# Examining Conditional Language Models

Claire Gardent

Joint work with Juliette Faille,

Albert Gatt, Quentin Brabant, Lina Rojas-Barahona and Gwénolé Lecorvé

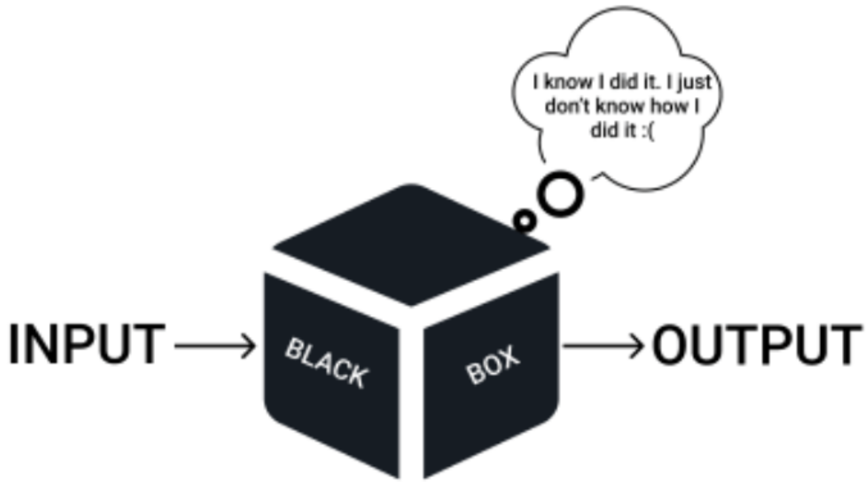
Interactive *Natural Language* Technology for eXplainable Artificial Intelligence

NL4XAI

Loria

Laboratoire lorrain de recherche en informatique et ses applications

# Language Models are not interpretable ...




***Interpretable*** : directly understandable by humans

# ... but we can try to make them explainable

Using

**Fine-Grained Evaluation Metrics** : to analyse, detect and quantify errors

**Probing** : to understand where the errors come from

**Explainable Models** : explain output based on non latent, interpretable, intermediate results

***Explainable*** : can be explained to a human

# Generation Tasks

## Knowledge Graph-to-Text Generation

- Semantic Adequacy
- Detecting, quantifying and analysing the source of semantic errors

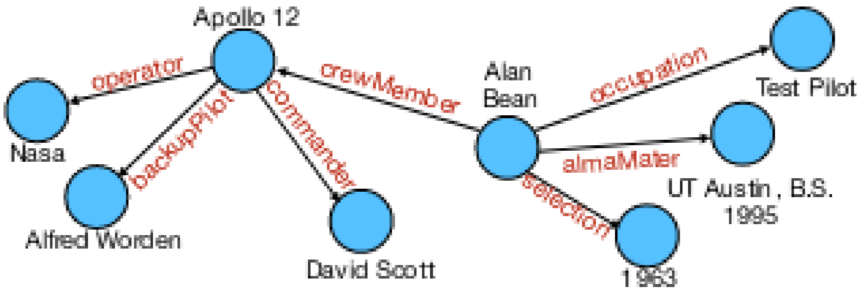
## Knowledge-Based Dialog

- Coherence and Cohesion
- Using intermediate results to evaluate coherence and cohesion

# Knowledge Graph-to-Text Generation

Converting Knowledge Graphs to Text

# Example




$\Downarrow$

Alan Bean graduated from UT Austin in 1955 with a Bachelor of Science degree. He was hired by NASA in 1963 and served as a test pilot. Apollo 12's backup pilot was Alfred Worden and was commanded by David Scot

# Detecting Omissions

# Omissions

```
Lady_Anne_Monson | birthPlace | Darlington
Lady_Anne_Monson | birthDate | 1726-01-01
Lady_Anne_Monson | deathDate | 1776-02-18
Lady_Anne_Monson | birthPlace | Kingdom_of_England
Lady_Anne_Monson | residence | India
```

Born in the ***Kingdom of England*** in ***1726-01-01*** , and living in ***India*** , on the 18th of July, 1776, the country is the birth place of Joh Davutoglu.

***Omissions are entities with no corresponding mentions***

# RDF-to-Text Evaluation Metrics

## Global Metrics

***Scores the generated text***

- BLEU
- BERTScore
- METEOR
- Chrf++
- DataQuestEval
- BLEURT
- ...

## Fine-Grained Metric

***Scores the level of omissions***

- Entity-Based Semantic Adequacy (ESA)
- Entity-Based Semantic Inadequacy (ESI)

# Detecting Omissions

Given a (graph,text) pair, the algorithm returns a ***list of (graph entity, text span)*** pairs using:

- An Entity Linker: maps text spans to KB entities
- Approximate string matching: between text n-grams and graph entities
- Pronoun resolution: resolve and match antecedent with graph entities
- A Date parser: normalise and match

***Omissions = Graph entities with no corresponding text mention***

# Evaluating Omission Detection

## Quantitative Analysis

Benchmark

- WebNLG gold data 2017
- 25K (graph, text) pairs where entity mentions have been manually annotated

Precision: 0.83

Recall: 0.82

## Qualitative Analysis

WebNLG System Outputs

- 11 texts with automatically detected omissions but high human semantics score
- 10 with missing mentions
- 1 degenerate text

# Entity-based semantic Adequacy

$$ESA = \frac{count(InputEntitiesDetected)}{count(InputEntities)}$$

```
Lady_Anne_Monson | birthPlace | Darlington
Lady_Anne_Monson | birthDate | 1726-01-01
Lady_Anne_Monson | deathDate | 1776-02-18
Lady_Anne_Monson | birthPlace | Kingdom_of_England
Lady_Anne_Monson | residence | India
```

$$ESA_I = 0.5$$

*Born in the **Kingdom of England** in **1726-01-01** , and living in **India**, on the 18th of July, 1776, the country is the birth place of Joh Davutoglu.*

6 RDF Entities in the input

3 RDF entities detected in the generated text

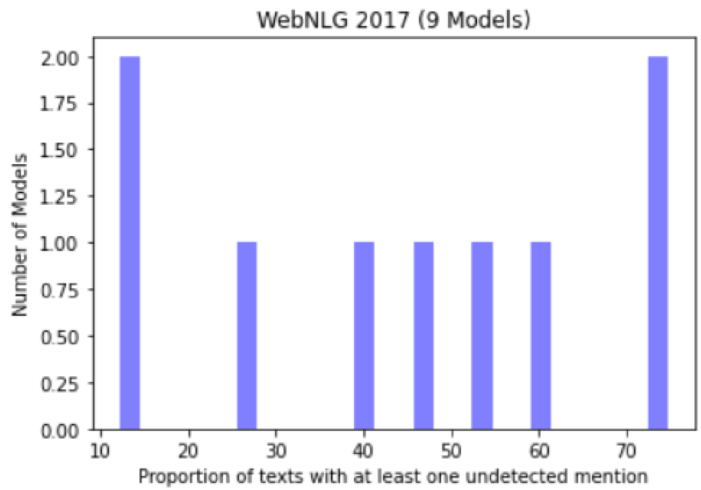
# Corpus Level Omission Metrics

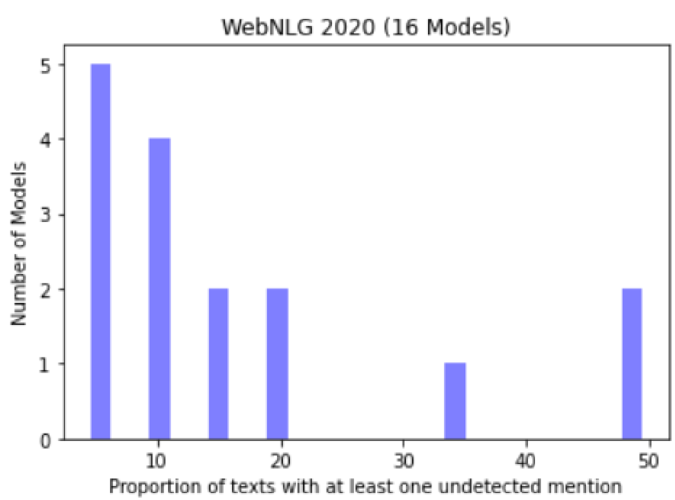
How well does a ***model*** handle a ***corpus*** ?

$ESA_C$ = Average ESA score on corpus

$$ESI_C^n = \frac{count(Text\ with\ at\ least\ n\ Undetected\ Entity)}{\mathrm{count}(Text)}$$

# Evaluating RDF-to-Text Generation Models




25 models from the WebNLG 2017 and 2020 challenges.

2017

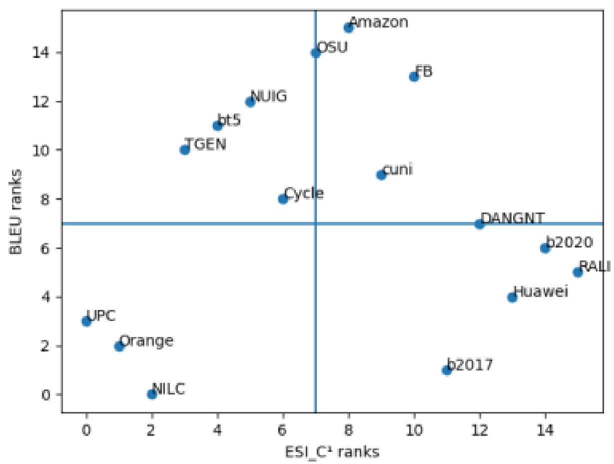
- 10 to 77% of the generated texts have at least one omission
- 6/9 models: 40% of the generated texts have at least one omission

2020

- 5 to 45% of the generated texts have at least one omission
- For the top 5 models, the generated text omits at least one entity 5% of the time
- the remaining 11 models omit at least one entity 10% of the time or more

# BLEU vs Entity Based Semantic Adequacy

WebNLG 2020 Models are ranked with respect to BLEU and $ESI_1$ score




***A High BLEU does not garantee that all entities are mentioned***

For the 8 models with highest BLEU rank

- only 3 have high ESI rank (Amazon, FB and cuni-ufal).
- the other 5 have an ESI score ranging between 10 and 22%. On average ***they fail to mention at least one of the input entities 10 to 22% of the time*** . (OSU, CycleGT, NUIG, TGen, bt5)

# Correlation with Human Judgments and other Metrics

Correlation with human judgement of Semantic Adequacy

- 2017: strong (R: 0.66)
- 2020: moderate (R: 0.53-0.57)

| | METEOR | TER | Fluency | Grammar | Semantics | $ESA_I$ |
|---|---|---|---|---|---|---|
| BLEU | 0.74 | -0.57 | 0.39 | 0.43 | 0.53 | 0.59 |
| METEOR | | -0.54 | 0.57 | 0.63 | 0.72 | **0.87** |
| TER | | | -0.42 | -0.45 | -0.4 | -0.42 |
| Fluency | | | | 0.89 | 0.51 | 0.49 |
| Grammar | | | | | 0.57 | 0.57 |
| Semantics | | | | | | **0.66** |

Correlation with global automatic metrics is moderate (R:0.39 - 0.87)

| | Bl-nltk | Met | chrf | TER | BSC-P | -R | -F1 | BL | Cor | Cov | Fl | REl | Str | **ESA** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AUTO** | | | | | | | | | | | | | | |
| BLEU | 0.97 | 0.71 | 0.82 | -0.67 | 0.69 | 0.66 | 0.71 | 0.49 | 0.42 | 0.3 | 0.34 | 0.33 | 0.31 | 0.41 |
| BLEU NLTK | | 0.77 | 0.87 | -0.74 | 0.74 | 0.72 | 0.77 | 0.54 | 0.45 | 0.34 | 0.39 | 0.36 | 0.36 | 0.39 |
| METEOR | | | 0.9 | -0.62 | 0.67 | 0.82 | 0.78 | 0.67 | 0.49 | 0.49 | 0.4 | 0.42 | 0.36 | **0.45** |
| chrF++ | | | | -0.69 | 0.74 | 0.82 | 0.82 | 0.6 | 0.51 | 0.46 | 0.41 | 0.43 | 0.37 | **0.45** |
| TER | | | | | -0.76 | -0.67 | -0.75 | -0.61 | -0.41 | -0.31 | -0.42 | -0.39 | -0.4 | -0.24 |
| BERT-score P | | | | | | 0.83 | 0.95 | 0.73 | 0.6 | 0.41 | 0.52 | 0.56 | 0.5 | 0.39 |
| BERT-score R | | | | | | | 0.95 | 0.75 | 0.57 | 0.52 | 0.49 | 0.49 | 0.45 | 0.43 |
| BERT-score F1 | | | | | | | | 0.77 | 0.61 | 0.49 | 0.53 | 0.55 | 0.5 | **0.44** |
| BLEURT | | | | | | | | | 0.62 | 0.54 | 0.52 | 0.59 | 0.5 | 0.43 |
| **HUMAN** | | | | | | | | | | | | | | |
| Correctness | | | | | | | | | | 0.75 | 0.71 | 0.83 | 0.67 | 0.56 |
| DataCoverage | | | | | | | | | | | 0.62 | 0.76 | 0.57 | **0.57** |
| Fluency | | | | | | | | | | | | 0.67 | 0.86 | 0.41 |
| Relevance | | | | | | | | | | | | | 0.65 | 0.53 |
| TextStructure | | | | | | | | | | | | | | 0.36 |

# Detecting Hallucinations

| Model | >1 | >1✓ | Dist | ↓ ESI$_C$[1] |
|---|---|---|---|---|
| RALI | 0 | 0 | 0 | 0% |
| B-2017 | 1 | 1 | 1 | 0.1% |
| B-2020 | 1 | 1 | 1 | 0.1% |
| NUIG | 4 | 3 | 3 | 0.2% |
| UPC | 4 | 4 | 3 | 0.2% |
| DANGNT | 5 | 5 | 5 | 0.3% |
| TGen | 8 | 7 | 2 | 0.5% |
| cuni-ufal | 9 | 7 | 6 | 0.5% |
| Amazon | 9 | 9 | 3 | 0.5% |
| FBConvAI | 17 | 11 | 6 | 1% |
| CycleGT | 19 | 18 | 10 | 1% |
| OSU | 20 | 19 | 3 | 1% |
| bt5 | 36 | 17 | 3 | 2% |
| Huawei | 48 | 47 | 28 | 3% |
| NILC | 117 | 99 | 66 | 7% |
| ORANGE | 288 | 288 | 60 | 16% |
| UIT | 1 | 0 | 1 | 0.1% |
| Tilburg SMT | 4 | 0 | 4 | 0.2% |
| Tilburg NMT | 11 | 4 | 7 | 0.6% |
| UPF | 12 | 8 | 4 | 0.6% |
| Tilburg Pl | 14 | 11 | 6 | 0.8% |
| Melbourne | 114 | 112 | 24 | 6% |
| Adapt | 241 | 234 | 151 | 13% |
| PKUWriter | 286 | 283 | 135 | 15% |
| Baseline | 754 | 144* | 147 | 40% |

Hallucinations: ***Mentions in the output text that have no corresponding RDF entity in the input graph*** (Entity linking only).

On 144 randomly chosen texts

1: Number of texts with at least one hallucination

1✓: Number of texts with at least one hallucination which are manually validated

# Qualitative Analysis

Three main causes for omissions: short texts, hallucination, degenerate output

**Short Text**
(*Liselotte_Grschebina*, *nationality*, *Israel*)
(*Israel*, *areaTotal*, *20769100000.0*)
(*Israel*, *officialLanguage*, *Modern_Standard_Arabic*)
(*Liselotte_Grschebina*, *birthPlace*, *German_Empire*)
(*Liselotte_Grschebina*, *training*, *School_of_Applied_Arts_in_Stuttgart*)

Liselotte Grschebina is a German national who was born in the German Empire and has a total area of 20769100000. 0.

# Qualitative Analysis

**Hallucination**

(*Lady_Anne_Monson*, *birthPlace*, *Darlington*)
(*Lady_Anne_Monson*, *birthDate*, *1726-01-01*)
(*Lady_Anne_Monson*, *deathDate*, *1776-02-18*)
(*Lady_Anne_Monson*, *birthPlace*, *Kingdom_of_England*)
(*Lady_Anne_Monson*, *residence*, *India*)

Born in the Kingdom of England in 1726-01-01, and living in India, on the 18th of July, 1776, **the country** is the birth place of **Joh Davutoglu**.

# Qualitative Analysis

**Degenerate Output**

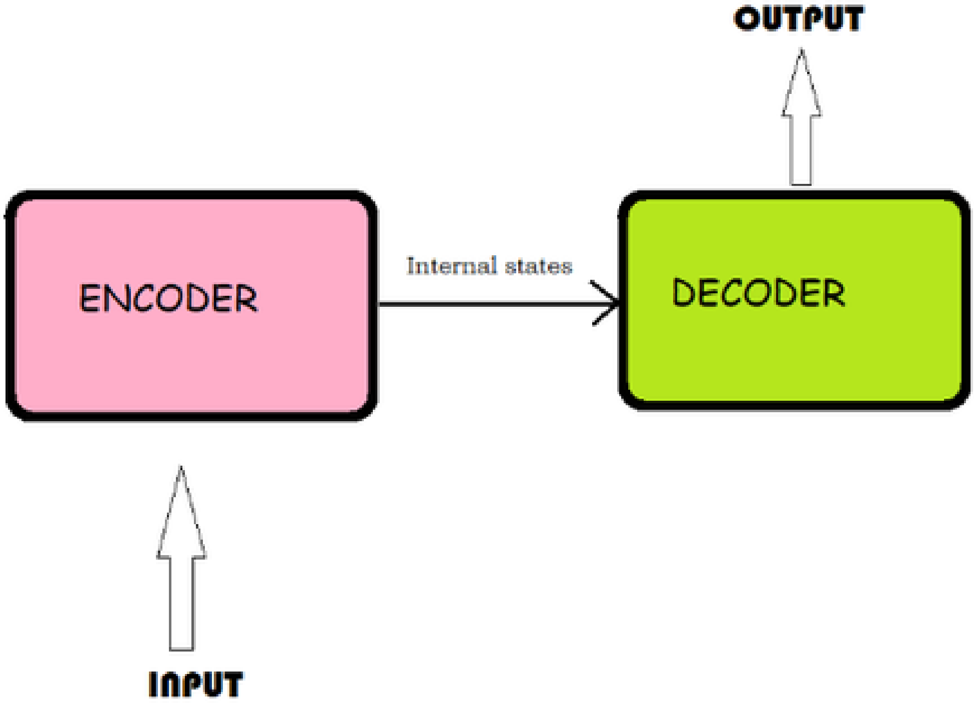
(*Lady_Anne_Monson*, *birthPlace*, *Darlington*)
(*Lady_Anne_Monson*, *birthDate*, *1726-01-01*)
(*Lady_Anne_Monson*, *deathDate*, *1776-02-18*)
(*Lady_Anne_Monson*, *birthPlace*, *Kingdom_of_England*)
(*Lady_Anne_Monson*, *residence*, *India*)

Born in the Kingdom of England, and died on 1776-02-18, on 1726-01-01, in the Kingdom of England, the prime minister of community of England is called, Germanic duties, and arrabbiata (born on the 18th of July, 1726-01-01).

# Analysing Omissions

Where do omissions come from ?

# Where do omissions come from ?



# Probing the Encoder

***Can we detect omissions in the encoder representations?***

Two probing methods

- Parametric: classifier probe
- Non parametric method based on encoding similarity

# Probing Experiments

Generate texts from graphs

Annotate generated text for omissions

Use annotated data to train and test a probe

# Generating Texts from Knowledge Graphs

Generation Model

- T5 and BART
- fine-tuned on the WebNLG training data, 47k (RDF graph, text) pairs where the RDF graphs are subgraphs of DBPedia and texts are crowd-sourced.

Creating (Graph,Text) Data

- 22,657 RDF input graphs
    - 16,657 RDF graphs from the WebNLG V3.0 dataset
    - 6k graphs from the KELM dataset (1k graphs for each graph size from 1 to 6 triples)
- permute input
- generate
- filter repeated output

***71,644 (graph, text) pairs***

# Creating Omission Data

Labelling (Graph,Text) pairs with omissions

- Automatic annotation
  - All 71K texts
- Manual annotation
  - 3 NLP MSc students
  - Kappa between each pair of annotators: 0.56 to 0.69
  - 13K texts
  - omissions and distortions

Data for probing experiments

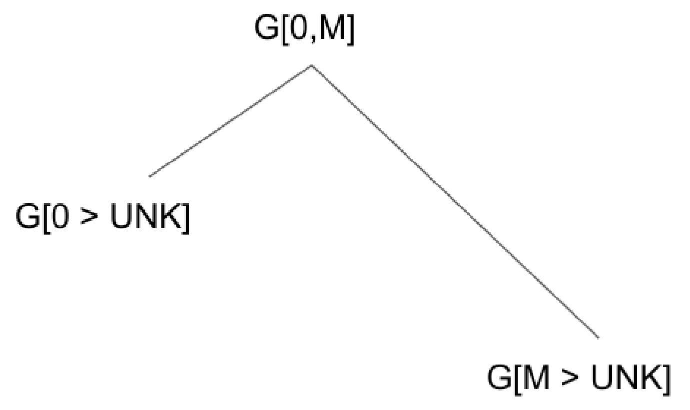
- Texts with at least one omission or distortion
- ***33,160 texts automatically labelled with omissions***
- ***6,249 texts manually labelled with omission, 6,518 with distortion***
- Train/dev/test: 70/15/15

# Example Distortions

| RDF Entity | Distortion |
|---|---|
| Olga_Bondareva | Olgaondarev |
| 177539.0 | 1777539 |
| Ciudad_Ayala | Ciudad Ayalatus |
| Lee Jae-hak | Lee Lee-hak |
| Doosan Bears | Donosan Bears |
| Lionsgate | Lionsburg |
| 1997 | 1996 |
| EGBF | EAWFB |
| Columbia_Records | The Columbus Records |
| 1929-06-11 | June 5th, 1929 |
| St._Louis,_Missouri | St Louis, Mississippi |
| 11.51147.0 | 11.5 |
| -6 | Delta 6 |

# Parameter free probing

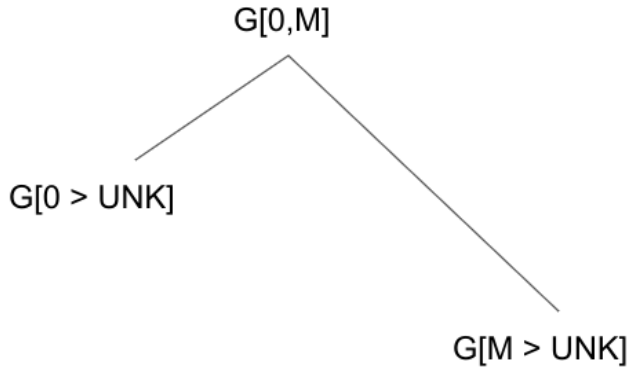
## Intuition




- Omitted entities have a weak signal
- Because it lacks specificity, the representation of an omitted entity is more similar to the representation of the unknown token UNK than the representation of an entity that is correctly verbalised in the output text.

# Parameter free probing

We compare the similarity between the encoder representation of a graph leading to an omission with two alternative representations




Average similarity for mentions:

$$cos(g, g^{\backslash M}) = \frac{1}{K_g} \sum_{k=1}^{K_g} sim(g, g^{\backslash m_k})$$

Ratio of graphs such that:

$$cos(g, g^{\backslash o}) > cos(g, g^{\backslash M})$$

# Parameter free probing Results

| | All | In Domain | | | OOD | |
|---|---|---|---|---|---|---|
| | | W-T | W-D | W-S | W-U | K |
| O | 0.68 | 0.64 | 0.72 | 0.61 | *0.52* | 0.77 |
| O+D | 0.54 | 0.66 | 0.70 | 0.46 | 0.38 | *0.50* |
| D | 0.44 | 0.70 | 0.68 | 0.47 | 0.45 | 0.47 |
| Auto | 0.66 | 0.83 | 0.85 | 0.56 | 0.44 | 0.65 |

Most results are statistically significant showing that encodings of graphs lieading to omissions are different from those that do not.

On average, the proportion of graphs for which $sim(g, g^{\setminus o}) > sim(g, g^{\setminus M})$ is

- 66% for the automatically annotated data
- 68% for the manually annotated data

The difference is less on OOD data as these have weaker signal than graphs seen during training.

# Parametric probe

Binary Classifier

- Two-layer Multi-layer Perceptron
- Trained on ***(encoding(graph), encoding(entity), label)***
- Label = 1 if the entity is not omitted, 0 otherwise

Aka entailment relation between a graph representation and an entity

$$1 \text{ if } g \models e, \text{ else } 0$$

| | |
|---|---|
| **Manual-O+D** | |
| F1 | **0.82** |
| **Manual-O** | |
| F1 | **0.69** |
| **Manual-D** | |
| F1 | **0.79** |

- The probe successfully classifies distortions and omissions
- Distortions are easier to detect
- Complementary to parameter-free probe

# Upper Bound

Binary Classifier

- Trained to distinguish entities present in a graph from ***entities absent from that graph***
- Trained on 18k graphs and 198K entities
- Entity not present in the input graph viewed as an extreme case of omission
- Input: encoding(Graph), encoding(entity)
- Label: 1 if the entity is in graph, 0 otherwise

| | |
|---|---|
| **Manual-O+D** | |
| F1 | **0.82** |
| **Manual-O** | |
| F1 | **0.69** |
| **Manual-D** | |
| F1 | **0.79** |
| **Upper-Bound** | |
| F1 | **0.97** |

***F1 on class 0 is high***

- The probe can detect whether or not an entity is present from the embedding of an RDF graph.
- Absent entities are easier to spot than omitted or distorted entities

# Control Task

Is the probe really evaluating the embeddings or does it memorise the training data?

Training set with random labels

| | |
|---|---|
| **Manual-O+D** | |
| F1 | **0.82** |
| $C_{F1}$ | **0.00** |
| $S_{F1}$ | 0.82 |
| **Manual-O** | |
| F1 | **0.69** |
| $C_{F1}$ | **0.00** |
| $S_{F1}$ | 0.69 |
| **Manual-D** | |
| F1 | **0.79** |
| $C_{F1}$ | **0.00** |
| $S_{F1}$ | 0.79 |
| **Upper-Bound** | |
| F1 | **0.97** |

Selectivity = drop in performance between the probe (trained on the original dataset) and the control probe (trained on the randomised dataset).

***Selectivity is high, our probe is not memorising the data***

# Testing on Hard Examples

- Entities that are sometimes omitted, sometimes mentioned /or and sometimes distorted
- Permits testing whether probe classifies omissions/distortions/mentions or graph that contain specific entities

| Training Data | Test Data | % Data | F1 (B.Acc) |
|---|---|---|---|
| Manual-O | M&O | 13% | 0.81 (0.74) |
| Manual-D | M&D | 14% | 0.84 (0.81) |
| Manual-O+D | M&O&D | 9% | 0.78 (0.82) |
| Manual-O+D | M&O | 13% | 0.82 (0.82) |
| Manual-O+D | M&D | 14% | 0.78 (0.81) |

***The probe also performs well on difficult examples.***

# Generalising to Other RDF-to-Text Models

| | # T | # T(O) | # O |
|---|---|---|---|
| **WebNLG** | | | |
| Train | 36,704 | 7,064 (19%) | 7,824 |
| Dev | 4,658 | 882 (19%) | 993 |
| Test | 6,173 | 2,286 (37%) | 2,855 |
| **KELM** | 24,963 | 17,852 (72%) | 29,596 |
| **ALL** | 72,498 | 28,084(39%) | 41,268 |

| | **All** | **In Domain** | | | **OOD** | |
|---|---|---|---|---|---|---|
| | | W-T | W-D | W-S | W-U | K |
| **NP.P** | 0.89 | 0.84 | 0.84 | 0.88 | 0.81 | 0.91 |
| **P.P** | | | | | | |
| F1 | 0.8 | 0.84 | 0.83 | 0.79 | 0.7 | 0.78 |
| B.Acc | 0.85 | 0.88 | 0.88 | 0.83 | 0.77 | 0.81 |

## Probing T5

- T5 fine tuned on same data
- Automatic annotation of omissions
- Higher results than for BART

***In both cases, the embeddings of graphs leading to omissions differ from those that do not.***

# Summary

Omissions are frequent

The encoder plays a role in determining whether content is omitted in the output.

# Analysing Dialog Coherence and Cohesion

# Knowledge-Based Dialog

**Knowledge Graph**

(*Elsa Morante, place of birth, Rome*)
(*Elsa Morante, cause of death, myocardial infarction*)
(*Elsa Morante, spouse, Alberto Moravia*)
(*Elsa Morante, manner of death, natural causes*)
(*natural causes, inv. opposite of, unnatural death*)
(*Rome, inv. airline hub, Norwegian Air Shuttle*)
(*Rome, inv. enclave within, Vatican City*)
(*Rome, official language, Italian*)
(*Alberto Moravia, inv. founded by, Nuovi Argomenti*)
(*Alberto Moravia, place of death, Rome*)

**Dialog Context**

Where was Elsa Morante born?
Rome
What is Rome's administrative territory?
Vatican City
Who was Morante married to?
Alberto Moravia
Which communication medium was founded by Moravia?
Nuovi Argomenti
Where did Moravia take his last breath ?
Rome

**Generation**

Was Morante's death an accident or a suicide?
Natural causes

# Challenges

Dialog coherence

- Relevant turn (Content Selection)
- No repetition

Factuality

- Factually correct question (KB Fact)

Dialog Cohesion

- Appropriate anaphors

# Explainability by Design

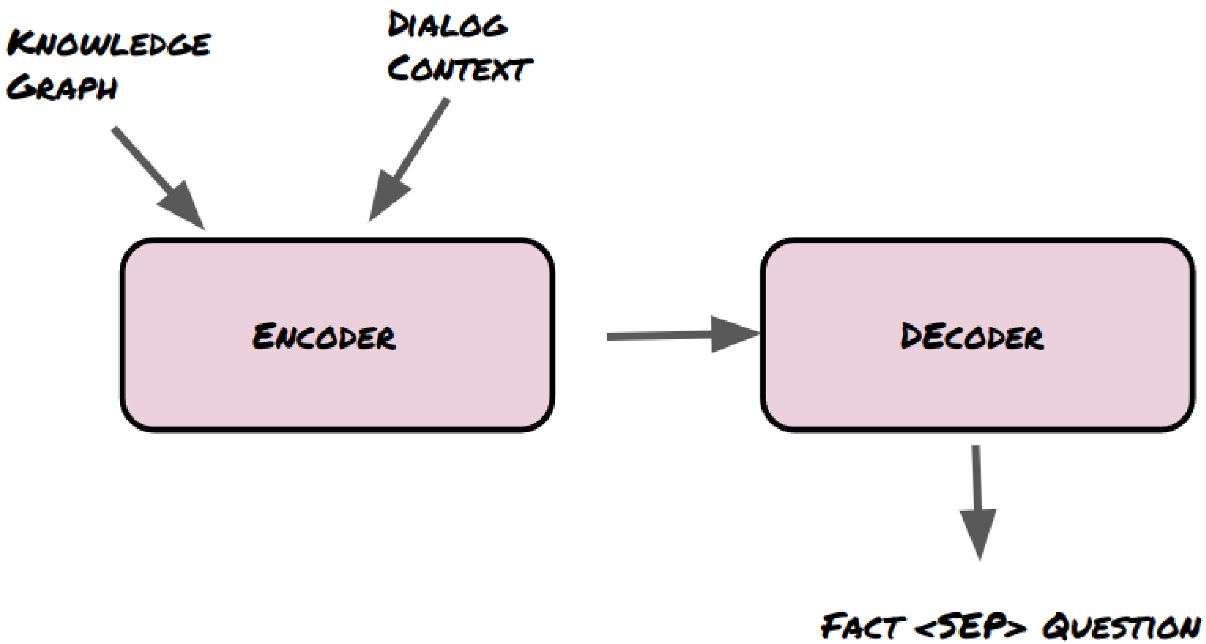
**Knowledge Graph**
(*Elsa Morante, place of birth, Rome*)
(*Elsa Morante, cause of death, myocardial infarction*)
(*Elsa Morante, spouse, Alberto Moravia*)
(*Elsa Morante, manner of death, natural causes*)
(*natural causes, inv. opposite of, unnatural death*)
(*Rome, inv. airline hub, Norwegian Air Shuttle*)
(*Rome, inv. enclave within, Vatican City*)
(*Rome, official language, Italian*)
(*Alberto Moravia, inv. founded by, Nuovi Argomenti*)
(*Alberto Moravia, place of death, Rome*)

**Dialog Context**
Where was Elsa Morante born?
Rome
What is Rome's administrative territory?
Vatican City
Who was Morante married to?
Alberto Moravia
Which communication medium was founded by Moravia?
Nuovi Argomenti
Where did Moravia take his last breath ?
Rome

**Generation**
(*Elsa Morante, cause of death, myocardial infarction*)
Was Morante's death an accident or a suicide?

# Knowledge-Guided Response Generation

T5 trained on KGConv dataset



# Analysing Generation

Factuality

- Is the ***predicted fact*** true (is it in the KB)?
- Does the question match the predicted fact
- If both are true the question is factual

Dialog Coherence

- Is the predicted fact different from those already predicted ? (New information)
- Is it relevant ? (Content Selection)

Dialog Cohesion

- Are pronouns correct and unambiguous ?
- Does the genre of the pronoun match that of the corresponding entity in the predicted triple ?
- Does the pronoun denote the last entity with matching genre ?

# The KGConv Dialogs

T (*Sitara Achakzai, field of work, feminism*)
Q `What was Sitara Achakzai's field of work?`
A `feminism`

T (*Sitara Achakzai, death manner, murder*)
Q `What was the cause of death of Achakzai?`
A `homicide`

T (*Sitara Achakzai, birthplace, Afghanistan*)
Q `Where was she born ?`
A `Afghanistan`

T (*Afghanistan, capital, Kabul*)
Q `What is the capital of Afghanistan?`
A `Kabul`

T (*Afghanistan, lowest point, Amu Darya*)
Q `What is the lowest point of Afghanistan?`
A `Amu Darya`

- 70,956 English Dialogs, 143K Wikidata triples
- Each dialog $D$ is associated with a Knowledge-Graph $K_D$
- A dialog is a sequence of question/response pairs
- Each question/response pair is grounded in a Wikidata fact

# Content Selection / Relevance

$K_D$ extended with three types of distractors

- Out-of-Scope triples (entity)
  - triples whose subject is of the same Wikidata category as the dialog root entity .
- Out-of-Scope triples (property)
  - triples whose property appears in $K_D$.
- Noise triples
  - Triples that are not in KGConv (and most of time not in Wikidata) but whose subject, property and object are in KGConv

# Dialog Context

4 types

- Natural Language only (NL)
- Triples only (KL)
- Natural Language Questions only (Q)
- NL + Triples (Hybrid)

# Analysing Coherence

| Context Type | $D_{QA_{nl}}$ | (%) | $D_{Q_{nl}}$ | (%) | $D_{kl}$ | (%) | $D_{QA_{nl+kl}}$ | (%) |
|---|---|---|---|---|---|---|---|---|
| # test examples | 313583 | | 321270 | | 315815 | | 313865 | |
| # distinct generated triples | 16519 | | 18146 | | 17875 | | 16597 | |
| **Correct triples** | 303723 | 97 | 286439 | 89 | 301970 | 96 | 304794 | 0,97 |
| Exact match with target | 123684 | 39 | 109031 | 34 | 123605 | 39 | 131453 | 42 |
| Other triple from input RDF | 180039 | 57 | 177408 | 55 | 178365 | 56 | 173341 | 55 |
| **Incorrect triples** | 9860 | 3 | 34831 | 11 | 13845 | 4 | 9071 | 3 |
| Repetitions | 1788 | 1 | 23149 | 7 | 1308 | 0 | 1705 | 1 |
| Out-of-scope (entity) triples | 305 | 0 | 640 | 0 | 340 | 0 | 398 | 0 |
| Out-of-scope (property) triples | 5327 | 2 | 6987 | 2 | 6437 | 2 | 5448 | 2 |
| Noise triples | | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Ill-formed triples | 460 | 0 | 2033 | 1 | 1663 | 1 | 710 | 0 |
| Triples not in KGConv | 5514 | 2 | 7403 | 2 | 7761 | 2 | 4977 | 2 |

***The model selects relevant facts***

- Few OOS and Noise triples (0-2%)

# Analysing Coherence

| Context Type | $D_{QA_{nl}}$ | (%) | $D_{Q_{nl}}$ | (%) | $D_{kl}$ | (%) | $D_{QA_{nl}+kl}$ | (%) |
|---|---|---|---|---|---|---|---|---|
| # test examples | 313583 | | 321270 | | 315815 | | 313865 | |
| # distinct generated triples | 16519 | | 18146 | | 17875 | | 16597 | |
| **Correct triples** | 303723 | 97 | 286439 | 89 | 301970 | 96 | 304794 | 0,97 |
| Exact match with target | 123684 | 39 | 109031 | 34 | 123605 | 39 | 131453 | 42 |
| Other triple from input RDF | 180039 | 57 | 177408 | 55 | 178365 | 56 | 173341 | 55 |
| **Incorrect triples** | 9860 | 3 | 34831 | 11 | 13845 | 4 | 9071 | 3 |
| Repetitions | 1788 | 1 | 23149 | 7 | 1308 | 0 | 1705 | 1 |
| Out-of-scope (entity) triples | 305 | 0 | 640 | 0 | 340 | 0 | 398 | 0 |
| Out-of-scope (property) triples | 5327 | 2 | 6987 | 2 | 6437 | 2 | 5448 | 2 |
| Noise triples | | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Ill-formed triples | 460 | 0 | 2033 | 1 | 1663 | 1 | 710 | 0 |
| Triples not in KGConv | 5514 | 2 | 7403 | 2 | 7761 | 2 | 4977 | 2 |

Content Selection

***The model selects relevant facts***

- Few OOS and Noise triples (0-2%)

***Some fake facts***

- Triples not in KGConv (2%)

# Analysing Coherence

| Context Type | $D_{QA_{nl}}$ | (%) | $D_{Q_{nl}}$ | (%) | $D_{kl}$ | (%) | $D_{QA_{nl}+kl}$ | (%) |
|---|---|---|---|---|---|---|---|---|
| # test examples | 313583 | | 321270 | | 315815 | | 313865 | |
| # distinct generated triples | 16519 | | 18146 | | 17875 | | 16597 | |
| **Correct triples** | 303723 | 97 | 286439 | 89 | 301970 | 96 | 304794 | 0,97 |
| Exact match with target | 123684 | 39 | 109031 | 34 | 123605 | 39 | 131453 | 42 |
| Other triple from input RDF | 180039 | 57 | 177408 | 55 | 178365 | 56 | 173341 | 55 |
| **Incorrect triples** | 9860 | 3 | 34831 | 11 | 13845 | 4 | 9071 | 3 |
| Repetitions | 1788 | 1 | 23149 | 7 | 1308 | 0 | 1705 | 1 |
| Out-of-scope (entity) triples | 305 | 0 | 640 | 0 | 340 | 0 | 398 | 0 |
| Out-of-scope (property) triples | 5327 | 2 | 6987 | 2 | 6437 | 2 | 5448 | 2 |
| Noise triples | | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Ill-formed triples | 460 | 0 | 2033 | 1 | 1663 | 1 | 710 | 0 |
| Triples not in KGConv | 5514 | 2 | 7403 | 2 | 7761 | 2 | 4977 | 2 |

***High relevance***

- Few incorrect triples for most models (3%)
- Answers matter: Q generates more incorrect triples (11%), often repeating previous turns

# Analysing Coherence

| Context Type | $D_{QA_{nl}}$ | (%) | $D_{Q_{nl}}$ | (%) | $D_{kl}$ | (%) | $D_{QA_{nl}+kl}$ | (%) |
|---|---|---|---|---|---|---|---|---|
| # test examples | 313583 | | 321270 | | 315815 | | 313865 | |
| # distinct generated triples | 16519 | | 18146 | | 17875 | | 16597 | |
| **Correct triples** | 303723 | 97 | 286439 | 89 | 301970 | 96 | 304794 | 0,97 |
| Exact match with target | 123684 | 39 | 109031 | 34 | 123605 | 39 | 131453 | 42 |
| Other triple from input RDF | 180039 | 57 | 177408 | 55 | 178365 | 56 | 173341 | 55 |
| **Incorrect triples** | 9860 | 3 | 34831 | 11 | 13845 | 4 | 9071 | 3 |
| Repetitions | 1788 | 1 | 23149 | 7 | 1308 | 0 | 1705 | 1 |
| Out-of-scope (entity) triples | 305 | 0 | 640 | 0 | 340 | 0 | 398 | 0 |
| Out-of-scope (property) triples | 5327 | 2 | 6987 | 2 | 6437 | 2 | 5448 | 2 |
| Noise triples | | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Ill-formed triples | 460 | 0 | 2033 | 1 | 1663 | 1 | 710 | 0 |
| Triples not in KGCONV | 5514 | 2 | 7403 | 2 | 7761 | 2 | 4977 | 2 |

High relevance

- Few incorrect triples for most models (3%)
- Answers matter: Q generates more incorrect triples (11%), often repeating previous turns

High Semantic Adequacy

- GLEU(Question,triple): 0.73 - 0.76

***Most questions are relevant and factual***

# Analysing Cohesion

Gender

- each RDF entity is associated its "sex or gender" value from Wikidata
- A pronoun in a generated question has the correct gender if its gender is the same as the gender of its referent, i.e. the subject entity of the triple the question is conditioned on.

Ambiguity

- A pronoun with genre $g$ is ambiguous if the last entity of genre $g$ mentioned in the dialog context is not the referent of that pronoun.

**Dialog Context**
T: (*NGC 2539*, *discoverer or inventor*, *William Herschel*)
Q: Who found NGC 2423?
A: William Herschel

T: (*NGC 2539*, *constellation*, *Puppis*)
Q: What is the name of the constellation which NGC 2423 belongs?
A: Puppis

T: (*William Herschel*, *student of*, *Nevil Maskelyne*))
Q: What was the name of Herschel's teacher?
A: Nevil Maskelyne

**Generation**
(*William Herschel*, *place of burial*, *Westminster Abbey*)
Where was **he** buried?

he → William Herschel

# Analysing Cohesion

| Context Type | $D_{QA_{nl}}$ | $D_{Q_{nl}}$ | $D_{kl}$ | $D_{QA_{nl}+kl}$ |
|---|---|---|---|---|
| questions with a pronoun | 9% | 8% | 13% | 8% |
| "he" | 53% | 47% | 54% | 52% |
| "it" | 32% | 35% | 34% | 35% |
| "him" | 7% | 10% | 8% | 7% |
| "she" | 8% | 7% | 3% | 6% |
| "her" | <1% | 1% | 4% | <1% |
| pronouns with gender mistakes | 5% | 5% | 3% | 4% |
| "he" | 29% | 44% | 68% | 52 %% |
| "she" | 62 % | 39% | 18% | 34% |
| "him" | 4 % | 9 % | 9 % | 8% |
| "her" | 3% | 5 % | 2 % | 2% |
| "it" | 2% | 3% | 3 % | 4 % |
| ambiguous pronouns | 30% | 36% | 34% | 29% |
| "it" | 64% | 67% | 76% | 66% |
| "he" | 18% | 19% | 15% | 21% |
| "she" | 14% | 9% | 4% | 9% |
| "him" | 3% | 4% | 4% | 3% |
| "her" | 1% | 1% | 1% | 1% |
| pronominalized distinct triples | 22% | 19% | 24% | 19% |

- Good proportion of questions containing pronouns (between 8 and 13% of the test examples)
- The KL context induces a much higher rate of pronouns
- Strong bias for masculine pronouns

# Analysing Cohesion

| Context Type | $D_{QA_{nl}}$ | $D_{Q_{nl}}$ | $D_{kl}$ | $D_{QA_{nl}+kl}$ |
|---|---|---|---|---|
| questions with a pronoun | 9% | 8% | 13% | 8% |
| "he" | 53% | 47% | 54% | 52% |
| "it" | 32% | 35% | 34% | 35% |
| "him" | 7% | 10% | 8% | 7% |
| "she" | 8% | 7% | 3% | 6% |
| "her" | <1% | 1% | 4% | <1% |
| pronouns with gender mistakes | 5% | 5% | 3% | 4% |
| "he" | 29% | 44% | 68% | 52 %% |
| "she" | 62 % | 39% | 18% | 34% |
| "him" | 4 % | 9 % | 9 % | 8% |
| "her" | 3% | 5 % | 2 % | 2% |
| "it" | 2% | 3% | 3 % | 4 % |
| ambiguous pronouns | 30% | 36% | 34% | 29% |
| "it" | 64% | 67% | 76% | 66% |
| "he" | 18% | 19% | 15% | 21% |
| "she" | 14% | 9% | 4% | 9% |
| "him" | 3% | 4% | 4% | 3% |
| "her" | 1% | 1% | 1% | 1% |
| pronominalized distinct triples | 22% | 19% | 24% | 19% |

- Good diversity of the triples giving rise to pronominal questions (about 2% of the dataset triples).

# Analysing Cohesion

| Context Type | $D_{QA_{nl}}$ | $D_{Q_{nl}}$ | $D_{kl}$ | $D_{QA_{nl}+kl}$ |
|---|---|---|---|---|
| questions with a pronoun | 9% | 8% | 13% | 8% |
| "he" | 53% | 47% | 54% | 52% |
| "it" | 32% | 35% | 34% | 35% |
| "him" | 7% | 10% | 8% | 7% |
| "she" | 8% | 7% | 3% | 6% |
| "her" | <1% | 1% | 4% | <1% |
| pronouns with gender mistakes | 5% | 5% | 3% | 4% |
| "he" | 29% | 44% | 68% | 52 %% |
| "she" | 62 % | 39% | 18% | 34% |
| "him" | 4 % | 9 % | 9 % | 8% |
| "her" | 3% | 5 % | 2 % | 2% |
| "it" | 2% | 3% | 3 % | 4 % |
| ambiguous pronouns | 30% | 36% | 34% | 29% |
| "it" | 64% | 67% | 76% | 66% |
| "he" | 18% | 19% | 15% | 21% |
| "she" | 14% | 9% | 4% | 9% |
| "him" | 3% | 4% | 4% | 3% |
| "her" | 1% | 1% | 1% | 1% |
| pronominalized distinct triples | 22% | 19% | 24% | 19% |

- Antecedent/Pronoun Genre agreement is high (95%-96%)

# Analysing Cohesion

| Context Type | $D_{QA_{nl}}$ | $D_{Q_{nl}}$ | $D_{kl}$ | $D_{QA_{nl+kl}}$ |
|---|---|---|---|---|
| questions with a pronoun | 9% | 8% | 13% | 8% |
| "he" | 53% | 47% | 54% | 52% |
| "it" | 32% | 35% | 34% | 35% |
| "him" | 7% | 10% | 8% | 7% |
| "she" | 8% | 7% | 3% | 6% |
| "her" | <1% | 1% | 4% | <1% |
| pronouns with gender mistakes | 5% | 5% | 3% | 4% |
| "he" | 29% | 44% | 68% | 52 %% |
| "she" | 62 % | 39% | 18% | 34% |
| "him" | 4 % | 9 % | 9 % | 8% |
| "her" | 3% | 5 % | 2 % | 2% |
| "it" | 2% | 3% | 3 % | 4 % |
| ambiguous pronouns | 30% | 36% | 34% | 29% |
| "it" | 64% | 67% | 76% | 66% |
| "he" | 18% | 19% | 15% | 21% |
| "she" | 14% | 9% | 4% | 9% |
| "him" | 3% | 4% | 4% | 3% |
| "her" | 1% | 1% | 1% | 1% |
| pronominalized distinct triples | 22% | 19% | 24% | 19% |

- The proportion of ambiguous pronouns is quite high, ranging between 29% and 36%

# Ablating the Knowledge Graph

| Context Type | $D_{QA_{nl}}$ | $D_{Q_{nl}}$ | $D_{kl}$ | $D_{QA_{nl}+kl}$ |
|---|---|---|---|---|
| # Test examples | 323k | 302k | 323k | 323k |
| Incorrect triple | 92% | 92% | 91% | 91% |
| Repetition | 2% | 1% | 2% | 1% |
| Triple not in KGCONV | 84% | 81% | 83% | 82% |
| Subject not in KGCONV | 13% | 28% | 17% | 15% |
| Property not in KGCONV | 14% | 33% | 17% | 16% |
| Object not in KGCONV | 13% | 29% | 17% | 15% |

***Conditioning question generation not only on the dialog context but also on a knowledge graph helps generating factually correct dialogs***

- 91%-92% of generated triples are incorrect
- Almost all of them (81-84%) are hallucinated triples not belonging to the set of KGConv triples, a large set of 132K Wikidata triples.

.

# Ablating the Dialog Context

| | # | % |
|---|---|---|
| # test examples | 323765 | |
| Correct triples | 166716 | 51 |
| Exact match with target | 36474 | 11 |
| Other triple from input RDF | 130242 | 40 |
| Incorrect triples | 157049 | 49 |
| Repetitions | 149363 | 46 |
| Out-of-scope (entity) triples | 327 | 0 |
| Out-of-scope (property) triples | 8713 | 3 |
| Noise triples generated | 0 | 0 |
| Ill-formed triples | 182 | 0 |
| Triples with a property not in KGConv | 6989 | 2 |

Unsurprisingly, ablating the dialog context

- drastically reduces the proportion of correct triples (51%) and
- increases the ratio of repetitions (46%).

.

# Conclusion

Like hallucinations, omissions impact seamntic adequacy

- More work is need to identify, quantify and explain omissions in other generation tasks and for other languages

Grounding Dialog Models in Knowledge helps getting a detailed picture of their coherence, factuality and cohesion

- Can the approach be extended to more complex questions, to other languages andto open domain dialogs ?

Questions ?