifically.

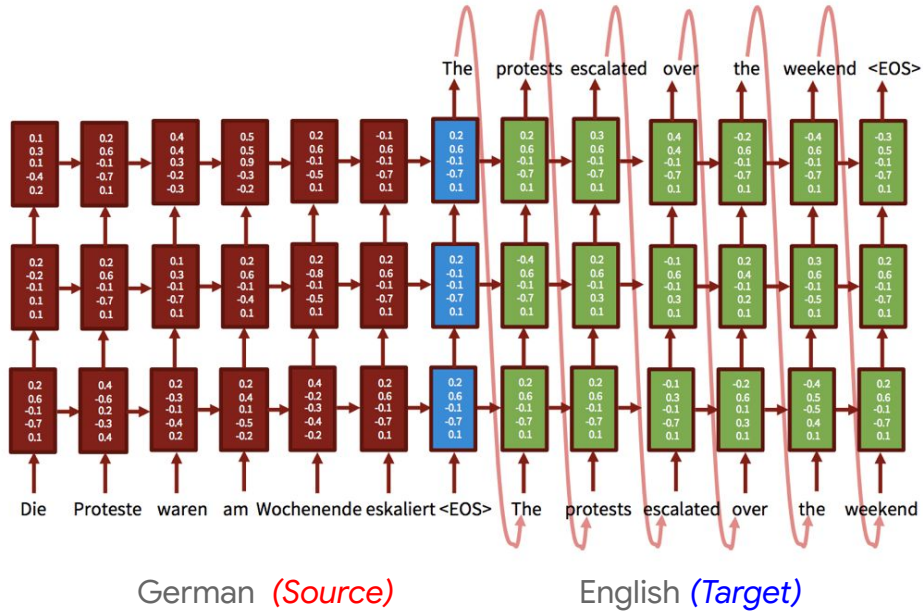
Chelsea star Eden Hazard is set to make his 100th top-flight appearance. Santi Cazorla should hit the same milestone when Arsenal meet Burnley. Both players have impressed since moving to the Premier League in 2012. Hazard has more goals this season but Cazorla has one more assist. Sports-mail's reporters choose the player who has excited them the most.

Summary (Target)

Develop trustworthy natural language generation models for communicative scenarios



Develop trustworthy natural language generation models for communicative scenarios



Sequence to sequence models
(Sutskever et al. 2014, Bahdanau et al. 2014)

Background:

- Popular modeling framework for conditional generation models
- Pre-trained models (e.g. BERT, PEGASUS, T5, GPT) have resulted in further progress
- Advantages
 - End-to-end training and inference
 - Fluent target output
 - Pre-training+fine tuning paradigm is attractive

Develop trustworthy natural language generation models for communicative scenarios

Chelsea's Eden Hazard and Arsenal's Santi Cazorla are set to reach a Premier League milestone this weekend when they each make their 100th appearance. Both players have been hugely influential since they moved to London in the summer of 2012, but who has been the most exciting import to watch? Here, Sportsmail's reporters choose the player they most enjoy seeing in action. Eden Hazard (L) and Santi Cazorla are both set to make their 100th Premier League appearance this weekend. Lee Clayton. Cazorla has wonderful balance. So does Hazard. Cazorla scores important goals. So does Hazard. Cazorla is two-footed. So is Hazard. Cazorla dances past opponents. So does Hazard. So, while there is not a lot to choose between them and Hazard is likely to get the most picks in this article, I am going for Cazorla. It's a personal choice. He is a wonderful footballer. I have paid to watch them both (and I will pay to watch them both again), but the little Spanish magician edges it for me. VERDICT: CAZORLA. Cazorla, pictured in action against Burnley, has been an influential part of Arsenal's midfield this season. Ian Ladyman. I remember when Manchester City balked at paying Hazard's wages when the Belgian was up for grabs in 2012. Back then City thought the young forward had a rather high opinion of his own worth for a player who was yet to play in a major European league. In the early days of his time at Chelsea, it looked as though City may have been right. He showed flashes of brilliance but also looked rather too easy to push off the ball. Roll forward to 2015, however, and the 24-year-old has developed in to one of the most important players in the Barclays Premier League. Brave, strong and ambitious, Hazard plays on the front foot and with only one thought in this mind. Rather like Cristiano Ronaldo, he has also developed in to the type of player ever defender hates, simply because he gets back up every time he is knocked to the ground. He would get in every team in the Premier League and is one of the reasons Chelsea will win the title this season. VERDICT: HAZARD. Hazard controls the ball under pressure from Stoke midfielder Stephen Ireland at Stamford Bridge. Dominic King. It has to be Hazard. I saw him play for Lille twice in the season before he joined Chelsea – once against St Etienne, the other was what proved to be his final appearance against Nancy. He scored two in the first match, a hat-trick the latter and played a different game to those around him. He hasn't disappointed since arriving here and I love the nonchalance with which he takes a penalty, his low centre of gravity and the way he can bamboozle defenders. If there is such a thing as £32million bargain, it is Hazard. VERDICT: HAZARD. Hazard celebrates after scoring a fine individual goal in Chelsea's 3-2 win against Hull in March. Nick Harris. Now this is a tricky one because while Eden Hazard will frequently embark on a dribble or dink in a pass that will make you nod in appreciation, he'll also miss a penalty and make you groan. Whereas the older Cazorla, less flashy but no less of a technical master, is to my mind more of a fulcrum, more important relatively to the sum of Arsenal's parts than Hazard is to Chelsea. You'll gasp at Hazard but Cazorla's wow factor is richer. That's not to dismiss either: both are brilliant footballers, contributing goals, assists and flair. Any neutral would bite your hand off to have either playing in your team. Forced to pick though, it's Cazorla, for his consistency and crucially doing it in the biggest games. Exhibit A would be Manchester City 0 Arsenal 2 in January; goal, assist, all-round brilliance, against a big team, at an important time. VERDICT: CAZORLA. Cazorla scores from the penalty spot in Arsenal's 2-0 away win at Manchester City in January. Riath Al-Samarrai. Eden Hazard for me. Cazorla is an utter delight, a little pinball of a man who is probably the most two-footed player I've seen. Put him in a tight space and then you see what makes him rare among the best. But Hazard is the top player in the Premier League, in my opinion. This is the sixth of his eight seasons as a professional where he has reached double figures and yet he offers so much more than goals (36 in 99 in the Premier League for Chelsea). He can beat a man and, better still, you sense he likes doing it. Technically, his passing and shooting are excellent and he also has a mind capable of sussing out the shapes and systems in front of him. That intelligence, more specifically.

Hallucinations in
state-of-the-art sequence to
sequence models (PEGASUS):

Eden Hazard and Santi Cazorlag will each make their 100th Premier League appearance this weekend. **nightstandapplication.com**. Sportsmail's **hovercraft** reporters choose their **man of the match** **countermeasures**.

Summary (Target)

 = hallucinations, not attributable to source

Develop trustworthy natural language generation models for communicative scenarios

Frank Lino

FBI surveillance photo

Birth date October 30, 1938

Birth place Gravesend, Brooklyn, New York,
United States


https://en.wikipedia.org/wiki/Frank_Lino

Wikipedia Infobox (*Source*)

Hallucinations in state-of-the-art
sequence to sequence models
(pointer-generator):

Frank Lino (born October 30, 1938 in Brooklyn, New York, United States) is an American **criminal defense attorney**.

Biography sentence (*Target*)

 = hallucination, not attributable to source

Develop trustworthy natural language generation models for communicative scenarios

Evaluation and Benchmarking

Handling Divergent Reference Texts when Evaluating Table-to-Text Generation (Dhingra et al., ACL 2019)

ToTTo: A Controlled Table-to-Text Generation Dataset (Parikh et al., EMNLP 2020)

BLEURT: Learning Robust Metrics for Text Generation (Sellam et al., ACL 2020)

Learning to Evaluate Translation Beyond English: BLEURT Submissions to the WMT Metrics 2020 Shared Task (Pu et al., WMT 2020)

Learning Compact Metrics for MT (Pu et al., EMNLP 2021)

The GEM Benchmark: Natural Language Generation, Evaluation and Metrics (Gehrmann et al., ACL 2021 Workshop, *living benchmark*)

Measuring Attribution in Natural Language Generation Models (Rashkin et al., 2022, working paper)

Controllable Models

Text Generation with Exemplar-based Adaptive Decoding (Peng et al., NAACL 2019)

Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation (Tian et al., 2019, arXiv)

Increasing Faithfulness in Knowledge-Grounded Dialogue with Controllable Features (Rashkin, et al., ACL 2021)

Planning with Learned Entity Prompts for Abstractive Summarization (Narayan et al., TACL 2022)

A Well-Composed Text is Half Done! Composition Sampling for Diverse Conditional Generation (Narayan et al., ACL 2022)

Conditional Generation with a Question-Answering Blueprint (Narayan et al., 2022, working paper)

Evaluation and Benchmarking

ToTTo: A Data-to-Text Generation Benchmark

120K training examples

Table Title: Robert **Craig** (American football)

Section Title: **National Football League** statistics

Table Description: None

YEAR	TEAM	Rushing					Receiving				
		ATT	YDS	AVG	LNG	TD	NO.	YDS	AVG	LNG	TD
1983	SF	176	725	4.1	71	8	48	427	8.9	23	4
1984	SF	155	649	4.2	28	4	71	675	9.5	64	3
1985	SF	214	1,050	4.9	62	9	92	1,016	11.0	73	6
1986	SF	204	830	4.1	25	7	81	624	7.7	48	0
1987	SF	215	815	3.8	25	3	66	492	7.5	35	1
1988	SF	310	1,502	4.8	46	9	76	534	7.0	22	1
1989	SF	271	1,054	3.9	27	6	49	473	9.7	44	1
1990	SF	141	439	3.1	26	1	25	201	8.0	31	0
1991	RAI	162	590	3.6	15	1	17	136	8.0	20	0
1992	MIN	105	416	4.0	21	4	22	164	7.5	22	0
1993	MIN	38	119	3.1	11	1	19	169	8.9	31	1
Totals	—	1,991	8,189	4.1	71	56	566	4,911	8.7	73	17

Craig finished his eleven **NFL** seasons with 8,189 rushing yards and 566 receptions for 4,911 receiving yards.

One sentence description (*Target*)

Table, metadata, set of highlighted cells (*Source*)

ToTTo: A Data-to-Text Generation Benchmark

120K training examples

Table Title: Robert Craig (American football)

Section Title: National Football League statistics

Table Description:None

YEAR	Rushing						Receiving				
	TEAM	ATT	YDS	AVG	LNG	TD	NO.	YDS	AVG	LNG	TD
1983	SF	176	725	4.1	71	8	48	427	8.9	23	4
1984	SF	155	649	4.2	28	4	71	675	9.5	64	3
1985	SF	214	1,050	4.9	62	9	92	1,016	11.0	73	6
1986	SF	204	830	4.1	25	7	81	624	7.7	48	0
1987	SF	215	815	3.8	25	3	66	492	7.5	35	1
1988	SF	310	1,502	4.8	46	9	76	534	7.0	22	1
1989	SF	271	1,054	3.9	27	6	49	473	9.7	44	1
1990	SF	141	439	3.1	26	1	25	201	8.0	31	0
1991	RAI	162	590	3.6	15	1	17	136	8.0	20	0
1992	MIN	105	416	4.0	21	4	22	164	7.5	22	0
1993	MIN	38	119	3.1	11	1	19	169	8.9	31	1
Totals	—	1,991	8,189	4.1	71	56	566	4,911	8.7	73	17

Craig finished his **eleven** NFL seasons with 8,189 rushing yards and 566 receptions for 4,911 receiving yards.

One sentence description (*Target*)

Table, metadata, set of highlighted cells (*Source*)

ToTTo: A Data-to-Text Generation Benchmark

120K training examples

Table Title: Robert Craig (American football)

Section Title: National Football League statistics

Table Description:None

YEAR	TEAM	Rushing					Receiving				
		ATT	YDS	AVG	LNG	TD	NO.	YDS	AVG	LNG	TD
1983	SF	176	725	4.1	71	8	48	427	8.9	23	4
1984	SF	155	649	4.2	28	4	71	675	9.5	64	3
1985	SF	214	1,050	4.9	62	9	92	1,016	11.0	73	6
1986	SF	204	830	4.1	25	7	81	624	7.7	48	0
1987	SF	215	815	3.8	25	3	66	492	7.5	35	1
1988	SF	310	1,502	4.8	46	9	76	534	7.0	22	1
1989	SF	271	1,054	3.9	27	6	49	473	9.7	44	1
1990	SF	141	439	3.1	26	1	25	201	8.0	31	0
1991	RAI	162	590	3.6	15	1	17	136	8.0	20	0
1992	MIN	105	416	4.0	21	4	22	164	7.5	22	0
1993	MIN	38	119	3.1	11	1	19	169	8.9	31	1
Totals	—	1,991	8,189	4.1	71	56	566	4,911	8.7	73	17

Craig finished his eleven NFL seasons with 8,189 rushing yards and 566 receptions for 4,911 receiving yards.

One sentence description (Target)

Table, metadata, set of highlighted cells (Source)

ToTTo: A Data-to-Text Generation Benchmark

120K training examples

Table Title: Robert Craig (American football)

Section Title: National Football League statistics

Table Description:None

YEAR	TEAM	Rushing						Receiving			
		ATT	YDS	AVG	LNG	TD	NO.	YDS	AVG	LNG	TD
1983	SF	176	725	4.1	71	8	48	427	8.9	23	4
1984	SF	155	649	4.2	28	4	71	675	9.5	64	3
1985	SF	214	1,050	4.9	62	9	92	1,016	11.0	73	6
1986	SF	204	830	4.1	25	7	81	624	7.7	48	0
1987	SF	215	815	3.8	25	3	66	492	7.5	35	1
1988	SF	310	1,502	4.8	46	9	76	534	7.0	22	1
1989	SF	271	1,054	3.9	27	6	49	473	9.7	44	1
1990	SF	141	439	3.1	26	1	25	201	8.0	31	0
1991	RAI	162	590	3.6	15	1	17	136	8.0	20	0
1992	MIN	105	416	4.0	21	4	22	164	7.5	22	0
1993	MIN	38	119	3.1	11	1	19	169	8.9	31	1
Totals	—	1,991	8,189	4.1	71	56	566	4,911	8.7	73	17

Craig finished his eleven NFL seasons with 8,189 rushing yards and 566 receptions for 4,911 receiving yards.

One sentence description (Target)

Table, metadata, set of highlighted cells (Source)

Annotation Process

Annotators do the following:

- Highlight cells that support sentence
- Iteratively revise the sentence so that it is faithful to the table and standalone

Table Title: Gabriele Becker
Section Title: International competitions

Year	Competition	Venue	Position	Event	Notes
Representing Germany					
1992	World Junior Championships	Seoul, South Korea	10th (semis)	100 m	11.83
1993	European Junior Championships	San Sebastián, Spain	7th	100 m	11.74
			3rd	4×100 m relay	44.60
1994	World Junior Championships	Lisbon, Portugal	12th (semis)	100 m	11.66 (wind: +1.3 m/s)
			2nd	4×100 m relay	44.78
1995	World Championships	Gothenburg, Sweden	7th (q-finals)	100 m	11.54
			3rd	4×100 m relay	43.01

Original Sentence

After winning the German under-23 100 m title, she was selected to run at the 1995 World Championships in Athletics both individually and in the relay.

After Deletion

~~After winning the German under-23 100 m title, she was selected to run at the 1995 World Championships in Athletics both individually and in the relay.~~

After

Decontextualization

Gabriele Becker competed at the 1995 World Championships in both individually and in the relay.

After Grammar

Gabriele Becker competed at the 1995 World Championships in both individually and in the relay.

ToTTo: A Data-to-Text Generation Benchmark

Novelty:

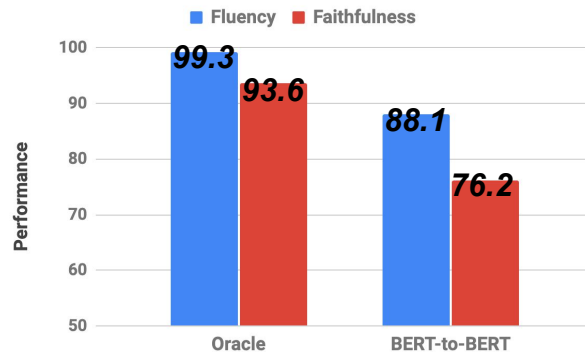
- Task Design: “Controlled generation”: Set of highlighted cells gives guidance as to what to generate.
 - Previous datasets such as Rotowire (Wiseman et al.) and Wikibio (Lebret et al.) contained significant noise in targets.
- Annotation process: Annotators iteratively revise natural sentences on Wikipedia so they are attributable to the table.
- Large dataset:
 - 120K training examples
 - 7500 dev examples
 - 7500 test examples

ToTTo: A Data-to-Text Generation Benchmark

Leaderboard

Model	Link	Uses Wiki	Overall		
			BLEU	PARENT	BLEURT
SKY	in preparation	yes	49.9	59.8	0.212
CoNT	[An et al., 2022]	yes	49.1	58.9	0.238
Supervised+NLPO	[Ramamurthy et al. 2022]	yes	47.4	59.6	0.192
Anonymous 3	in preparation	yes	49.3	58.8	0.235
ProEdit	Paper in preparation	yes	48.6	59.18	0.202
Anonymous 2	Paper in preparation	yes	49.4	59.0	0.253
PlanGen (University of Cambridge, Apple)	[Su et al. 2021]	yes	49.2	58.7	0.249
T5-based (Google)	[Kale, 2020]	yes	49.5	58.4	0.230
BERT-to-BERT (Wiki+Books)	[Rothe et al., 2019]	yes	44.0	52.6	0.121

Human evaluation

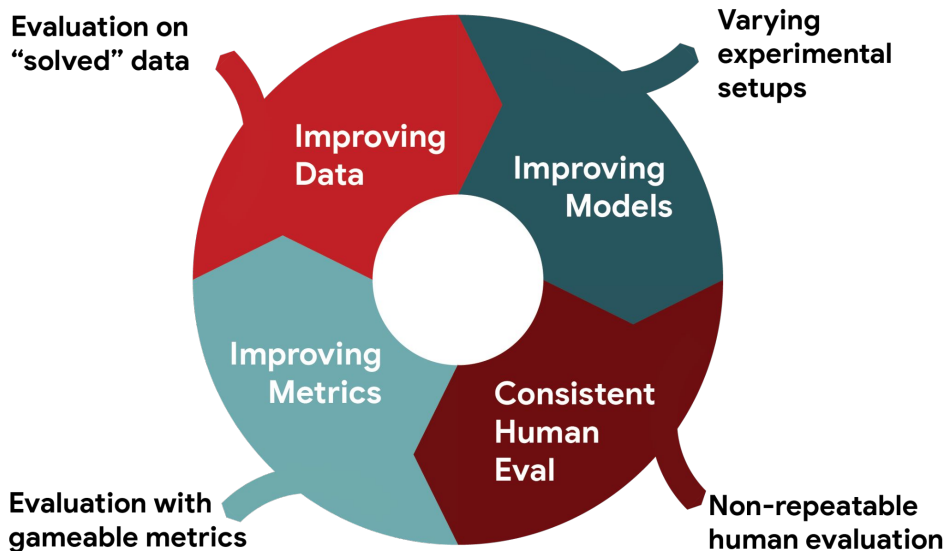


The GEM Benchmark

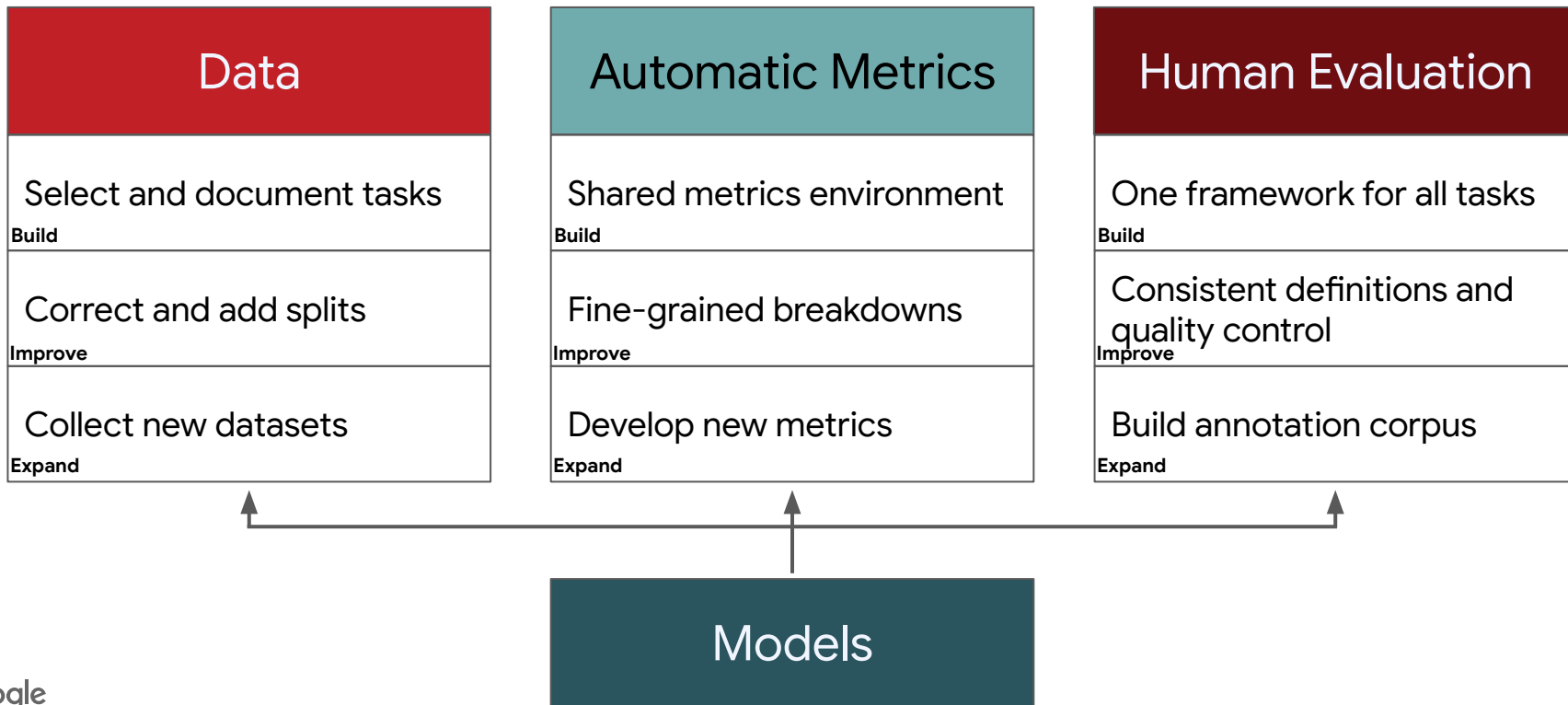
NLG Evaluation is

- ... challenging
- ... fragmented
- ... constantly evolving

As a result, we can't identify whether and how our models **fail**, or whether failure is **attributable** to the data, model, or metrics.



The GEM Benchmark



The GEM Benchmark

Set of tasks (after lengthy deliberation)

Dialog, Summarization, Simplification,
Surface Realization

18 languages

Sizes from 5k to 500k

Various Input Formats

Dataset	Communicative Goal	Language(s)	Size	Input Type
CommonGEN (Lin et al., 2020)	Produce a likely sentence which mentions all of the source concepts.	en	67k	Concept Set
Czech Restaurant (Dušek and Jurčiček, 2019)	Produce a text expressing the given intent and covering the specified attributes.	cs	5k	Meaning Representation
DART (Radev et al., 2020)	Describe cells in a table, covering all information provided in triples.	en	82k	Triple Set
E2E clean (Novikova et al., 2017) (Dušek et al., 2019)	Describe a restaurant, given all and only the attributes specified on the input.	en	42k	Meaning Representation
MLSum (Scialom et al., 2020)	Summarize relevant points within a news article	*de/es	*520k	Articles
Schema-Guided Dialog (Rastogi et al., 2020)	Provide the surface realization for a virtual assistant	en	*165k	Dialog Act
ToTTo (Parikh et al., 2020)	Produce an English sentence that describes the highlighted cells in the context of the given table.	en	136k	Highlighted Table
XSum (Narayan et al., 2018)	Highlight relevant points in a news article	en	*25k	Articles
WebNLG (Gardent et al., 2017)	Produce a text that verbalises the input triples in a grammatical and natural way.	en/ru	50k	RDF triple
WikiAuto + Turk/ASSET (Jiang et al., 2020) (Alva-Manchego et al., 2020)	Communicate the same information as the source sentence using simpler words and grammar.	en	594k	Sentence
WikiLingua (Ladhak et al., 2020)	Produce high quality summaries of an instructional article.	*ar/cs/de/en es/fr/hi/id/it ja/ko/nl/pt/ru th/tr/vi/zh	*550k	Article

The GEM Benchmark: Evaluation

What should our results tell us about a model?

✗ System Foo performs the best.

✓ System Foo leads to **consistent performance** increases in **Bar-type metrics** on challenges **that measure Baz** while maintaining equal performance on most metrics of **type Qux**.

Multiple Experiments

Specific Claims

Acknowledge Limitations

Multiple Metrics

Multiple Metrics

The GEM Benchmark: Results

There is no clear best model.

Winner

English Tasks

T5-Base	1
T5-XL	1
mT5-Base	2
mT5-Large	1
mT5-XL	2

Winner

Non-English Tasks

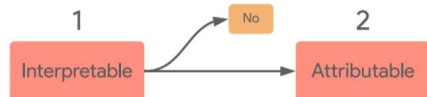
T5-XL	2
mT5-Base	1
mT5-XL	5

Measuring Attribution across Generation Models

Goal: assess when an utterance generated by a system is **attributable to identified sources (AIS)**

Dimensions: Interpretability, AIS (attributable to identified sources)

Task design: Cascading assessment: one item is evaluated for all dimensions by the same rater in the same session; negative responses filter out the item from the subsequent dimensions



Task	IAA	
	PA	α
QReCC	.89	.83
WoW	.75	.75
CNN/DM	.83	.56

Inter-annotator agreement

Model	Flag	Int	AIS
WoW Baseline (Dinan et al., 2019)	4.0	84.4*	19.8*
Dodeca (Shuster et al., 2020)	8.5	100.0	60.1*
T5 (Raffel et al., 2020b)	5.5	98.4	39.8*
T5 w/ ctrls (Rashkin et al., 2021)	7.5	99.5	92.4
<i>Reference</i>	4.0	100.0	15.6*

Evaluations show that grounded generation systems perform better on AIS on conversational question answering

Query: when did joe nieuwendyk play with the dallas stars?

System: in 1995

Query: did joe nieuwendyk start his career with the dallas stars?

System: joe nieuwendyk was a second round selection of the calgary flames.

Passage: joe nieuwendyk - wikipedia centralnotice joe nieuwendyk from wikipedia, the free encyclopedia jump to navigation jump to search joe nieuwendyk hockey hall of fame , 2011 nieuwendyk at the 2011 heritage classic alumni game born (1966-09-10) september 10, 1966 (age 52) oshawa , ontario , canada height 6 ft 2 in (188 cm) weight 195 lb (88 kg; 13 st 13 lb) position centre shot left played for calgary flames dallas stars new jersey devils toronto maple leaves florida panthers national team canada nhl draft 27th overall, 1985 calgary flames playing career 1987–2007 joseph "joe" nieuwendyk (born september 10, 1966) is a canadian former national hockey league (nhl) player. he was a second round selection of the calgary flames , 27th overall, at the 1985 nhl entry draft and played 20 seasons for the flames, dallas stars , new jersey devils , toronto maple leaves , and florida panthers . he is one of only 11 players in nhl history to win the stanley cup with three or more different teams, winning titles with calgary in 1989, dallas in 1999 and new jersey in 2003. a two-time olympian , nieuwendyk won a gold medal with team canada at the 2002 winter games . he was inducted into the hockey hall of fame in 2011 and his uniform number 25 was honoured by the flames in 2014. joe nieuwendyk was inducted into the ontario sports hall of fame in 2014. in 2017 nieuwendyk was named one of the ' 100 greatest nhl players ' in history.

Automatic Metrics for Natural Language Generation

Metrics are a Bottleneck to Progress: low correlation with human judgments for metrics such as BLEU, ROUGE, METEOR

Prediction

pete kmetovic from stanford university took third place at stanford university.

the 1956 grand prix motorcycle racing season consisted of eight grand prix races in six classes: 500cc, 350cc, 250cc, 125cc and sidecars 500cc.

kelce finished the 2012 season with 45 receptions for 722 yards (16.0 average) and eight touchdowns.

Reference

the eagles first pick, and third overall, was pete kmetovic, a halfback from stanford university.

the 1956 grand prix motorcycle racing season consisted of six grand prix races in five classes: 500cc, 350cc, 250cc, 125cc and sidecars 500cc.

in travis kelce's last collegiate season, he set personal career highs in receptions (45), receiving yards (722), yards per receptions (16.0) and receiving touchdowns (8).

Comments

incorrect and word overlap low

incorrect but word overlap high

Correct, but reference is more informative

Proposal: Learnt Metrics (for every NLG use case)

- Task-specific success criteria (usually evaluated by human annotators) can only be captured by learned metrics.
- Naive learnt metric: Fine-Tune BERT on human ratings data.

BERT [Devlin et al. 2018]



Human Ratings Data

—	—	0.1
—	—	0.7
...		
—	—	0.4

+

=

**Learnt
Metric**

- **Problem:** Brittle, requires lots of fine-tuning data for every new dataset/task.

Proposal: Learnt Metrics (for every NLG use case)

BLEURT

- Additional pretraining step based on synthetic data.
- Makes model robust to train/test skew and enables fast adaptation to other domains.

BERT [Devlin et al. 2018]



+

Synthetic Data

— —	0.2
— —	0.9
...	
— —	0.1

+

Human Ratings Data

— —	0.1
— —	0.7
...	
— —	0.4

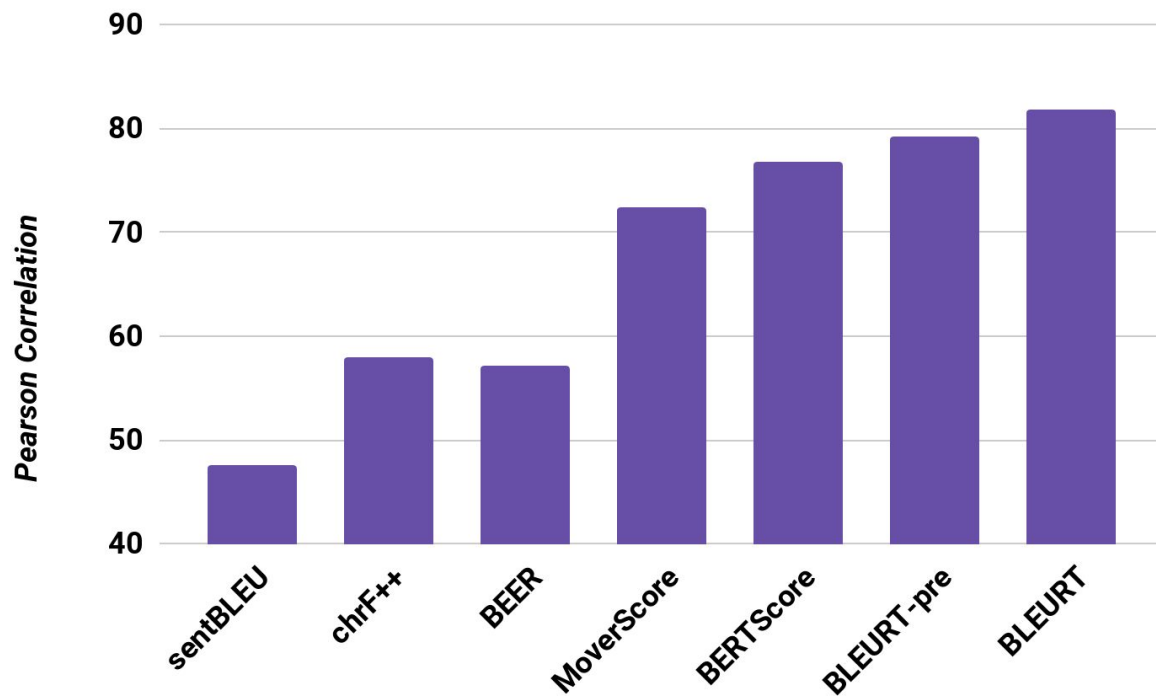
=

BLEURT

- State of the art results for Machine Translation: WMT 2017, 2018, 2019 and structure-to-text task: WebNLG

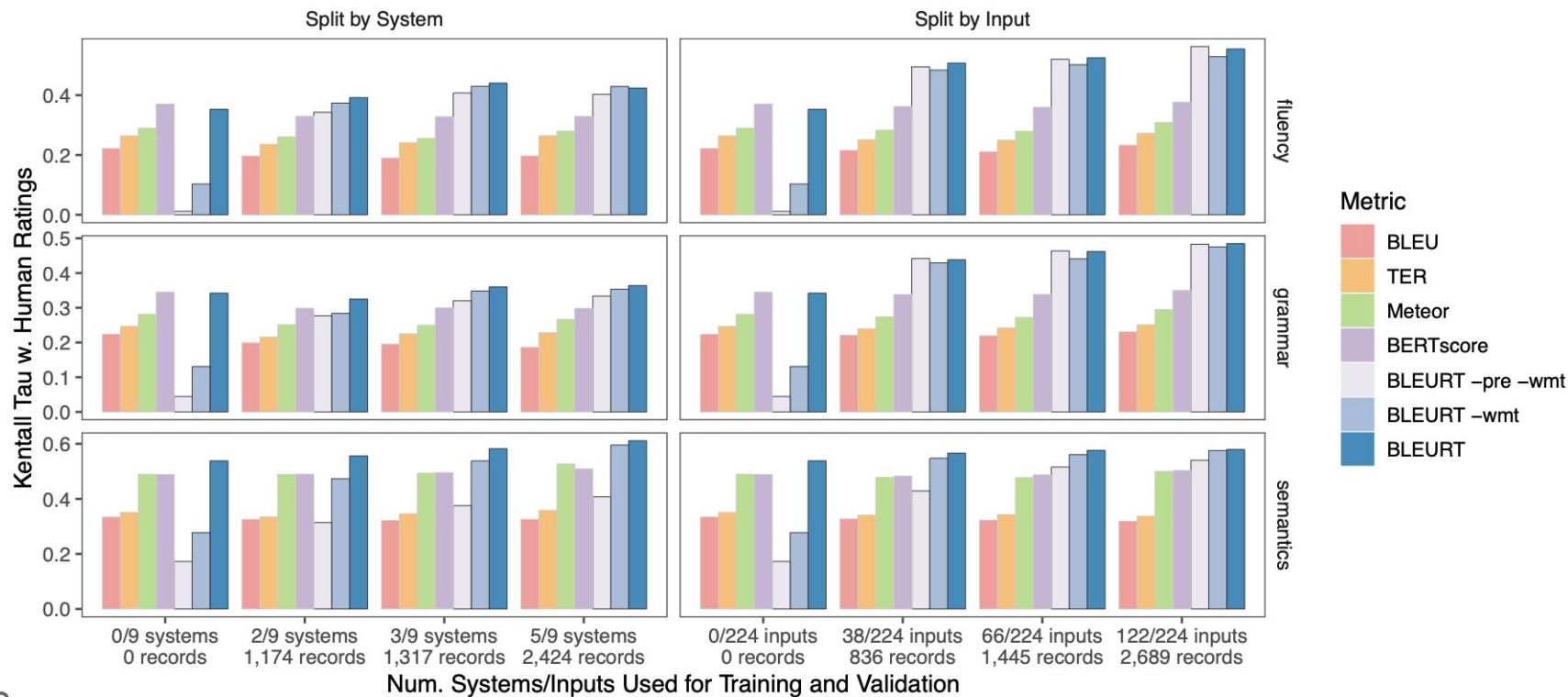
Learnt Metrics (machine translation)

BLEURT Results - WMT2017



Learnt Metrics (structure-to-text generation)

WebNLG Results



Learnt Metrics: Outlook

Two directions:

- Develop robust learnt metrics for a wide variety of generation tasks
 - Recent work for detection of factual consistency ([Honovich et al., 2022](#))
- Integrate learnt metrics to improve learning and inference:
 - Reinforcement learning towards BLEURT during training MT models ([Shu et al., 2021](#))
 - Minimum-Bayes-Risk Decoding via BLEURT ([Freitag et al., 2021](#))

Controllable Models

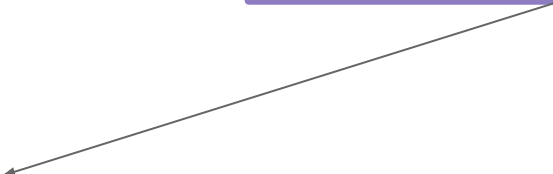
```
graph TD; A[Controllable Models] --> B[Controllable Response Generation for Knowledge Grounded Dialogue (Rashkin et al., 2021)]; A --> C[Controllable Entity-Based Planning for Summarization (Narayan et al., 2021)]; A --> D[Conditional Generation with a Question-Answering Blueprint (Narayan et al., 2022)];
```

Controllable Response
Generation for Knowledge
Grounded Dialogue
(Rashkin et al., 2021)

Controllable Entity-Based
Planning for Summarization
(Narayan et al., 2021)

Conditional Generation with
a Question-Answering
Blueprint
(Narayan et al., 2022)

Controllable Models



Controllable Response
Generation for Knowledge
Grounded Dialogue
(Rashkin et al., 2021)

Response Generation in Knowledge-Grounded Dialogue

Informative Dialogue Agents

Helping a user learn more about a topic



What foods can I feed my dog?

Is supported by documents

Carrots and apples are safe.

Not supported by documents

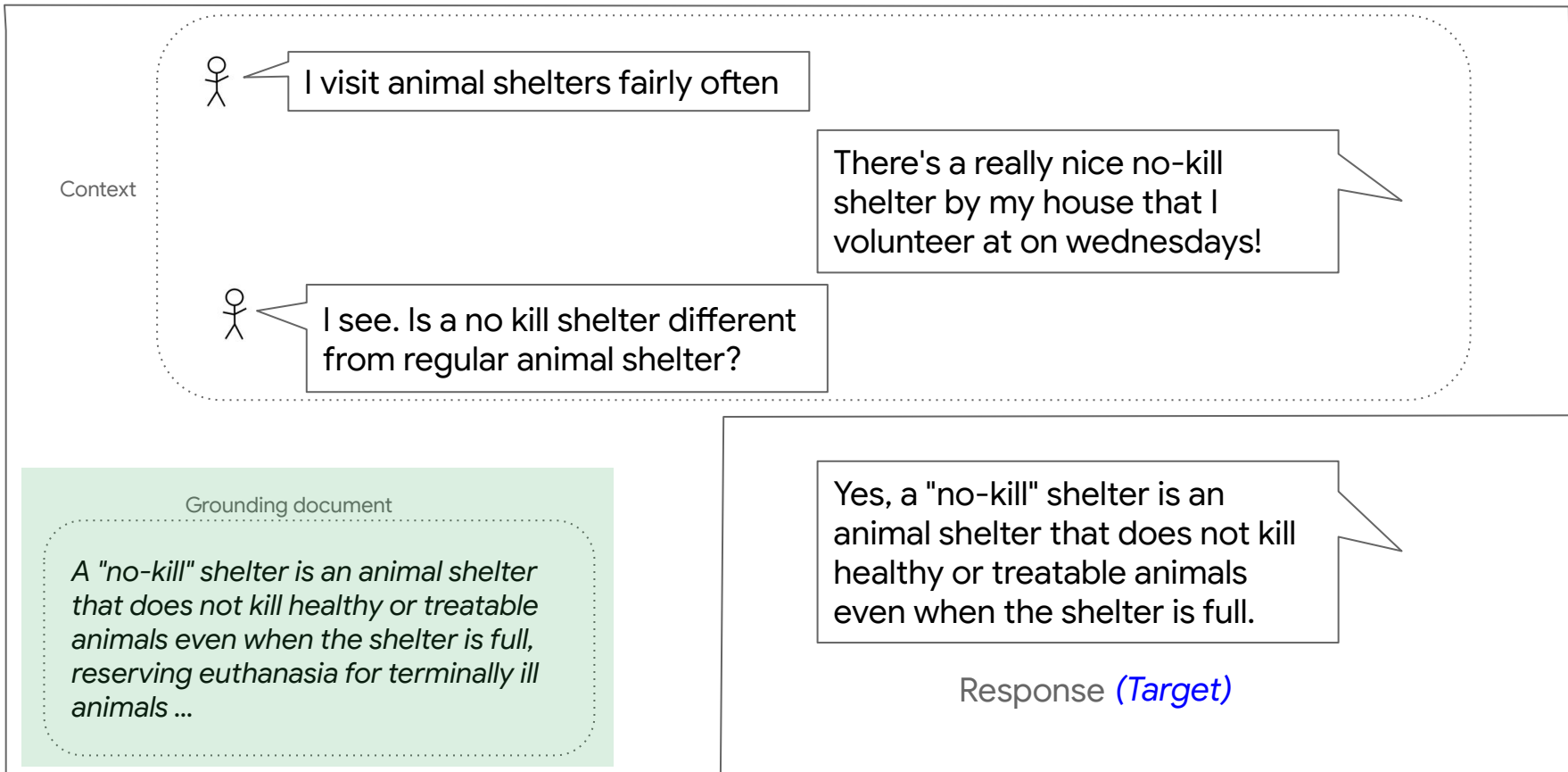
Most love chocolate.

Subjective information

I eat apples with my dog everyday.

LMs can hallucinate this type of information

Response Generation in Knowledge-Grounded Dialogue



Response Generation in Knowledge-Grounded Dialogue

General knowledge grounded dialogue

(Dinan et al., 2019; Qin et al., 2019; Ghazvininejad et al., 2018; Tian et al., 2020; Gopalakrishnan et al., 2019; Moghe et al., 2018; Wu et al. 2020)

Hallucinations in text generation

(Maynez et al., 2020; Zhao et al., 2020; Cao et al., 2018; Falke et al., 2019; Puduppully et al., 2019; Filippova 2020)

Increasing faithfulness in informative dialogue

Response Generation in Knowledge-Grounded Dialogue

Proposed Evaluation Measures

Goal 1:
**Response is
objective**

First-person
detector

Goal 2:
**Response
content is
derived from
evidence**

Lexical precision
w.r.t. evidence

Goal 3:
**Response is
inferable from
the evidence**

Roberta MNLI
entailment
classifier

Response Generation in Knowledge-Grounded Dialogue

Wizard of Wikipedia (Dinan et al., 2019)

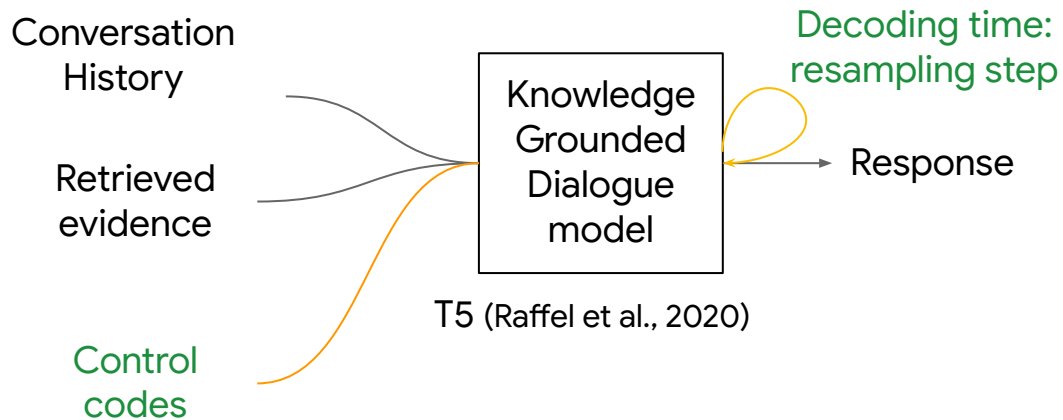
- Large scale dataset (75k ex of Wizard responses)
- Multi-turn, knowledge grounded
- Each turn - gold labeled evidence spans

1rst person: 44% of *Wiz. Utterances*
Avg Lex Prec: 0.43 on *Wiz. Utterances*
Non-entailing: 77% of *Wiz. Utterances*

Controllable
Text Generation

Response Generation in Knowledge-Grounded Dialogue

Approach



Response Generation in Knowledge-Grounded Dialogue

Control Token Sequence

Response in training data

Jerry Garcia passed away in 1995.

- Lexical prec w.r.t. evidence
- Contains 1st person
- NLI prediction

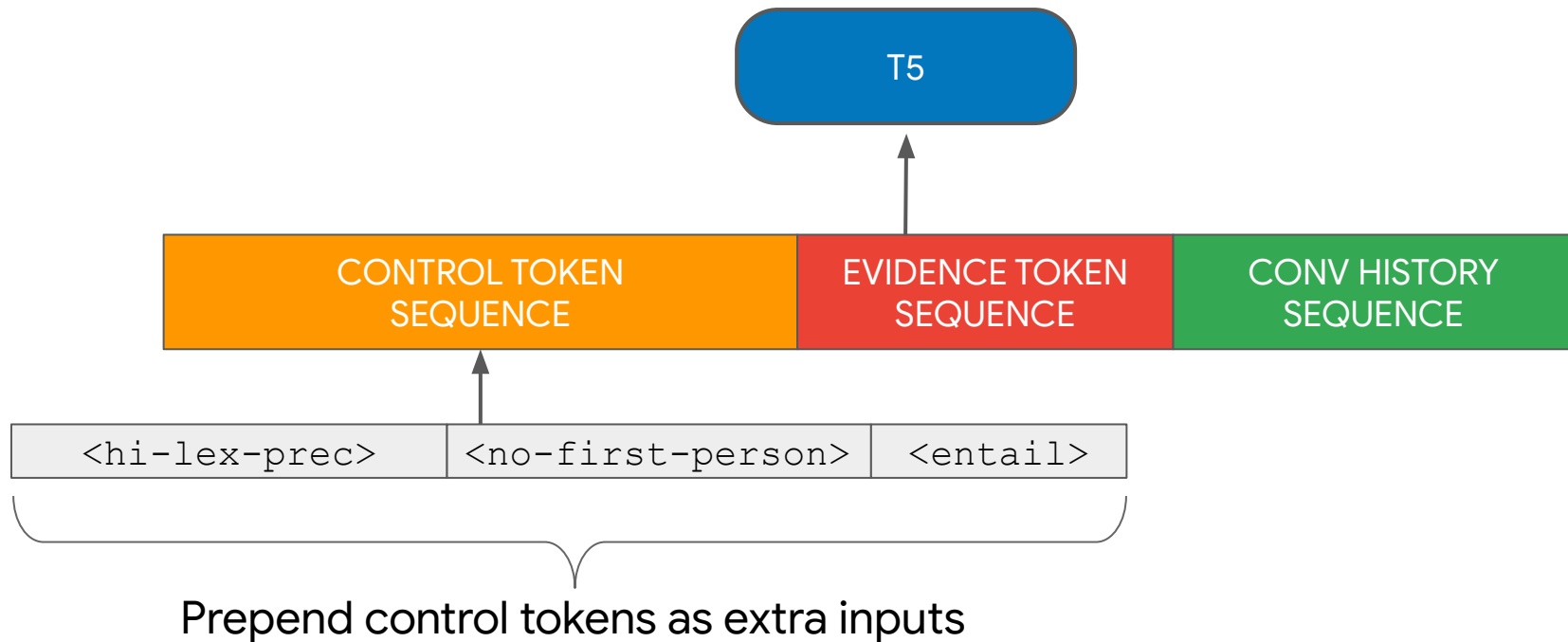
<hi-lex-prec>

<no-first-person>

<entail>

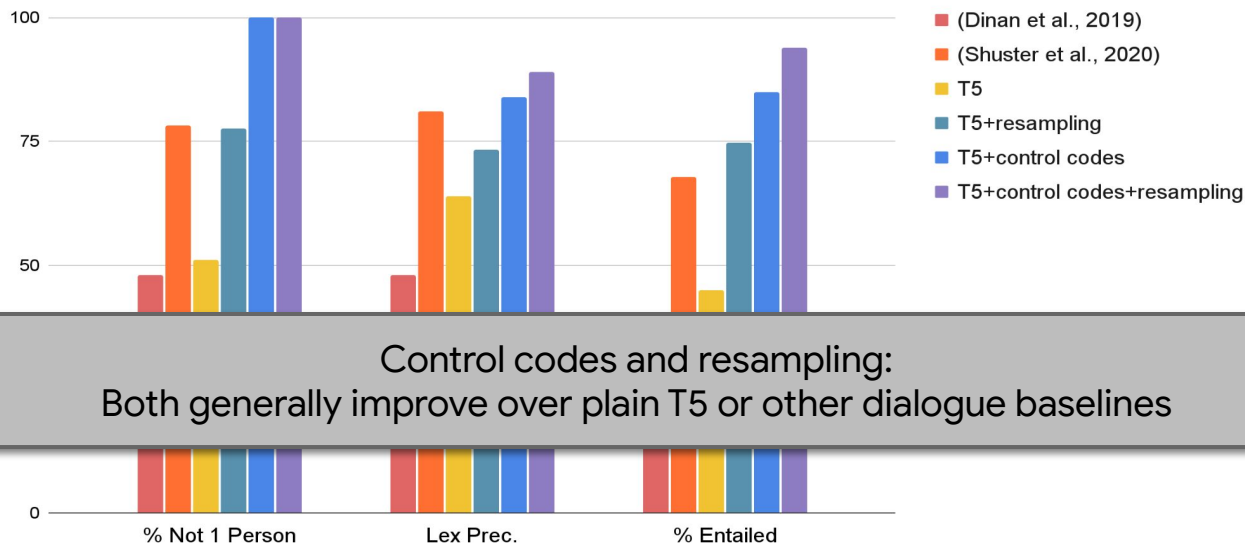
CONTROL TOKEN SEQUENCE

Response Generation in Knowledge-Grounded Dialogue



Response Generation in Knowledge-Grounded Dialogue

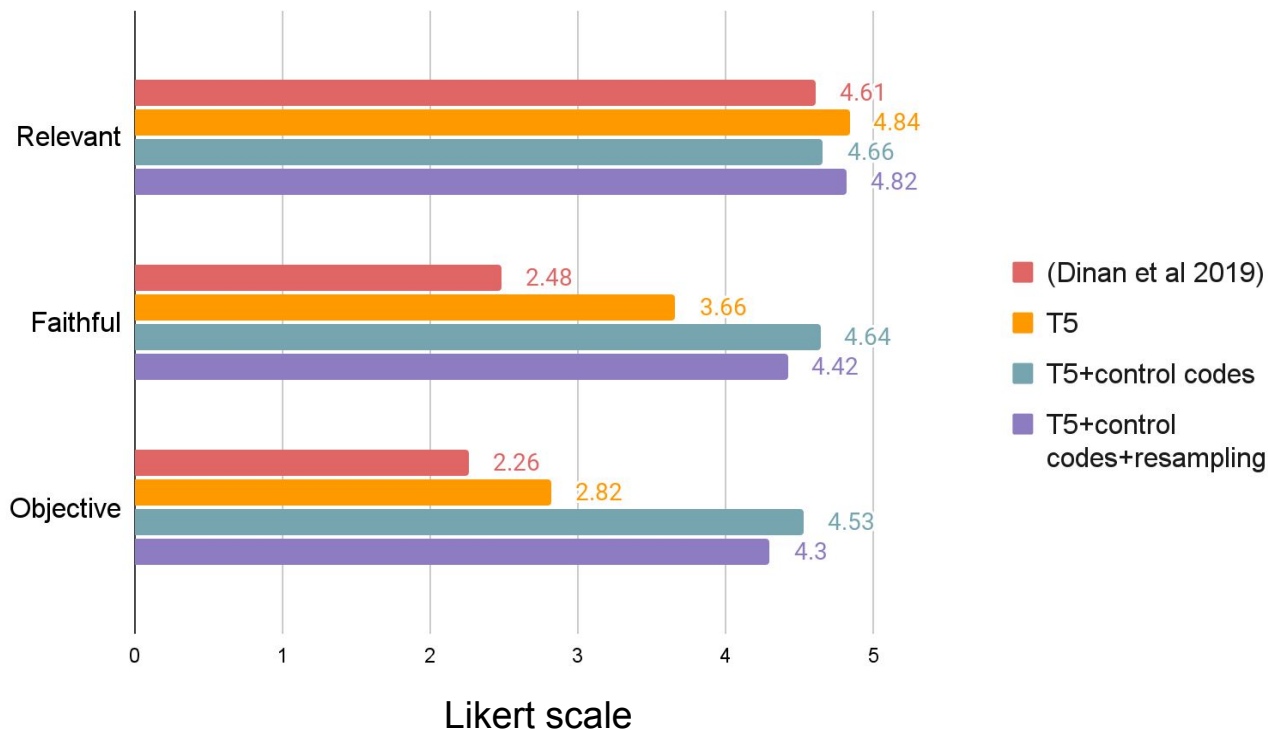
WoW Test Set Performance



Response Generation in Knowledge-Grounded Dialogue

Human Results

Controls → generally better **faithfulness** and **objectivity**



Controllable Models



```
graph TD; A[Controllable Models] --> B[Controllable Entity-Based Planning for Summarization (Narayan et al., 2021)];
```

Controllable Entity-Based
Planning for Summarization
(Narayan et al., 2021)

Controllable Entity-Based Planning for Summarization

Controllable Entity-Based Planning for Summarization

Chelsea's Eden Hazard and Arsenal's Santi Cazorla are set to reach a Premier League milestone this weekend when they each make their 100th appearance. Both players have been hugely influential since they moved to London in the summer of 2012, but who has been the most exciting import to watch? Here, Sportsmail's reporters choose the player they most enjoy seeing in action. Eden Hazard (L) and Santi Cazorla are both set to make their 100th Premier League appearance this weekend. Lee Clayton. Cazorla has wonderful balance. So does Hazard. Cazorla scores important goals. So does Hazard. Cazorla is two-footed. So is Hazard. Cazorla dances past opponents. So does Hazard. So, while there is not a lot to choose between them and Hazard is likely to get the most picks in this article, I am going for Cazorla. It's a personal choice. He is a wonderful footballer. I have paid to watch them both (and I will pay to watch them both again), but the little Spanish magician edges it for me. VERDICT: CAZORLA. Cazorla, pictured in action against Burnley, has been an influential part of Arsenal's midfield this season. Ian Ladyman. I remember when Manchester City balked at paying Hazard's wages when the Belgian was up for grabs in 2012. Back then City thought the young forward had a rather high opinion of his own worth for a player who was yet to play in a major European league. In the early days of his time at Chelsea, it looked as though City may have been right. He showed flashes of brilliance but also looked rather too easy to push off the ball. Roll forward to 2015, however, and the 24-year-old has developed in to one of the most important players in the Barclays Premier League. Brave, strong and ambitious, Hazard plays on the front foot and with only one thought in this mind. Rather like Cristiano Ronaldo, he has also developed in to the type of player ever defender hates, simply because he gets back up every time he is knocked to the ground. He would get in every team in the Premier League and is one of the reasons Chelsea will win the title this season. VERDICT: HAZARD. Hazard controls the ball under pressure from Stoke midfielder Stephen Ireland at Stamford Bridge. Dominic King. It has to be Hazard. I saw him play for Lille twice in the season before he joined Chelsea – once against St Etienne, the other was what proved to be his final appearance against Nancy. He scored two in the first match, a hat-trick the latter and played a different game to those around him. He hasn't disappointed since arriving here and I love the nonchalance with which he takes a penalty, his low centre of gravity and the way he can bamboozle defenders. If there is such a thing as £32million bargain, it is Hazard. VERDICT: HAZARD. Hazard celebrates after scoring a fine individual goal in Chelsea's 3-2 win against Hull in March. Nick Harris. Now this is a tricky one because while Eden Hazard will frequently embark on a dribble or dink in a pass that will make you nod in appreciation, he'll also miss a penalty and make you groan. Whereas the older Cazorla, less flashy but no less of a technical master, is to my mind more of a fulcrum, more important relatively to the sum of Arsenal's parts than Hazard is to Chelsea. You'll gasp at Hazard but Cazorla's wow factor is richer. That's not to dismiss either: both are brilliant footballers, contributing goals, assists and flair. Any neutral would bite your hand off to have either playing in your team. Forced to pick though, it's Cazorla, for his consistency and crucially doing it in the biggest games. Exhibit A would be Manchester City 0 Arsenal 2 in January; goal, assist, all-round brilliance, against a big team, at an important time. VERDICT: CAZORLA. Cazorla scores from the penalty spot in Arsenal's 2-0 away win at Manchester City in January. Riath Al-Samarrai. Eden Hazard for me. Cazorla is an utter delight, a little pinball of a man who is probably the most two-footed player I've seen. Put him in a tight space and then you see what makes him rare among the best. But Hazard is the top player in the Premier League, in my opinion. This is the sixth of his eight seasons as a professional where he has reached double figures and yet he offers so much more than goals (36 in 99 in the Premier League for Chelsea). He can beat a man and, better still, you sense he likes doing it. Technically, his passing and shooting are excellent and he also has a mind capable of sussing out the shapes and systems in front of him. That intelligence, more specifically.

Typical source-target scenario

Chelsea star Eden Hazard is set to make his 100th top-flight appearance. Santi Cazorla should hit the same milestone when Arsenal meet Burnley. Both players have impressed since moving to the Premier League in 2012. Hazard has more goals this season but Cazorla has one more assist. Sports-mail's reporters choose the player who has excited them the most.

Human summary (Target)

Controllable Entity-Based Planning for Summarization

Chelsea's Eden Hazard and Arsenal's Santi Cazorla are set to reach a Premier League milestone this weekend when they each make their 100th appearance. Both players have been hugely influential since they moved to London in the summer of 2012, but who has been the most exciting import to watch? Here, Sportsmail's reporters choose the player they most enjoy seeing in action. Eden Hazard (L) and Santi Cazorla are both set to make their 100th Premier League appearance this weekend. Lee Clayton. Cazorla has wonderful balance. So does Hazard. Cazorla scores important goals. So does Hazard. Cazorla is two-footed. So is Hazard. Cazorla dances past opponents. So does Hazard. So, while there is not a lot to choose between them and Hazard is likely to get the most picks in this article, I am going for Cazorla. It's a personal choice. He is a wonderful footballer. I have paid to watch them both (and I will pay to watch them both again), but the little Spanish magician edges it for me. VERDICT: CAZORLA. Cazorla, pictured in action against Burnley, has been an influential part of Arsenal's midfield this season. Ian Ladyman. I remember when Manchester City balked at paying Hazard's wages when the Belgian was up for grabs in 2012. Back then City thought the young forward had a rather high opinion of his own worth for a player who was yet to play in a major European league. In the early days of his time at Chelsea, it looked as though City may have been right. He showed flashes of brilliance but also looked rather too easy to push off the ball. Roll forward to 2015, however, and the 24-year-old has developed in to one of the most important players in the Barclays Premier League. Brave, strong and ambitious, Hazard plays on the front foot and with only one thought in this mind. Rather like Cristiano Ronaldo, he has also developed in to the type of player ever defender hates, simply because he gets back up every time he is knocked to the ground. He would get in every team in the Premier League and is one of the reasons Chelsea will win the title this season. VERDICT: HAZARD. Hazard controls the ball under pressure from Stoke midfielder Stephen Ireland at Stamford Bridge. Dominic King. It has to be Hazard. I saw him play for Lille twice in the season before he joined Chelsea – once against St Etienne, the other was what proved to be his final appearance against Nancy. He scored two in the first match, a hat-trick the latter and played a different game to those around him. He hasn't disappointed since arriving here and I love the nonchalance with which he takes a penalty, his low centre of gravity and the way he can bamboozle defenders. If there is such a thing as £32million bargain, it is Hazard. VERDICT: HAZARD. Hazard celebrates after scoring a fine individual goal in Chelsea's 3-2 win against Hull in March. Nick Harris. Now this is a tricky one because while Eden Hazard will frequently embark on a dribble or dink in a pass that will make you nod in appreciation, he'll also miss a penalty and make you groan. Whereas the older Cazorla, less flashy but no less of a technical master, is to my mind more of a fulcrum, more important relatively to the sum of Arsenal's parts than Hazard is to Chelsea. You'll gasp at Hazard but Cazorla's wow factor is richer. That's not to dismiss either: both are brilliant footballers, contributing goals, assists and flair. Any neutral would bite your hand off to have either playing in your team. Forced to pick though, it's Cazorla, for his consistency and crucially doing it in the biggest games. Exhibit A would be Manchester City 0 Arsenal 2 in January; goal, assist, all-round brilliance, against a big team, at an important time. VERDICT: CAZORLA. Cazorla scores from the penalty spot in Arsenal's 2-0 away win at Manchester City in January. Riath Al-Samarrai. Eden Hazard for me. Cazorla is an utter delight, a little pinball of a man who is probably the most two-footed player I've seen. Put him in a tight space and then you see what makes him rare among the best. But Hazard is the top player in the Premier League, in my opinion. This is the sixth of his eight seasons as a professional where he has reached double figures and yet he offers so much more than goals (36 in 99 in the Premier League for Chelsea). He can beat a man and, better still, you sense he likes doing it. Technically, his passing and shooting are excellent and he also has a mind capable of sussing out the shapes and systems in front of him. That intelligence, more specifically.

Chelsea star Eden Hazard is set to make his 100th top-flight appearance. Santi Cazorla should hit the same milestone when Arsenal meet Burnley. Both players have impressed since moving to the Premier League in 2012. Hazard has more goals this season but Cazorla has one more assist. Sports-mail's reporters choose the player who has excited them the most.

Human summary

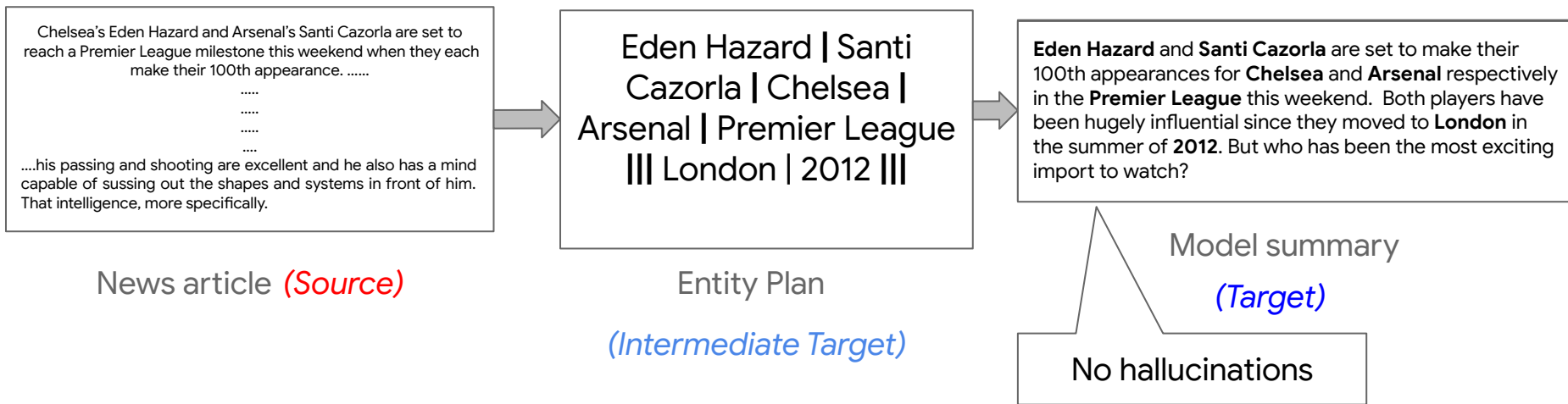
Eden Hazard and Santi **Cazorlag** will each make their 100th Premier League appearance this weekend. **nightstandapplication.com**. Sportsmail's **hovercraft** reporters choose their **man of the match** **countermeasures**.

Model summary
under nucleus
sampling

 = hallucinations, not attributable to source

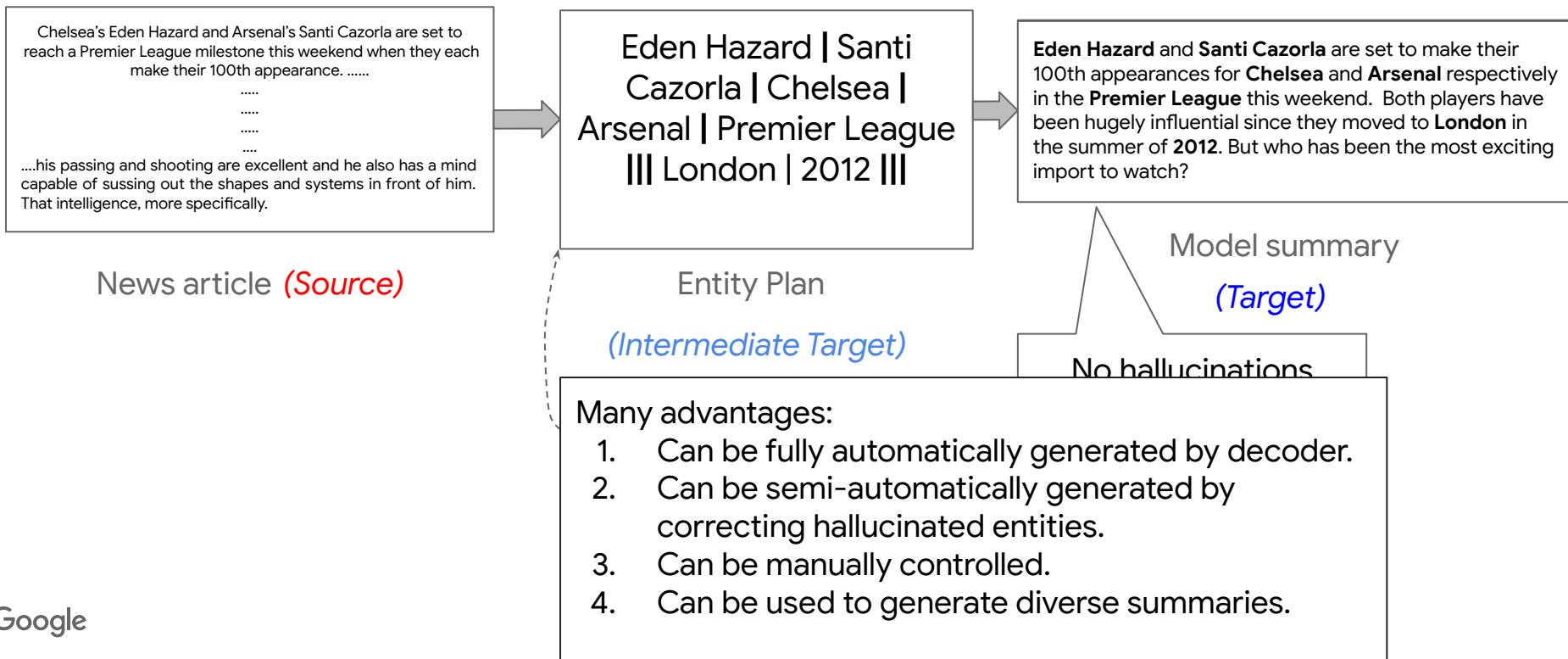
Controllable Entity-Based Planning for Summarization

Proposal: Break the summarization step into two parts (**FROST**).



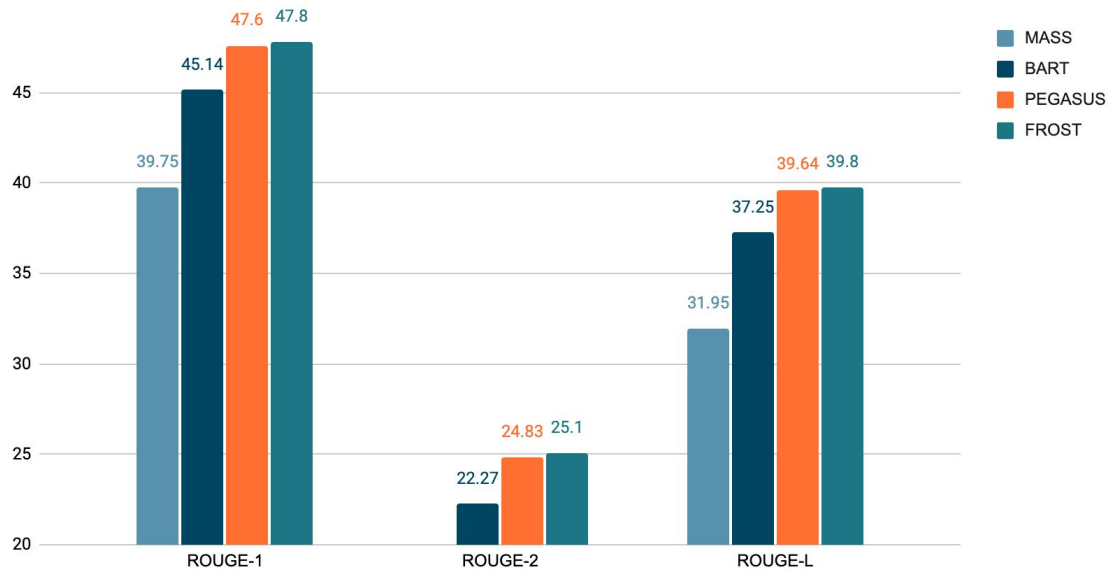
Controllable Entity-Based Planning for Summarization

Proposal: Break the summarization step into two parts (**FROST**).



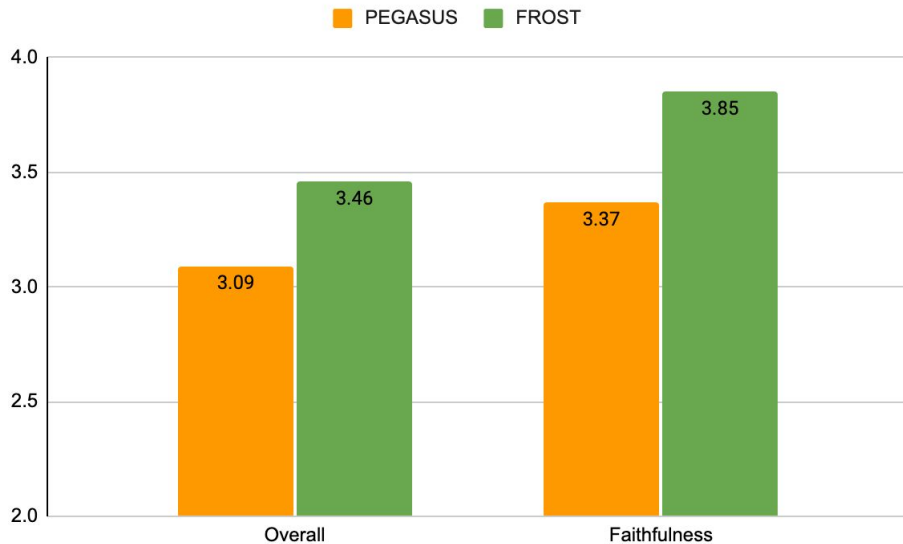
Controllable Entity-Based Planning for Summarization

Results on XSUM (Automatic Metrics)



Controllable Entity-Based Planning for Summarization

Summarization results by
controlling entity chains
automatically
(human evaluations)



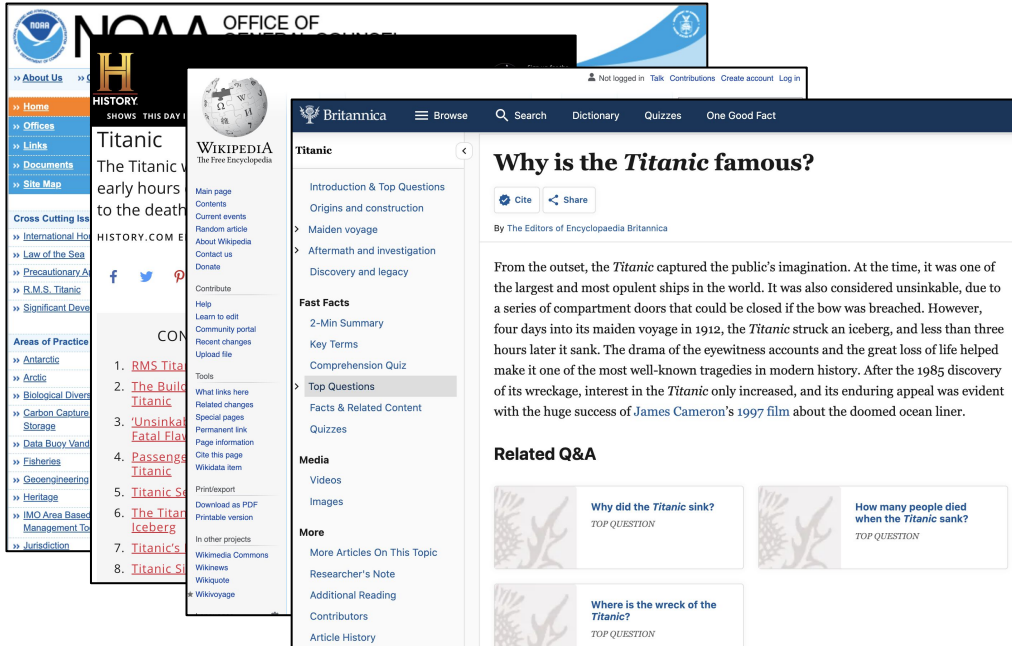
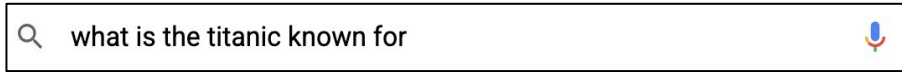
Controllable Models



```
graph TD; A[Controllable Models] --> B[Conditional Generation with a Question-Answering Blueprint (Narayan et al., 2022)];
```

Conditional Generation with
a Question-Answering
Blueprint
(Narayan et al., 2022)

blueprint: Query-focused Multi-document Summarization



Challenges

- Some queries can only be answered through a **long-form** answer
- These queries may require **multiple documents** to be answered
- Plans using entities alone become **less interpretable**
- **Attribution** is a bigger problem

blueprint: Question-Answer Pairs as Content Plans

what is the titanic known for



Q: What kind of ship was the RMS Titanic?
A: A British passenger liner

Q: What line operated the RMS Titanic?
A: the White Star Line

~~Q: How many ships did the white star line make?
A: three ships~~

Q: In what ocean did the RMS Titanic sink?
A: ~~Pacific Ocean~~ North Atlantic Ocean

Q: During what voyage did the RMS Titanic hit an iceberg?
A: its maiden voyage from Southampton, UK, to New York City



Idea: Use a [question-answering blueprint](#) as an intermediate planning stage for conditional text generation.

Connection to [Questions Under Discussion](#) (QUD) theory of discourse structure

Results:

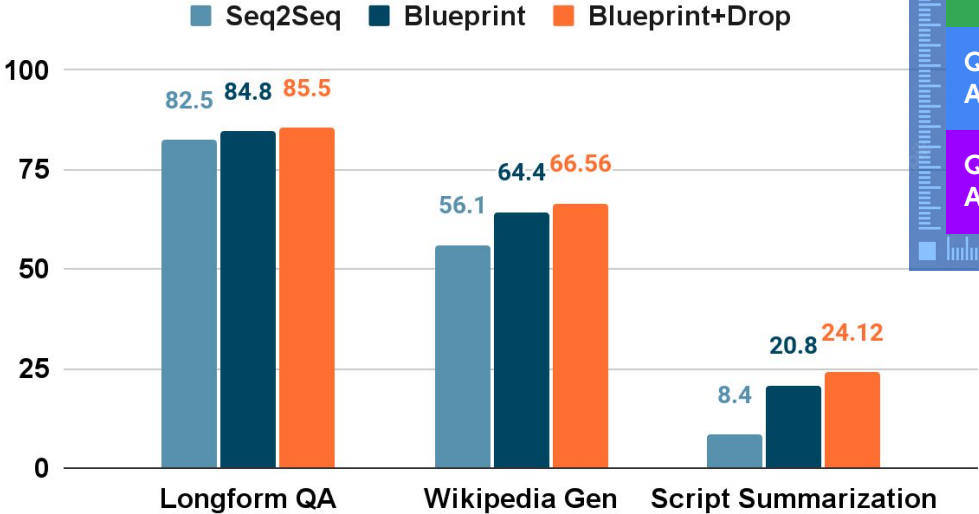
- ✓ Reduced factuality errors
- ✓ Increased controllability and explainability
- ✓ Better long-form summaries

RMS Titanic was a British passenger liner, operated by the White Star Line, which sank in the North Atlantic Ocean after striking an iceberg during its maiden voyage from Southampton, UK, to New York City.

Answer Summary

blueprint: Grounded and Controllable Generation to Improve Faithfulness

% of responses entailed by the input



Q: What kind of ship was the RMS Titanic?
A: A British passenger liner

Q: What line operated the RMS Titanic?
A: the White Star Line

~~Q: How many ships did the white star line make?
A: three ships~~

Q: In what ocean did the RMS Titanic sink?
A: ~~Pacific Ocean~~

Q: During what voyage did the RMS Titanic hit an iceberg?
A: its maiden voyage from Southampton, UK, to New York City

RMS Titanic was a British passenger liner, operated by the White Star Line, which sank in the North Atlantic Ocean after striking an iceberg during its maiden voyage from Southampton, UK, to New York City.

Summary

- Methods for standardization of data creation and human evaluation for NLG problems
 - ToTTo, GEM, AIS
- Introduction of Learned Metrics
 - BLEURT+
- Controllable models
 - Controllable encoders (dialogue responses) and decoders (FROST, Blueprint)

Future Work

- Tighter loop between models, automatic metrics and human intervention
 - Adaptable learned metrics for all NLG tasks
 - Reusable human evaluation methods (AIS+)
 - Planning based attributed generation
- Few-shot approaches with LLMs
 - See new paper “Query Refinement Prompts for Closed-Book Long-Form Question Answering” from Amplayo et al., (2022).

Thank you