R² Data Labs | CITY UNIVERSITY OF LONDON EST 1894

Rolls-Rolls & City, University of London Internship Project

Converting Natural Language Question to Graph Query:

Data generation & knowledge-based question answering (NL2GQ) systems in aerospace sector

https://www.rolls-royce.com/products-and-services/r2datalabs/ecosystem.aspx

Student: **Junru Xiong** | 27th, Sep, 2021

Supervisors from Rolls-Royce: **Muhannad Alomari** and **Dionisios Korovilas** | Supervisor from City University: **Tillman Weyde**

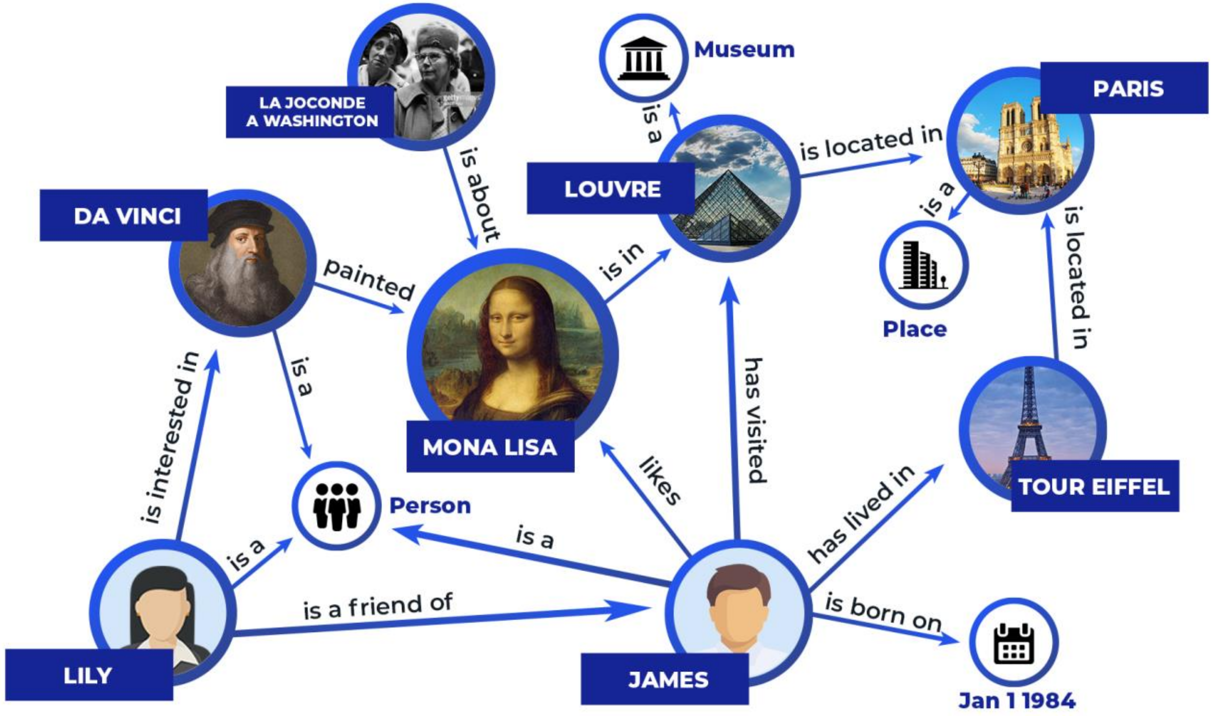
Background & objectives

Background, problem statement and goal:

- **Knowledge graph** (KG) makes engineers easier to **explore information**.
- The **barrier** to queries KG remains **high.**
- The **lack of domain dataset**.
- Need **quick access** to this resource.

Objectives:

- **Generating** question, answer and graph query **datasets** and **filtering** for further models (NL2GQ) training.
- Translate **free text questions into graph query language** (NL2GQ).

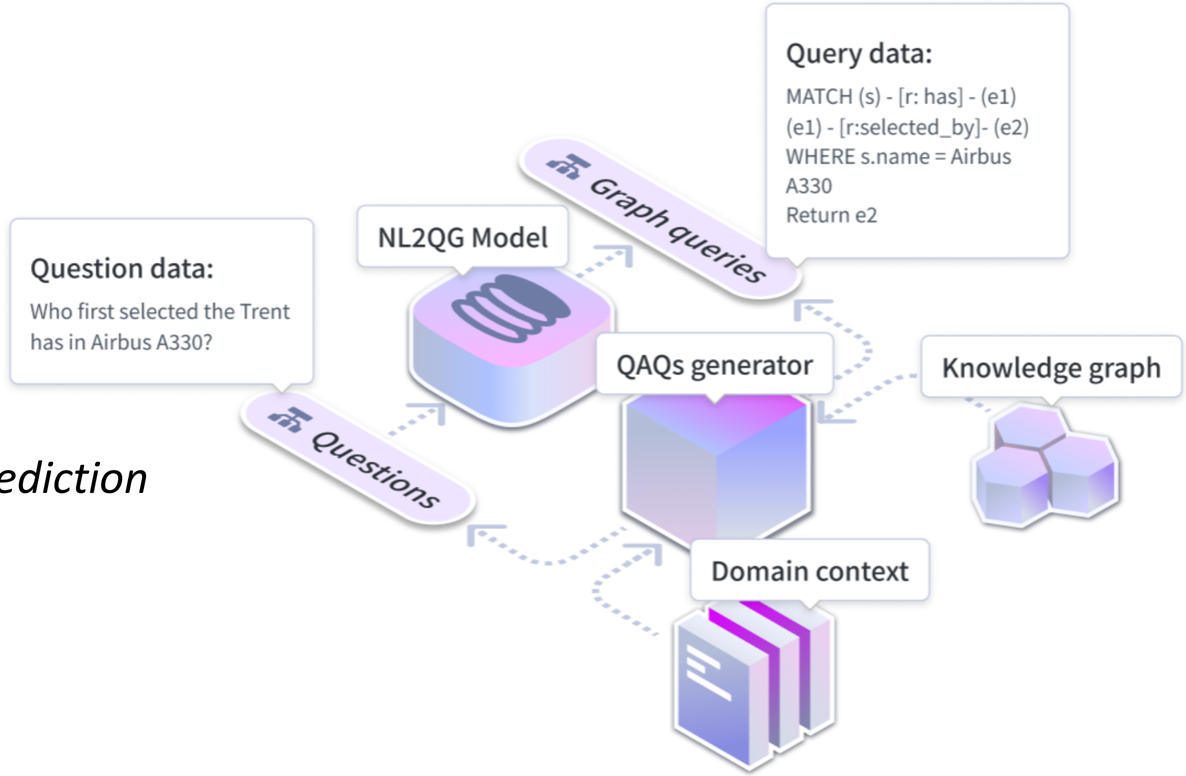


An example knowledge graph
(https://yashuseth.blog/2019/10/08/introduction-question-answering-knowledge-graphs-kgqa/)

Outline

1. **Aerospace question-answer-query pairs (QAQs) generation**
    1) *Raw dataset overview*
    2) *Pipeline overview*
    3) *Generated QAQs result table*
    4) *Filtering strategy*
    5) *Brief summary*

2. **Natural language to graph query (NL2GQ)**
    1) *Dataset overview*
    2) *Pipeline overview*
    3) *Name-entity recognition (NER) & One-hop relation prediction models*
    4) *Simple Multi-hop relation prediction models*
    5) *Models Learning*
    6) *Key metrics & summary*
    7) *Video showcase*



Query data:

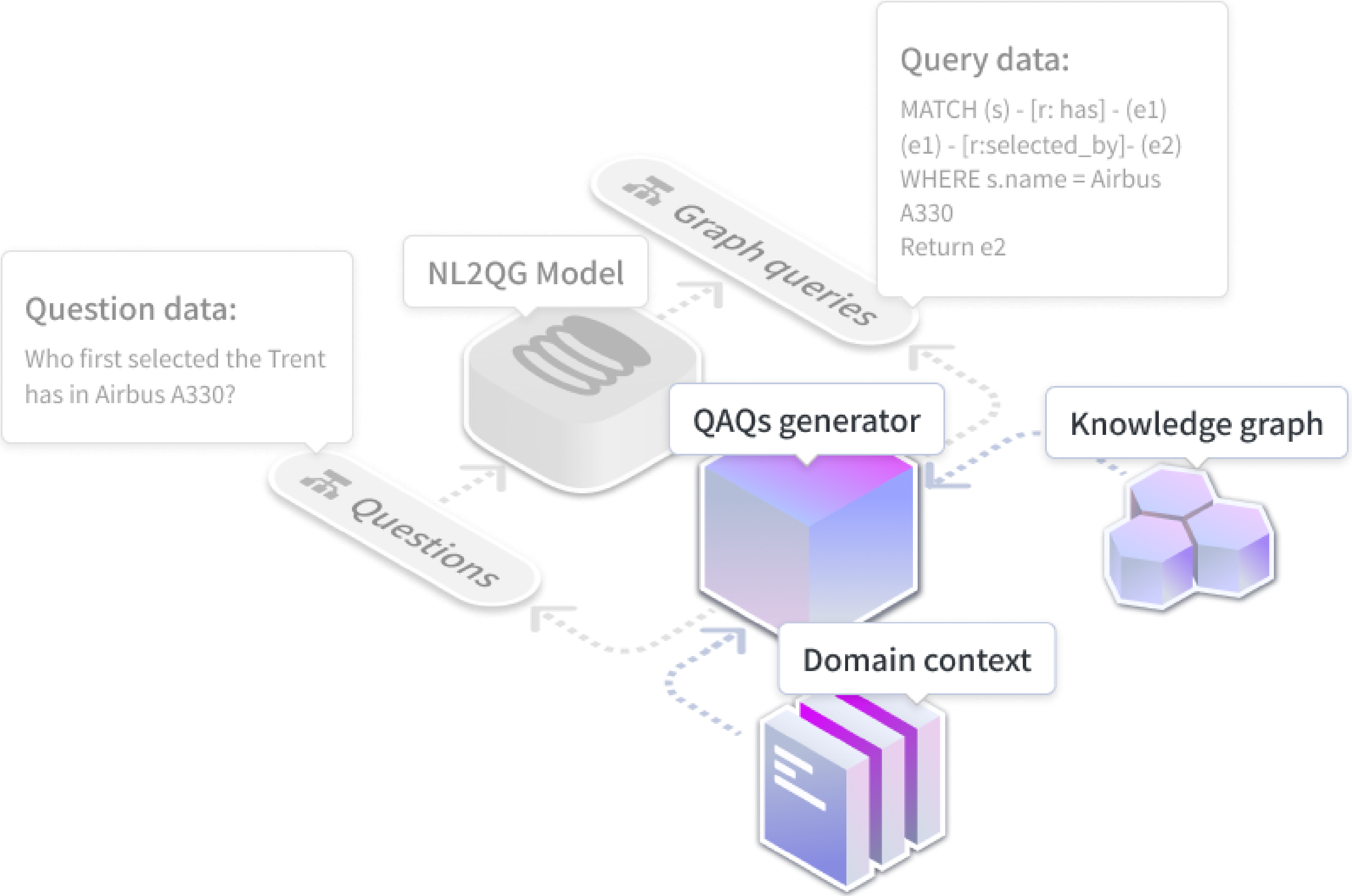
```
MATCH (s) - [r: has] - (e1)
(e1) - [r:selected_by]- (e2)
WHERE s.name = Airbus
A330
Return e2
```

Question data:

Who first selected the Trent has in Airbus A330?

Aerospace question-answer-query pairs (QAQs) generation

Where are the datasets for training NL2QG supervise learning models?



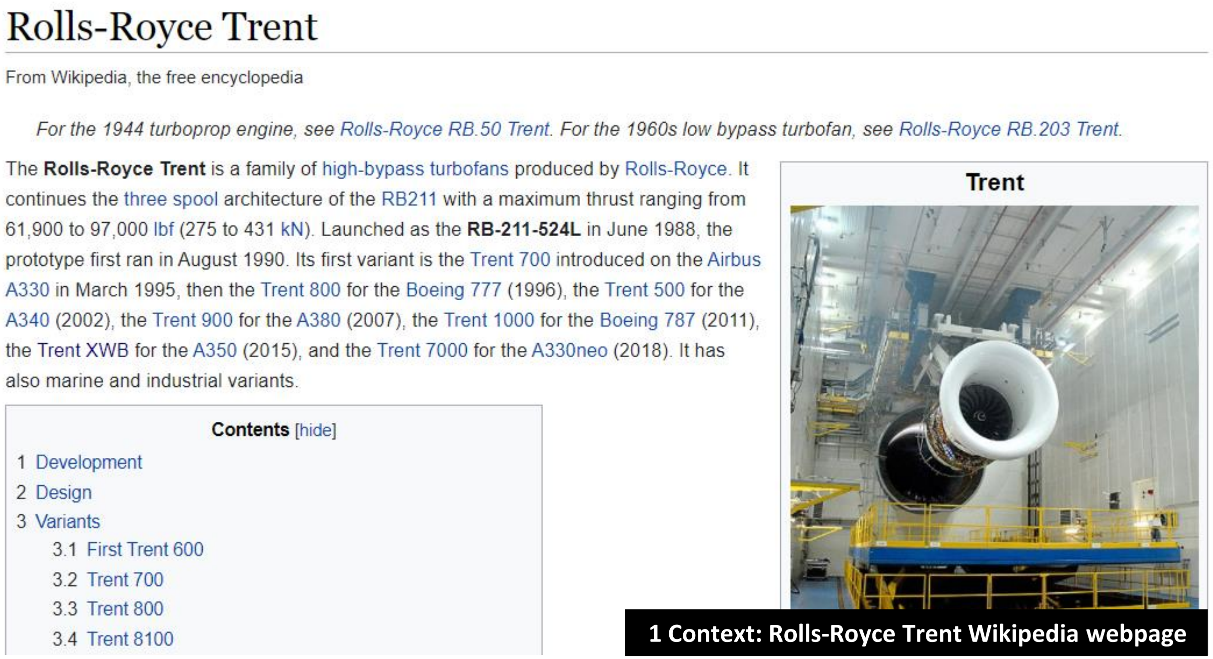
Query data:

MATCH (s) - [r: has] - (e1)
(e1) - [r:selected_by]- (e2)
WHERE s.name = Airbus
A330
Return e2

Question data:

Who first selected the Trent has in Airbus A330?

1.1 Datasets (context and KG) , and T5 (question generator) model for QAQs generation:



Rolls-Royce Trent

From Wikipedia, the free encyclopedia

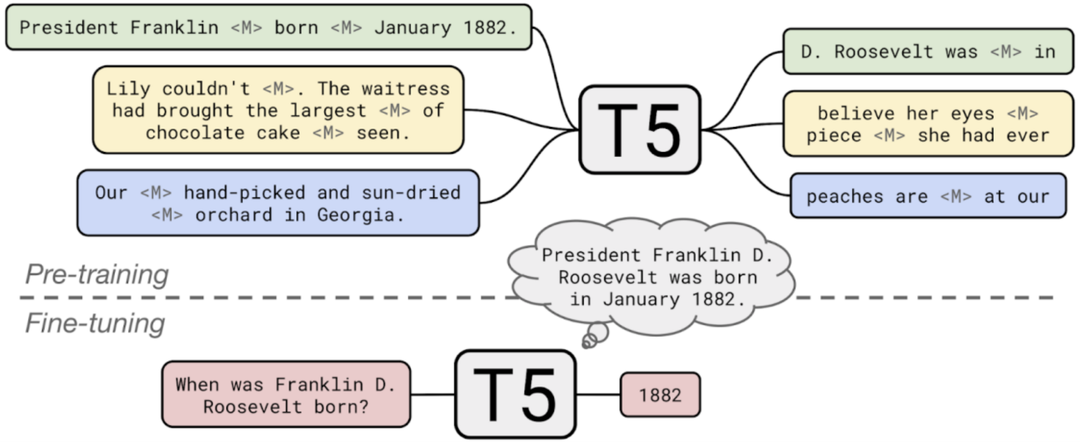
For the 1944 turboprop engine, see Rolls-Royce RB.50 Trent. For the 1960s low bypass turbofan, see Rolls-Royce RB.203 Trent.

The **Rolls-Royce Trent** is a family of high-bypass turbofans produced by Rolls-Royce. It continues the three spool architecture of the RB211 with a maximum thrust ranging from 61,900 to 97,000 lbf (275 to 431 kN). Launched as the **RB-211-524L** in June 1988, the prototype first ran in August 1990. Its first variant is the Trent 700 introduced on the Airbus A330 in March 1995, then the Trent 800 for the Boeing 777 (1996), the Trent 500 for the A340 (2002), the Trent 900 for the A380 (2007), the Trent 1000 for the Boeing 787 (2011), the Trent XWB for the A350 (2015), and the Trent 7000 for the A330neo (2018). It has also marine and industrial variants.

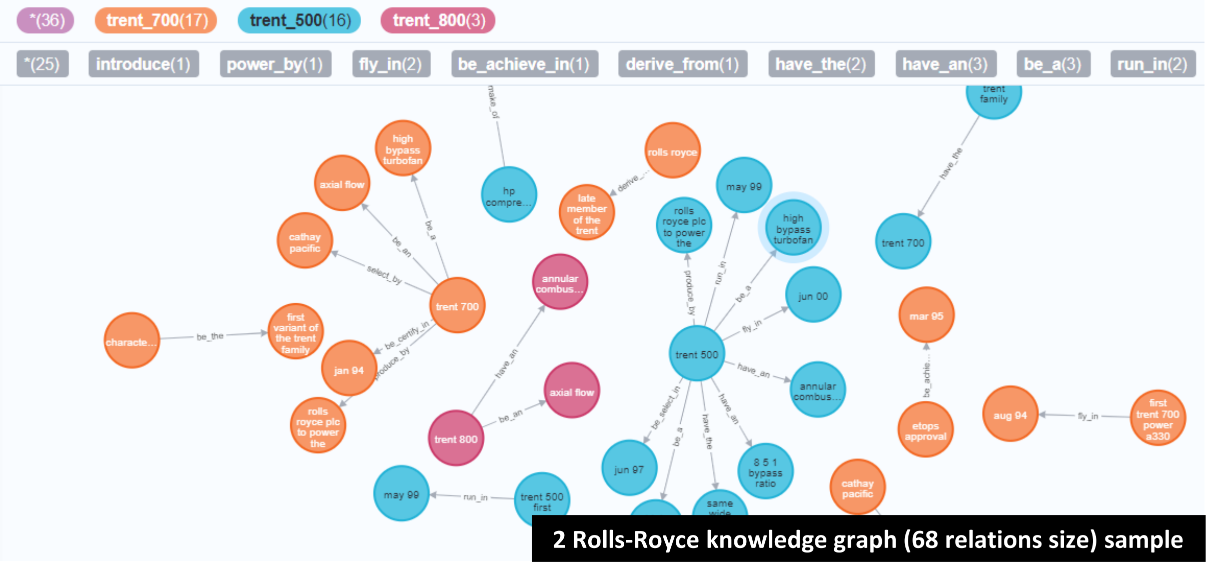
Contents [hide]

1 Development
2 Design
3 Variants
3.1 First Trent 600
3.2 Trent 700
3.3 Trent 800
3.4 Trent 8100

1 Context: Rolls-Royce Trent Wikipedia webpage


President Franklin <M> born <M> January 1882.
Lily couldn't <M>. The waitress had brought the largest <M> of chocolate cake <M> seen.
Our <M> hand-picked and sun-dried <M> orchard in Georgia.
T5
D. Roosevelt was <M> in
believe her eyes <M> piece <M> she had ever
peaches are <M> at our
Pre-training
President Franklin D. Roosevelt was born in January 1882.
Fine-tuning
When was Franklin D. Roosevelt born?
T5
1882

3 Fine-tuned T5 model for question prediction from the given context



2 Rolls-Royce knowledge graph (68 relations size) sample

The T5 Model

- **Text-to-Text Transfer Transformer** NLP model
- **Pre-trained** by masked language modelling using Colossal Clean **Crawled Corpus** (C4) dataset
- Our T5 question generator is **fine tuned** on **SQUAD** (Wikipedia-based) question answering dataset

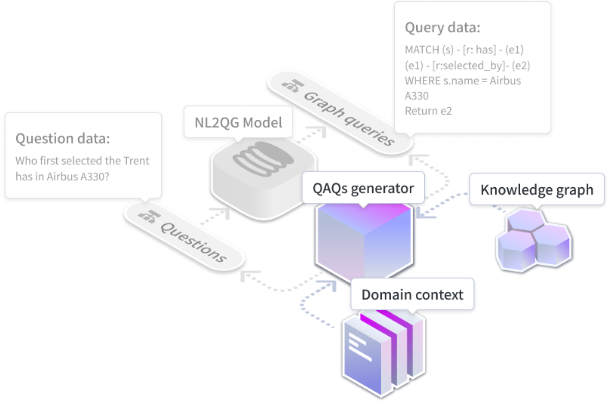
1.2 Question, answer and query generation pipeline:

Semantic matching

- Use **fuzzy search** to match entities in KG and entities in context.

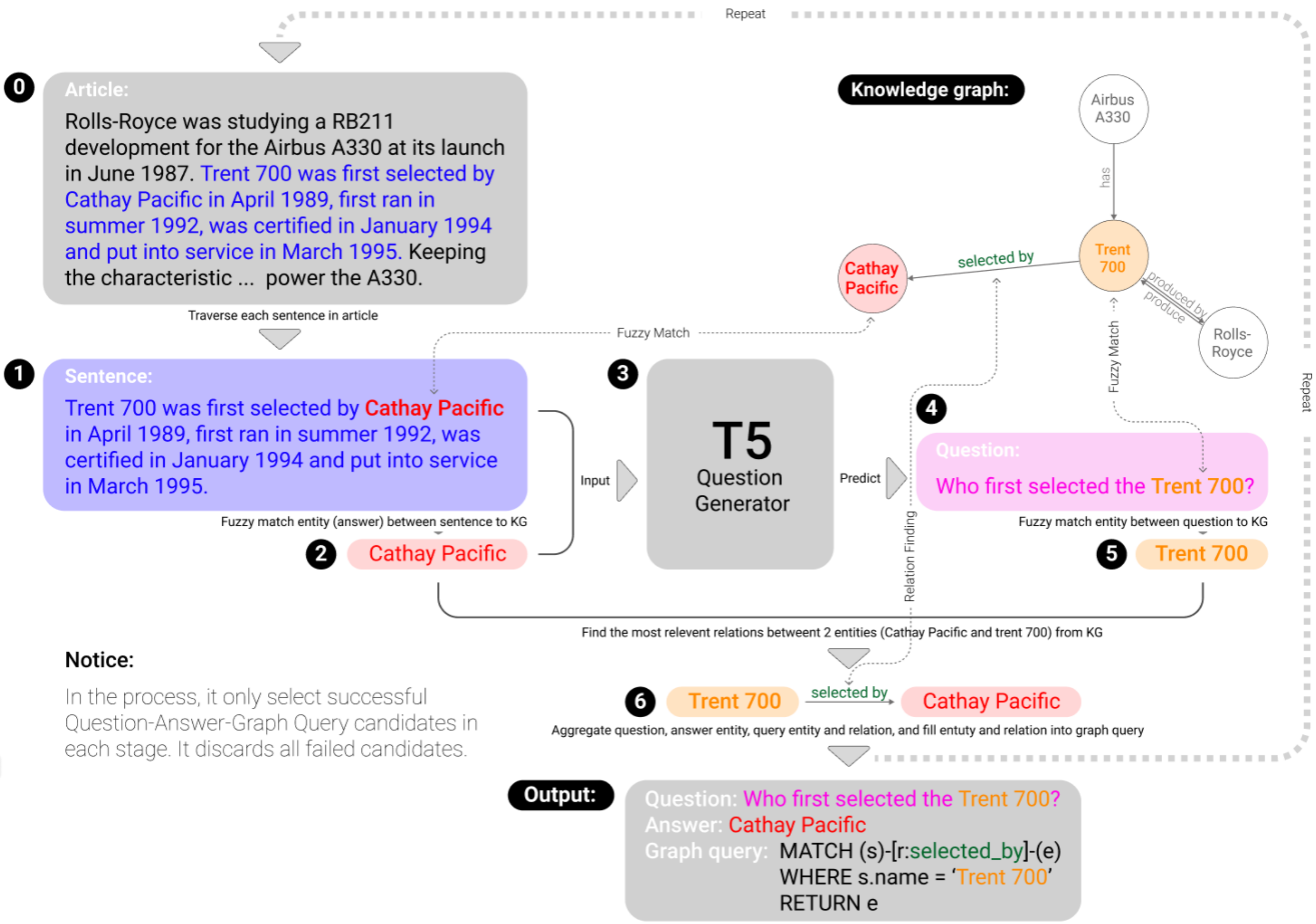
Many candidate relation edges in step 6?

- **Word embedding** to translate list of relations to vectors
- Compute their **cosine similarity**
- Choose the **highest score** relation edges



Query data:

```
MATCH (s) - [r: has] - (e1)
(e1) - [r:selected_by]- (e2)
WHERE s.name = Airbus
A330
Return e2
```



Article: Rolls-Royce was studying a RB211 development for the Airbus A330 at its launch in June 1987. Trent 700 was first selected by Cathay Pacific in April 1989, first ran in summer 1992, was certified in January 1994 and put into service in March 1995. Keeping the characteristic ... power the A330.

Sentence: Trent 700 was first selected by Cathay Pacific in April 1989, first ran in summer 1992, was certified in January 1994 and put into service in March 1995.

Fuzzy match entity (answer) between sentence to KG

Cathay Pacific

T5 Question Generator

Question: Who first selected the Trent 700?

Fuzzy match entity between question to KG

Trent 700

Find the most relevent relations between 2 entities (Cathay Pacific and trent 700) from KG

Trent 700 —selected by→ Cathay Pacific

Aggregate question, answer entity, query entity and relation, and fill entuty and relation into graph query

Output:

Question: Who first selected the Trent 700?

Answer: Cathay Pacific

Graph query:

```
MATCH (s)-[r:selected_by]-(e)
WHERE s.name = 'Trent 700'
RETURN e
```

Notice:

In the process, it only select successful Question-Answer-Graph Query candidates in each stage. It discards all failed candidates.

1.3 Generated questions, answers and queries results overview:

context	predicted questions (from lemmatized context)	predicted questions (from no lemmatized context)	answer entity	query entity	relations	Is_valid
the rolls royce trent be a family of high bypass turbofan produce by rolls royce	What type of turbofan is the rolls royce trent?	What type of turbofan is the rolls royce trent?	high bypass turbofan	rolls royce		NO
despite share the trent 700 hp system and ip turbine the overall pressure ratio of the trent 800 be high by increase the capacity of the ip compressor and the lp turbine	What is the hp system of the trent 800?	What HP system does the Trent 800 share?	trent 700	trent 800		NO
the trent 700 be first select by cathay pacific in april 1989 first run in summer 1992 be certify in january 1994 and put into service in march 1995	What was the first cathay pacific aircraft to be selected?	What was the first cathay pacific aircraft to be selected?	trent 700	cathay pacific		NO
the trent 700 was first select by cathay pacific in april 1989, first run in summer 1992 was certify in january 1994 and put into service in march 1995	Who first selected the trent 700?	Who selected the trent 700?	cathay pacific	trent 700	select by	YES
the trent 800 have the trent family three shaft architecture with a 280 cm 110 in fan	What is the trent 800's architecture?	What type of architecture is the trent 800?	trent family three shaft architecture	trent 800	have the	YES
Keeping the three spool architecture of the trent family, it has the trent 700 s 2.47 m (97.5 in) fan and a Trent 800 core scaled down	What is the name of the trent family?	What is the name of the spool fan that is used in the trent family?	trent 700	trent family	have the	YES
the rolls royce trent 7000 power exclusively the airbus a330neo	What is the name of the roll royce?	What is the name of the other model that shares components with the trent 700?	trent 700	rolls royce		NO
On 17 January 2008, a British Airways Boeing 777-236ER, operating as BA038 from Beijing to London, crash-landed at Heathrow after both Trent 800 engines lost power during the aircraft's final approach.		Where did the Birtish Airways Boeing 777 crash land?	heathrow	Birtish Airways		NO
singapore airlines have 58 trent 800 power 777s and 5 trent 500 powered a340 500s it also have a further 19 trent 700 power a330 300s 19 trent 900 power a380 800s and 20 trent xwb power a350 xwb 900s on order	What airline has 58 trent 800 power 777s?	What airline has 58 trent 800 power 777s?	singapore airlines	trent 800		NO
rolls royce plc	What is the name of the company that owns the roll royce plc?		rolls royce plc	rolls royce		NO
trent 700 series	What is the name of the series of cars that are part of the trent family?	What is the name of the series of cars that are part of the trent family?	trent 700	trent family	have the	NO

1.4 Filtering and result:

Filtering strategy

- Limit **minimum length** of context (e.g., Context must be a complete sentence) **&**
- ‘Answer entity’ and ‘query entity’ **must** have **linked relations** in knowledge graph **&**
- ‘Answer entity’ and ‘query entity’ **must appear in context at the same time**

(If someone has good idea for filtering, welcome to discuss)

(The process speed of entire pipeline is **1-2s per** question, answer and query pair)

context	predicted questions (from no lemmatized context)	answer entity	query entity	relations	graph query

the **trent 700** was first select by **cathay pacific** in april 1989, first run in summer 1992 was certify in january 1994 and put into service in march 1995	Who selected the **trent 700**?	**cathay pacific**	**trent 700**	select by	MATCH (s)-[r:selected_by]-(e)
					WHERE s.name = ‘Trent 700’
					RETURN e

the **trent 800** have the **trent family three shaft architecture**, with a 280 cm (110 in) fan	What type of architecture is the **trent 800**?	**trent family three shaft architecture**	**trent 800**	have the	MATCH (s)-[r:have_the]-(e)
					WHERE s.name = ‘Trent 800’
					RETURN e

Keeping the three spool architecture of the **trent family**, it has the **trent 700** s 2.47 m (97.5 in) fan and a Trent 800 core scaled down	What is the name of the spool fan that is used in the **trent family**?	**trent 700**	**trent family**	have the	MATCH (s)-[r:have_the]-(e)
					WHERE s.name = ‘Trent family’
					RETURN e

(Filtering results, it looks acceptable)

1.5 Question, answer and query pairs generation brief summary:

Summary

Pros

- An **end-to-end** way to **automatically** generate QAQ pairs from text and knowledge base.
- It can used for **general QA** or **KBQA** training or
- Can be applied in any **other data** generation in RR.

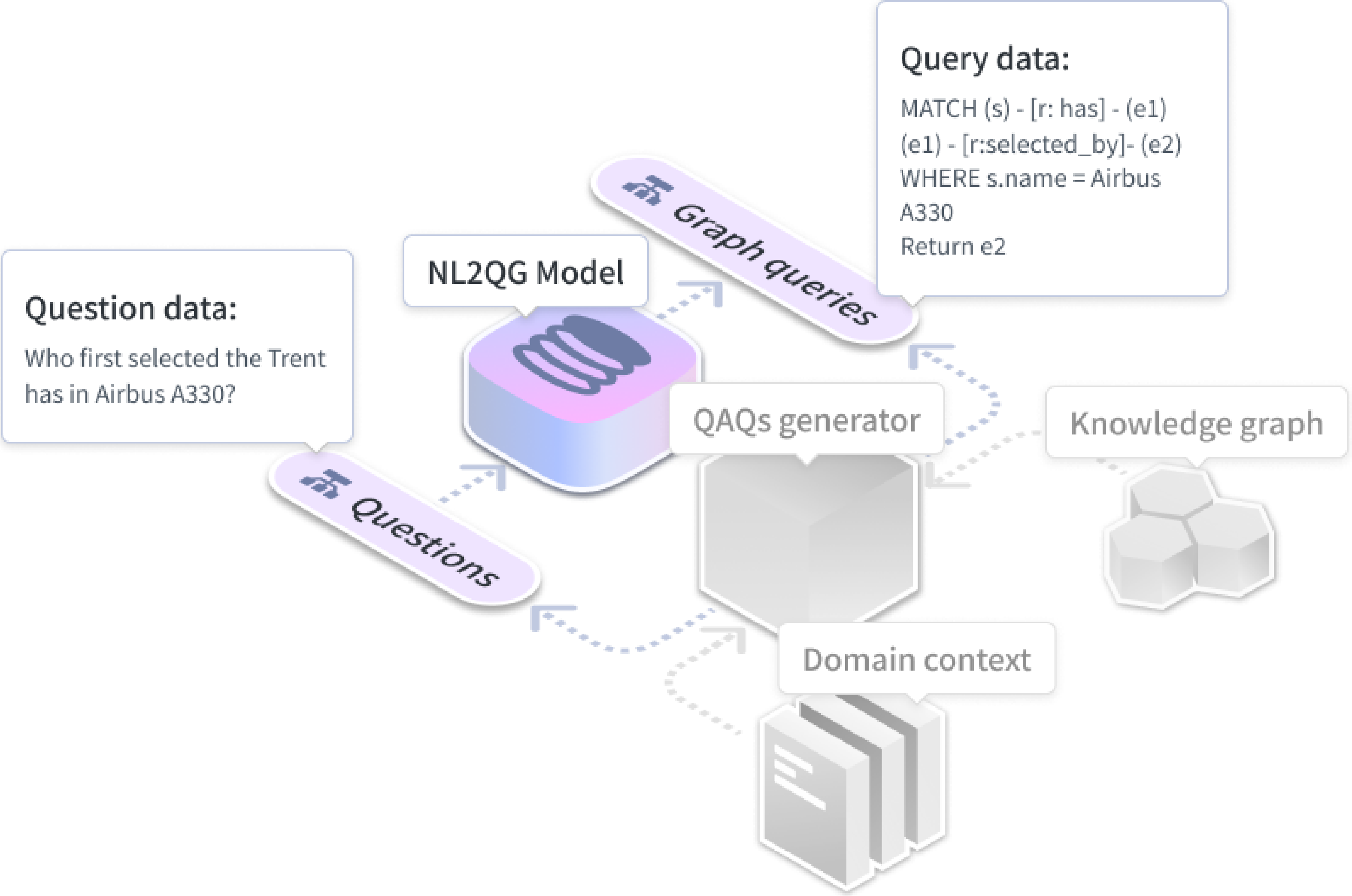
Cons

- Generated questions may **not be perfect** ,and somehow need **human judgement**
- It only able to generate **simple multi-hop graph query**.

context	predicted questions (from no lemmatized context)	answer entity	query entity	relations	graph query
the **trent 700** was first select by **cathay pacific** in april 1989, first run in summer 1992 was certify in january 1994 and put into service in march 1995	Who selected the **trent 700**?	**cathay pacific**	**trent 700**	select by	MATCH (s)-[r:selected_by]-(e) WHERE s.name = ‘Trent 700’ RETURN e
the **trent 800** have the **trent family three shaft architecture**, with a 280 cm (110 in) fan	What type of architecture is the **trent 800**?	**trent family three shaft architecture**	**trent 800**	have the	MATCH (s)-[r:have_the]-(e) WHERE s.name = ‘Trent 800’ RETURN e
Keeping the three spool architecture of the **trent family**, it has the **trent 700** s 2.47 m (97.5 in) fan and a Trent 800 core scaled down	What is the name of the spool fan that is used in the **trent family**?	**trent 700**	**trent family**	have the	MATCH (s)-[r:have_the]-(e) WHERE s.name = ‘Trent family’ RETURN e

Natural question to graph query (NL2GQ)

An knowledge-based question answering system



Question data:

Who first selected the Trent has in Airbus A330?

Query data:

```
MATCH (s) - [r: has] - (e1)
(e1) - [r:selected_by]- (e2)
WHERE s.name = Airbus
A330
Return e2
```

2.1 dataset overview:

SimpleQuestions – A Knowledge-based Question Answering open-domain dataset

SimpleQuestions consists **108,442 questions** written in natural language, formatted as (**subject, relationship, object**)

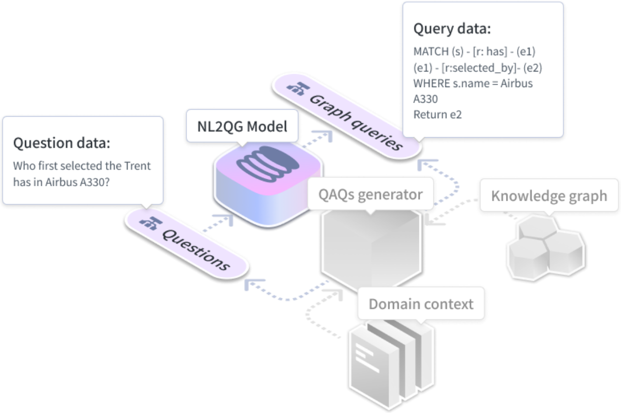
question	query entity_subject	subject_id	relationship	object_id
what is the book e about	e	04whkz5	book/written_work/subjects	01cj3p
to what release does the release track cardiac arrest come from	cardiac arrest	0tp2p24	music/release_track/release	0sjc7c1
what country was the film the debt from	the debt	04j0t75	film/film/country	07ssc
what songs have nobuo uematsu produced	nobuo uematsu	0ftqr	music/producer/tracks_produced	0p600l
who produced eve olution	eve olution	036p007	music/release/producers	0677ng
which artist recorded most of us are sad	most of us are sad	0ms5mg	music/recording/artist	0mjn2
what movie is produced by warner bros	warner bros	086k8	film/production_company/films	0278x5r
what is don graham known as	don graham	02vnx8y	common/topic/notable_types	01xljrf
what 's there to see in columbus	columbus	01smm	travel/travel_destination/tourist_attractions	0328cp
what album was tibet released on	tibet	0mgb6cl	music/release_track/release	0f4zk3j
who is a musician born in detroit	detroit	02dtg	location/location/people_born_here	01s8mcb
which city did the artist ryna originate in	ryna	0275d7v	music/artist/origin	052bw

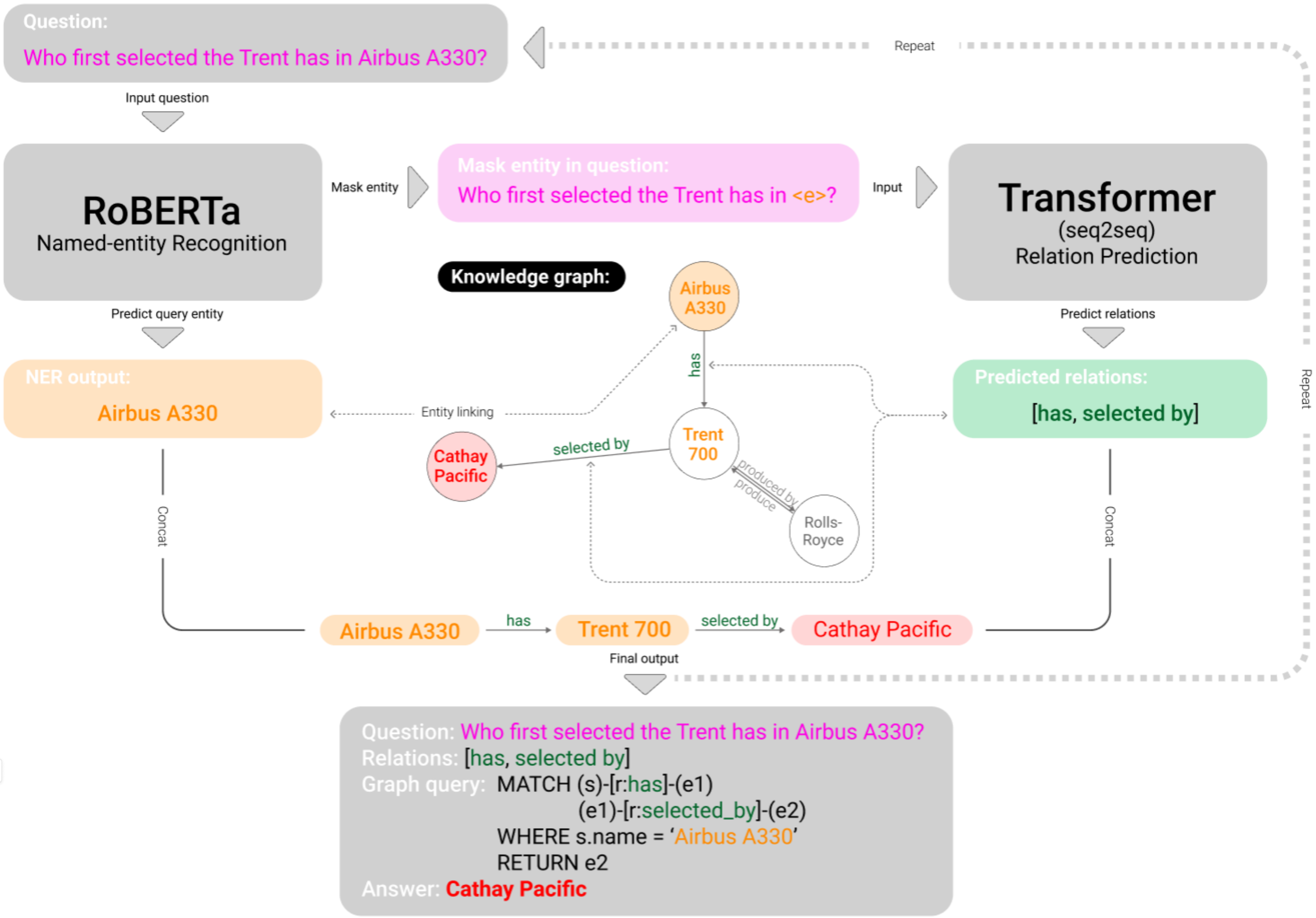
(https://github.com/davidgolub/SimpleQA/tree/master/datasets/SimpleQuestions)

2.2 Natural language question to graph query Pipeline:

How Entity Linking works?

Use lemmatized **exact search** to match entities in the knowledge graph and NER output entities.


Query data:
MATCH (s) - [r: has] - (e1)
(e1) - [r:selected_by]- (e2)
WHERE s.name = Airbus A330
Return e2
Graph queries
NL2QG Model
Question data:
Who first selected the Trent has in Airbus A330?
Questions
QAQs generator
Knowledge graph
Domain context



Question: Who first selected the Trent has in Airbus A330?
Relations: [has, selected by]
Graph query:

```
MATCH (s)-[r:has]-(e1)
          (e1)-[r:selected_by]-(e2)
WHERE s.name = 'Airbus A330'
RETURN e2
```

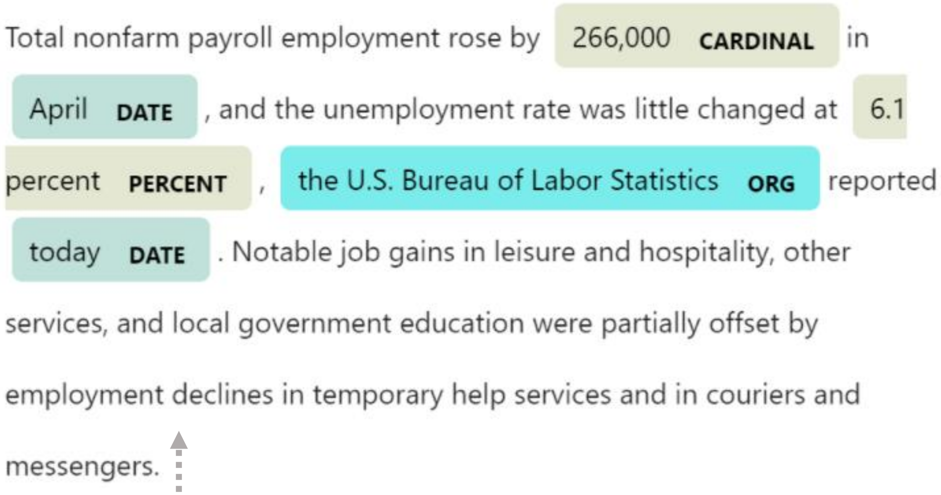
Answer: **Cathay Pacific**

2.3 NER and One-hop relation prediction models:

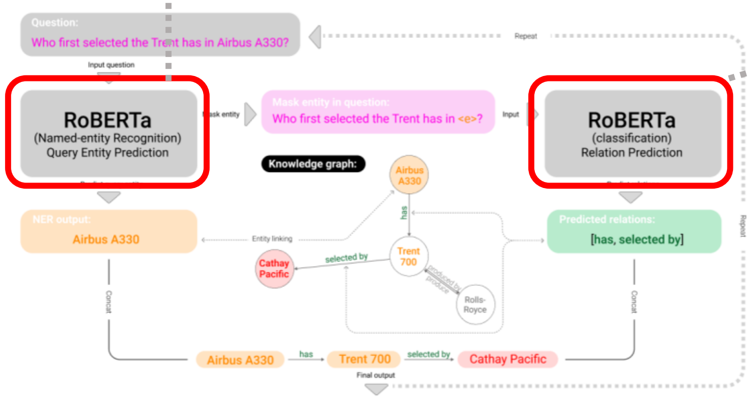
RoBERTa: A robustly optimized **pre-trained model** for pretraining natural language processing (NLP) systems that **improves on BERT**.

Downstream tasks:

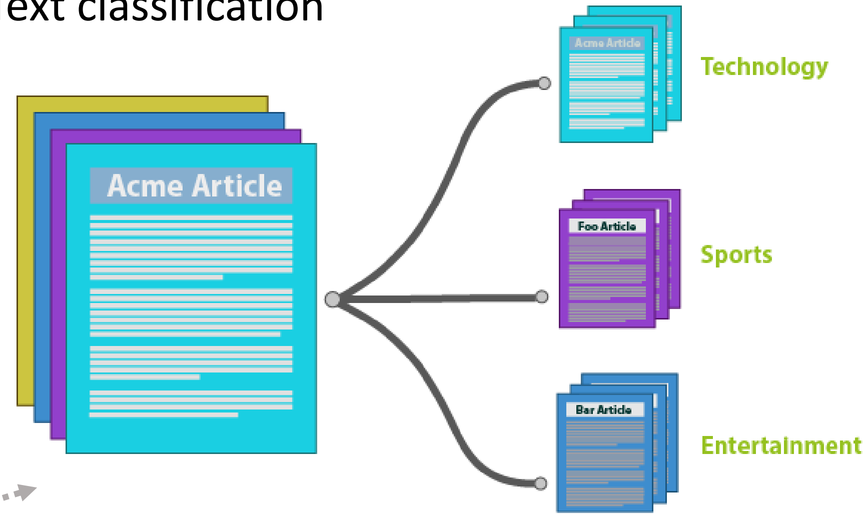
Named-entity recognition (NER)



Total nonfarm payroll employment rose by 266,000 CARDINAL in April DATE , and the unemployment rate was little changed at 6.1 percent PERCENT , the U.S. Bureau of Labor Statistics ORG reported today DATE . Notable job gains in leisure and hospitality, other services, and local government education were partially offset by employment declines in temporary help services and in couriers and messengers.



Text classification



(https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a)

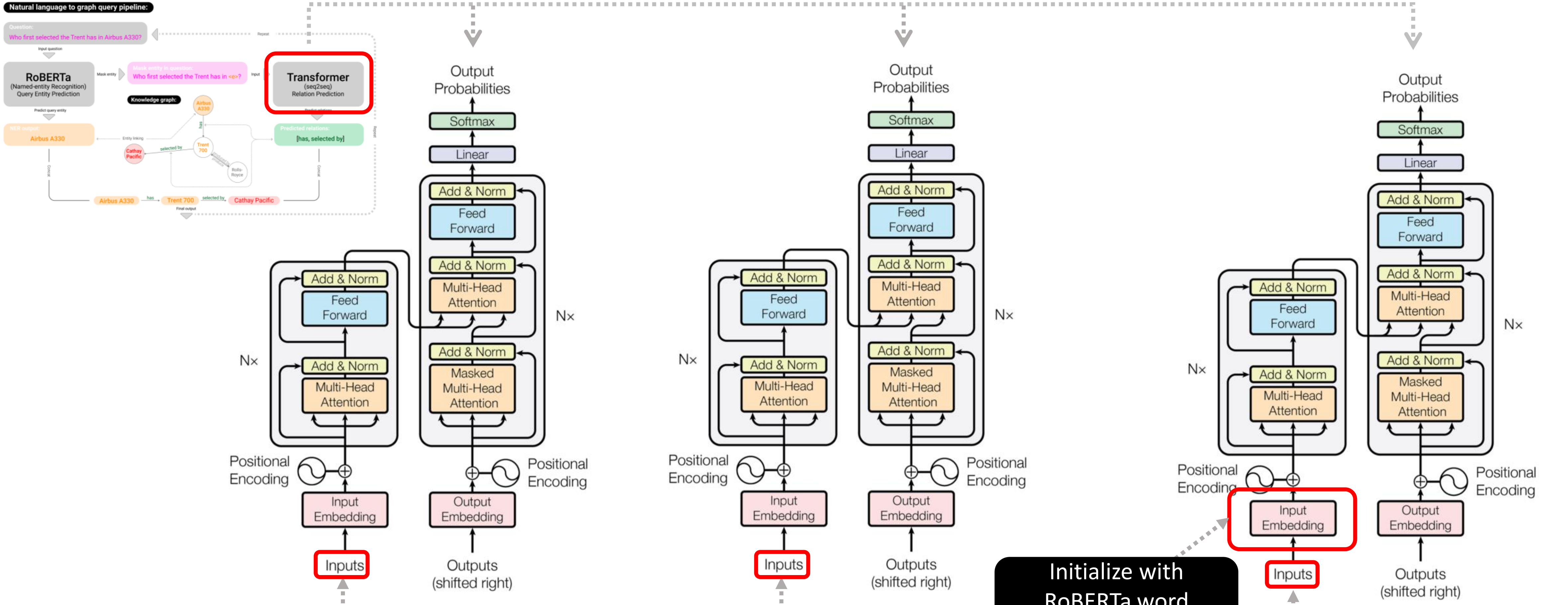
- **Fine tune RoBERTa** for Named-entity recognition (NER) to **predict query entity**
- **Fine tune RoBERTa** for text classification to **predict single relationship** in one-hop NL2GQ systems.

2.4 Multi-hop relation prediction (transformers) models:

Model 1: Seperator input

Model 2: Masked sentence input

Model 3: Masked sentence input and RoBERTa embedding



Initialize with RoBERTa word embedding

Who first selected the Trent has in A330? [sep] A330

Who first selected the Trent has in <e>?

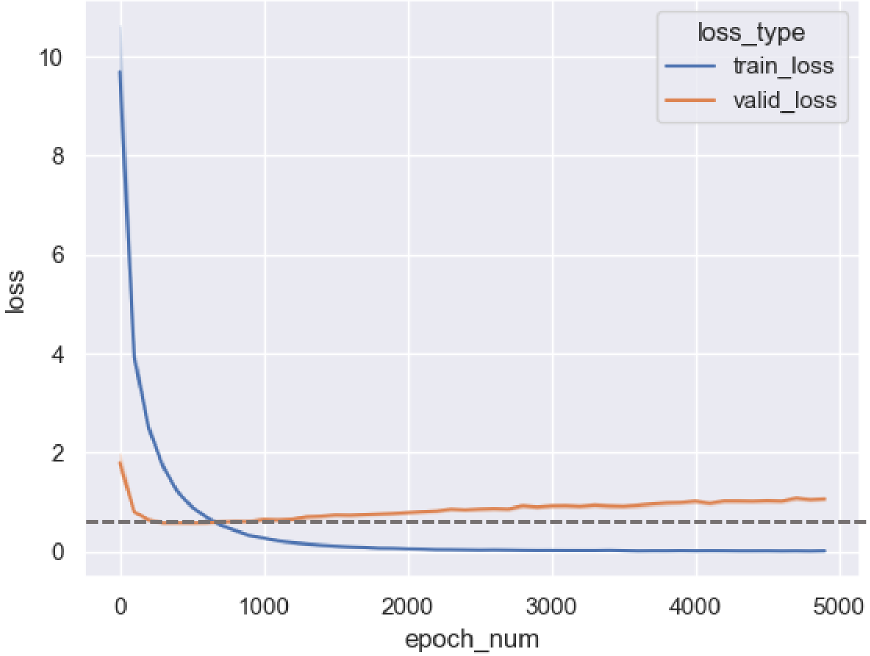
Who first selected the Trent has in <e>?

(Base images are from 'Attention is all you need' paper)

2.5 Multi-hop relation prediction models learning loss:

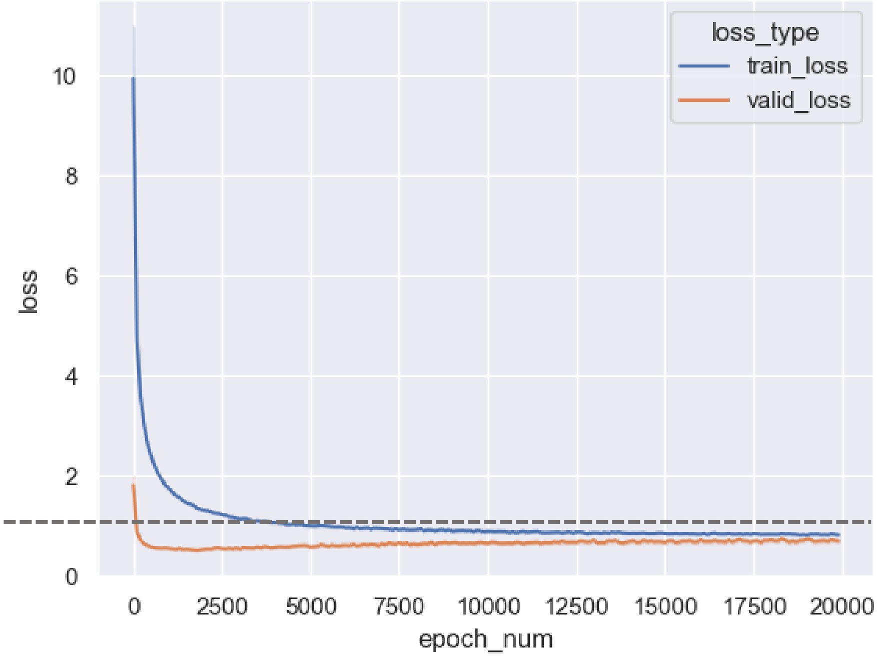
Model 1: : Seperator input

Input: Who first selected the Trent has in A330? [sep] A330



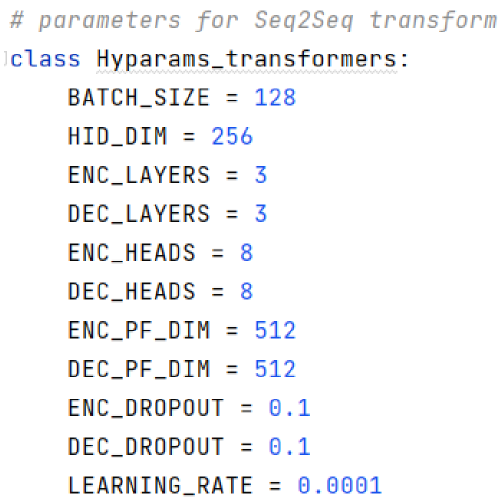
Model 2: Masked sentence input

Input: Who first selected the Trent has in <e>?



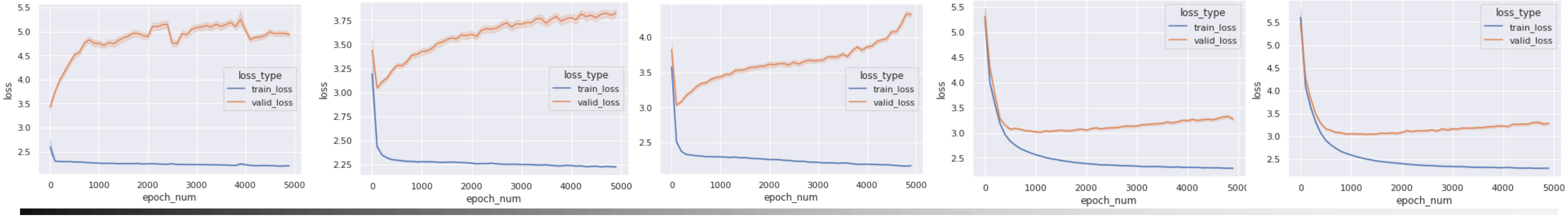
Best hyper parameters

Model 2: Masked sentence input



```
# parameters for Seq2Seq transform
class Hyparams_transformers:
    BATCH_SIZE = 128
    HID_DIM = 256
    ENC_LAYERS = 3
    DEC_LAYERS = 3
    ENC_HEADS = 8
    DEC_HEADS = 8
    ENC_PF_DIM = 512
    DEC_PF_DIM = 512
    ENC_DROPOUT = 0.1
    DEC_DROPOUT = 0.1
    LEARNING_RATE = 0.0001
```

Model 3: Masked sentence input and RoBERTa embedding with different hyper parameters



High Learning rate & model complexity

Low Learning rate & model complexity

2.6 Natural language question to graph query metrics on SimpleQuestion dataset:

Named Entity Recognition

	Precision	Recall	F1 score	test accuracy
RoBERTa (fine tuning)	0.96	0.96	0.95	0.96

Relation Prediction

	F1 macro	F1 micro	F1 weighted	test accuracy
One-hop: RoBERTa (fine tuning)	0.48	0.84	0.88	0.88
Multi-hop: transformers – Model 3 (Masked sentence input & RoBERTa embedding)	0.29	0.61	0.58	0.6
Multi-hop: transformers – Model 1 (Seperator sentence input)	0.31	0.69	0.67	0.69
Multi-hop: transformers – Model 2 (Masked sentence input)	0.42	0.76	0.75	0.76

Simple NL2GQ systems overall test accuracy

One-hop Natural language question to graph query	**0.8448**
Multi-hop Natural language question to graph query	**0.7296**

2.7 Natural language question to graph query summary:

Pros and cons

Pros

- Both one-hop and multi-hop NL2GQs have **good performance** on test dataset
- Can **deploy in any query language**, and **fast** query response time.

Cons

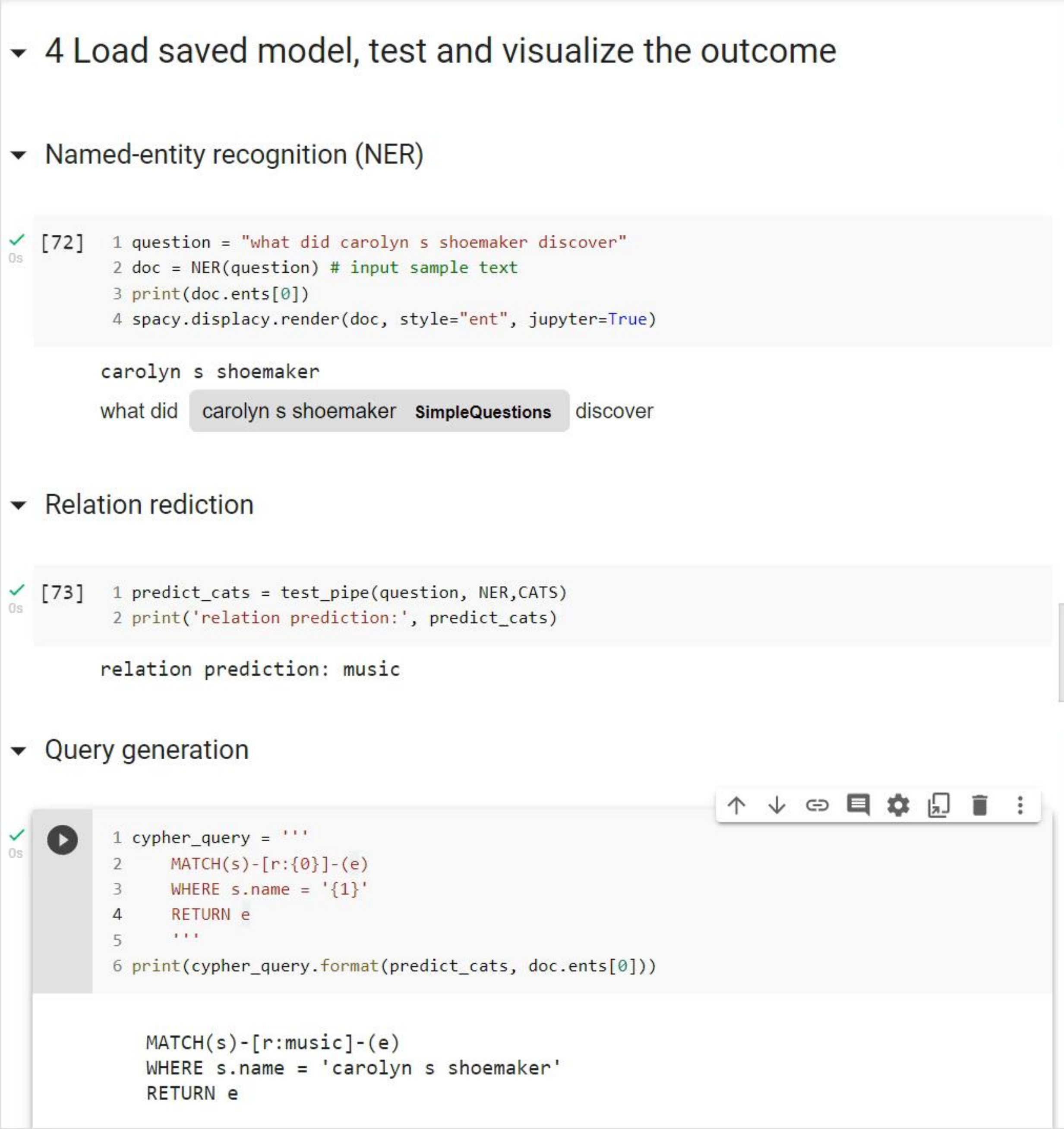
- It can only translate **simple multi-hop questions to queries**, can’t convert complex questions (E.g., Boolean, count)
- The best vanilla transformers’ vocabulary size is based on training data, which **may not query obscure words**.

Conclusion

- Our Simple one-hop & multi-hop NL2GQ experiments **achieved state-of-the-art** results in paperswithcode.com.
- There is still a **barrier** to translate **complex and strong logical** questions to graph query.
- Our results **only** represent the performance in **open-domain** NL2GQ systems, and it is **indeed further test** in **aerospace datasets**.
- In the future, we will try to use case-based NL2GQ t0 process complex query.

Simple NL2GQ systems overall test accuracy	
One-hop Natural language question to graph query	**0.8448**
Multi-hop Natural language question to graph query	**0.7296**

2.8 Natural language question to graph query showcase:



4 Load saved model, test and visualize the outcome

Named-entity recognition (NER)

[72]

```
question = "what did carolyn s shoemaker discover"
doc = NER(question) # input sample text
print(doc.ents[0])
spacy.displacy.render(doc, style="ent", jupyter=True)
```

```
carolyn s shoemaker
```

what did carolyn s shoemaker **SimpleQuestions** discover

Relation rediction

[73]

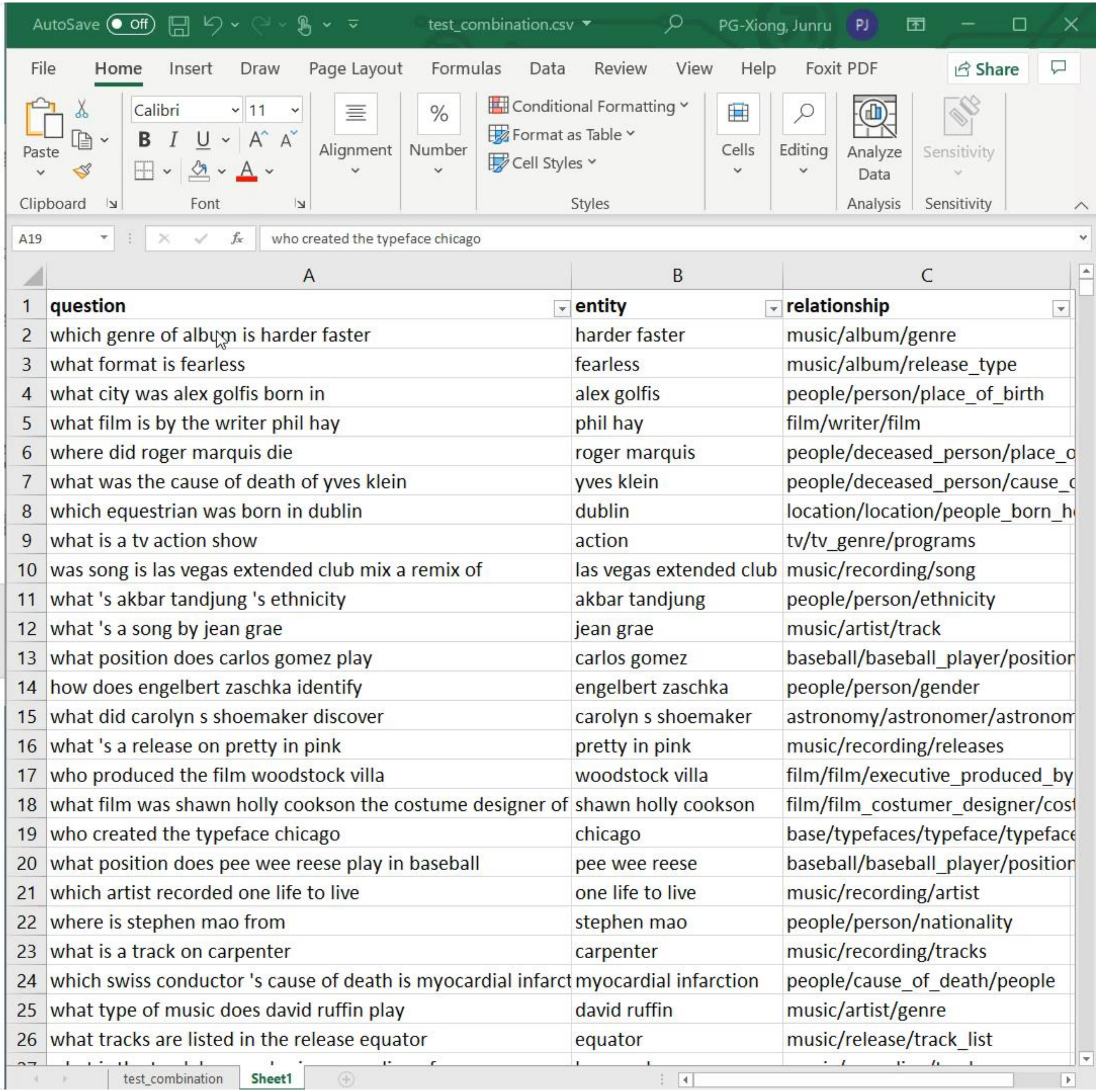
```
predict_cats = test_pipe(question, NER,CATS)
print('relation prediction:', predict_cats)
```

```
relation prediction: music
```

Query generation

```
cypher_query = '''
    MATCH(s)-[r:{0}]-(e)
    WHERE s.name = '{1}'
    RETURN e
    '''
print(cypher_query.format(predict_cats, doc.ents[0]))
```

```
MATCH(s)-[r:music]-(e)
WHERE s.name = 'carolyn s shoemaker'
RETURN e
```



A19: who created the typeface chicago

	A	B	C
1	question	entity	relationship
2	which genre of album is harder faster	harder faster	music/album/genre
3	what format is fearless	fearless	music/album/release_type
4	what city was alex golfis born in	alex golfis	people/person/place_of_birth
5	what film is by the writer phil hay	phil hay	film/writer/film
6	where did roger marquis die	roger marquis	people/deceased_person/place_o
7	what was the cause of death of yves klein	yves klein	people/deceased_person/cause_c
8	which equestrian was born in dublin	dublin	location/location/people_born_h
9	what is a tv action show	action	tv/tv_genre/programs
10	was song is las vegas extended club mix a remix of	las vegas extended club	music/recording/song
11	what 's akbar tandjung 's ethnicity	akbar tandjung	people/person/ethnicity
12	what 's a song by jean grae	jean grae	music/artist/track
13	what position does carlos gomez play	carlos gomez	baseball/baseball_player/position
14	how does engelbert zaschka identify	engelbert zaschka	people/person/gender
15	what did carolyn s shoemaker discover	carolyn s shoemaker	astronomy/astronomer/astronom
16	what 's a release on pretty in pink	pretty in pink	music/recording/releases
17	who produced the film woodstock villa	woodstock villa	film/film/executive_produced_by
18	what film was shawn holly cookson the costume designer of	shawn holly cookson	film/film_costumer_designer/cost
19	who created the typeface chicago	chicago	base/typefaces/typeface/typeface
20	what position does pee wee reese play in baseball	pee wee reese	baseball/baseball_player/position
21	which artist recorded one life to live	one life to live	music/recording/artist
22	where is stephen mao from	stephen mao	people/person/nationality
23	what is a track on carpenter	carpenter	music/recording/tracks
24	which swiss conductor 's cause of death is myocardial infarct	myocardial infarction	people/cause_of_death/people
25	what type of music does david ruffin play	david ruffin	music/artist/genre
26	what tracks are listed in the release equator	equator	music/release/track_list

Thanks