Chapter 6
Data to Text

Suppose you ask a question that can be answered using data in a database, and you use a text-to-data model to map the question to a formal query to the database. The query evaluated on a database returns the answer in a structured representation, such as a table in a relational database or a graph in a graph database. The full cycle of natural language interfaces to databases is not complete without mapping this structured data to natural language, which is the topic of this chapter.

The chapter discusses the process of converting structured data or formal representations to a naturally running text, which is conceptually the converse of the text-to-data material covered in Chapter 4. The chapter introduces both unconditional and conditional text generations, with a focus on conditional data-to-text generation, where the input is a structured representation (e.g., a table) and the output is plain text. Both domain specific data-to-text models and benchmarks (e.g., RDF to plain text, info-boxes to text, and summary of baseball score sheets) as well as domain-independent models and benchmarks (e.g., description of highlighted cells in a table, question answering and fact verification using tables, etc.) are reviewed and discussed.

6.1 Introduction

The term **Natural Language Generation** (NLG) is heavily used in the NLP literature but it means different things in different contexts. It can mean predicting (generating) the next word or sentence in a sequence using a Hidden Markov Model (HMM), a Recurrent Neural Network (RNN), or a pretrained language model such as CTRL [169], GPT [263], and GPT-3 [30]. In a more traditional sense, the term NLG describes the process of generating a natural running text from a *semantic representation*, which may be a tree or a database record, among other things.

Natural language generation can be done with or without a given context. When the output depends on the context, sometimes the term "controllable generation" is used. The context can be in the form of a user query, a prompt, a user profile, a document, etc. In this chapter we will talk about both of these types of generations, referred to as unconditional and conditional, with a focus on conditional data-to-text generation, where the input is a structured representation (e.g., a table) and the output is plain text.

6.1.1 Traditional generation

$$\mathcal{F}_4 \rightarrow \begin{bmatrix} cat: & s & \\ prot: & \boxed{1} & \begin{bmatrix} n: & "john" \end{bmatrix} \\ verb: & & \begin{bmatrix} v: & "like" \end{bmatrix} \\ goal: & \boxed{1} & \end{bmatrix}$$

Fig. 6.1: Sample FUF input for the sentence "John likes himself." (Image from [321]).

6.1.2 Data-to-Text Generation

FUF [101] is one of the early table-to-text systems, developed in late 1980s, that uses functional unification grammars to describe a set of key-value features. The features can be nested (e.g., the value of a key can be a feature structure). Also two keys (e.g., *prot* and *goal* in Figure 6.1) can have the same values. Another early example of data-to-text generation appears in [174]. The input, which comes from the ATIS airline reservation dataset [76], is represented in a structured data format, with nested attributes, e.g., the day of the departure is the 9th of August (Figure 6.2). One more example from the same time period is the work of Konstas and Lapata [175]. They use WeatherGov and WinHelp (Figure 6.3), two standard datasets, and generate multisentential text that describes the database records.

Database:

Flight		Day Number		Month		Condition			Search	
from	**to**	**number**	**dep/ar**	**month**	**dep/ar**	**arg1**	**arg2**	**type**	**type**	**what**
denver	boston	9	departure	august	departure	arrival_time	1600	<	query	flight

λ−expression: $\lambda x.flight(x) \wedge from(x, denver) \wedge to(x, boston) \wedge day\_number(x, 9) \wedge month(x, august) \wedge less\_than(arrival\_time(x), 1600)$

Text: Give me the flights leaving Denver August ninth coming back to Boston before 4pm.

Fig. 6.2: Structured data as input, along with a lambda expression representing its meaning, and the matching output text. (Image from [174]).

Other papers use as input attribute-value pairs extracted from Wikipedia biographical infoboxes, e.g., Figure 6.4. In this scenario, the generated output has to describe the entity in question using grammatical sentences, covering the most important facts in the input.

The Neural Wikipedian paper [322] and its follow up paper [323] introduce another interesting dataset that aligns sets of RDF triples to verbal summaries. An example data point from this dataset is shown in Figure 6.5. The text summary is a sentence that succinctly covers multiple facts presented in the input triples.

Database Records
temp(time:6-21, min:**9**, mean:**15**, max:**21**) wind-spd(time:6-21, min:**15**, mean:**20**, max:**30**) sky-cover(time:6-9, percent:*25-50*) sky-cover(time:9-12, percent:*50-75*) wind-dir(time:6-21, mode:SSE) gust(time:6-21, min:**20**, mean:**30**, max:**40**)
Output Text
Cloudy, with a high around 20. South southeast wind between 15 and 30 mph. Gusts as high as 40 mph.

Database Records
desktop(cmd:lclick, name:*start*, type:button) start(cmd:lclick, name:*settings*, type:button) start-target(cmd:lclick, name:*control panel*, type:button) win-target(cmd:dblclick, name:*users and passwords*, type:item) contMenu(cmd:lclick, name:*advanced*, type:tab) action-contMenu(cmd:lclick, name:*advanced*, type:button)
Output Text
Click start, point to settings, and then click control panel. Double-click users and passwords. On the advanced tab, click advanced.

Fig. 6.3: WeatherGov and WinHelp examples. (Image from [175]).

(a)

Born	Robert Joseph Flaherty February 16, 1884 Iron Mountain, Michigan, U.S.
Died	July 23, 1951 (aged 67) Dummerston, Vermont, U.S.
Cause of death	Cerebral thrombosis
Occupation	Filmmaker
Spouse(s)	Frances Johnson Hubbard

(b) Robert Joseph Flaherty, (February 16, 1884 July 23, 1951) was an American film-maker who directed and produced the first commercially successful feature-length documentary film, Nanook of the North (1922). The film made his reputation and nothing in his later life fully equalled its success, although he continued the development of this new genre of narrative documentary, e.g., with Moana (1926), set in the South Seas, and Man of Aran (1934), filmed in Ireland's Aran Islands. He is considered the "father" of both the documentary and the ethnographic film. Flaherty was married to writer Frances H. Flaherty from 1914 until his death in 1951. Frances worked on several of her husband's films, and received an Academy Award nomination for Best Original Story for Louisiana Story (1948).

(c) *Robert Joseph Flaherty, (February 16, 1884 July 23, 1951) was an American film-maker. Flaherty was married to Frances H. Flaherty until his death in 1951.*

Fig. 6.4: Sample Wikipedia biographical box, along with the matching text and verbalization (Image from [249]).

Triples	Atlas_Shrugged **literaryGenre** Science_fiction Atlas_Shrugged **country** United_States John_Galt **series** Atlas_Shrugged Atlas_Shrugged **publicationYear** "1957" Atlas_Shrugged **author** Ayn_Rand
Text Summary	Atlas Shrugged is a science fiction novel by Ayn Rand.

Fig. 6.5: Text summary aligned to a set of five RDF triples. Each of the triples includes a predicate (verb) in bold face and two arguments connected by that predicate. (Example from [323]).

6.1.3 Abstract Meaning Representation (AMR) for text generation

One representation suitable for meaning to text generation is AMR (Abstract Meaning Representation), previously discussed in the text-to-data chapter.

In natural language understanding, there is not much leeway in representing the content of the input sentence using a meaning representation. In contrast, in natural language generation, the expected output sentence for a given input is underspecified. For example, the AMR representation in Figure 6.6 indicates that a "break" event has taken or is taking place, and that its main arguments are "dog" and "window". This input can be translated to natural language sentences in multiple ways, e.g., "the dog broke the window", or "the dog will break the window", or "the window was broken by the dog", or even "the dog broke it."

```
(b / break-01
  :ARG0 (d / dog)
  :ARG1 (w / window))
```

Fig. 6.6: An example of meaning input underspecification. The input covers only some of the information needed to generate a full sentence. The rest of the features required to generate such a sentence have to come from other sources, e.g., discourse constraints, user preferences, etc.

In other words, natural language generation depends significantly on a series of choices made during the different stages mentioned earlier, e.g., content selection, planning, lexicalization, and surface realization. Such choices may be made in an arbitrary way, or they may depend on additional constraints. Such constraints can be based on knowledge about the time of the events (to determine the grammatical tense of the sentences) or the discourse structure (e.g., whether the dog or the window is the theme of the passage, or whether to replace one or more of the entities that have already been mentioned with a pronoun). Gatt and Kramer have a good survey that covers such topics [114].

Using AMR as a representation for NLG has been very popular over the recent years. The SemEval 2017 Task 9 [1] was on generating text from Abstract Meaning Representations. It includes two tasks, one of which is on AMR to English generation using (domain-independent) news and discussion text.

```
s / say-01
      :ARG0 (s2 / service
            :mod (e / emergency)
            :location (c / city :wiki ‘‘London’’
                  :name (n / name :op1 ‘‘London’’)))
      :ARG1 (s3 / send-01
            :ARG1 (p / person :quant 11)
            :ARG2 (h / hospital)
            :mod (a / altogether)
            :purpose (t / treat-03
                  :ARG1 p
                  :ARG2 (w / wound-01
                        :ARG1 p
                        :mod (m / minor)))))
```

Fig. 6.7: AMR input representation. This structure can be generated as “The London emergency services said that altogether 11 people had been sent to hospital for treatment due to minor wounds.”

6.1.4 Neural Generation

Many neural methods for natural language generation are based on an encoder-decoder model. Encoding was traditionally done using a recurrent neural network (RNN) that converts the input text into a formal meaning representation, which is then decoded. This is essentially the same machinery that we described in Chapter 4, with the output being plain text this time.

[1] https://alt.qcri.org/semeval2017/task9/

At each time step t, the model computes a vector of scores for each item in the output vocabulary and converts them to probabilities using the softmax function. During decoding, the algorithm decides what token to choose as the next item in the output. A simple way to perform decoding is to use *teacher forcing* (or maximum likelihood training) to produce, at each step, the most likely token, given the tokens generated so far. Beam decoding is a greedy algorithm that considers at each time step the k most likely candidates. Argmax decoding is an extreme case of a greedy algorithm with a beam size of $k = 1$.

Greedy methods tend to become repetitive and are usually avoided in favor of sampling methods such as top-k sampling or nucleus sampling. The amount of randomness can be regulated by adjusting the temperature parameter of the softmax. A survey of decoding methods in NLG is done by Zarriess et al. [377].

6.2 Domain Specific Table-to-Text

Research contributions in table-to-text generation have been split into two thrusts - domain-independent and domain-specific. A number of tasks have been created over the years in domain-specific table-to-text generation. We will now describe some of them.

6.2.1 SRST

Surface realization in NLP involves mapping structured data such as treebank annotations and meaning representaions to sentences. The First Multilingual Surface Realisation Shared Task (SR'18) [221, 222], presented at an ACL workshop as a follow up to a Surface Realisation Shared Task in 2011 [18], includes two tracks: (1) a shallow track (in ten languages) with universal dependency structures from CONLL where the lemmatized words appear in a shuffled order, and (2) a deep track (in three languages) where all function words, as well as most of the morphological information, are additionally removed. Figure 6.8 shows a sample input representation. A shallow input representation (from Track 1 of SR'18) is shown in Figure 6.9 and the corresponding deep input representation (from Track 2 of SR'18) is shown in Figure 6.10. The Third Multilingual Surface Realisation Shared Task (SR'20) [220] introduces two variants of each of the two tracks (shallow and deep). In one variant, only the data given as input for the track could be used for training whereas the second variant allowed the use of any source of data for training.

1	The	the	DET	DT	Definite=Def|PronType=Art	2	det
2	third	third	ADJ	JJ	Degree=Pos|NumType=Ord	5	nsubj_pass
3	was	be	AUX	VBD	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	5	aux
4	being	be	AUX	VBG	VerbForm=Ger	5	aux_pass
5	run	run	VERB	VBN	Tense=Past|VerbForm=Part|Voice=Pass	0	root
6	by	by	ADP	IN	_	8	case
7	the	the	DET	DT	Definite=Def|PronType=Art	8	det
8	head	head	NOUN	NN	Number=Sing	5	obl
9	of	of	ADP	IN	_	12	case
10	an	a	DET	DT	Definite=Ind|PronType=Art	12	det
11	investment	investment	NOUN	NN	Number=Sing	12	compound
12	firm	firm	NOUN	NN	Number=Sing	8	nmod
13	.	.	PUNCT	.	_	5	punct

Fig. 6.8: A universal dependency structure for English from SRST'18. (Image from [222]).

1	the	_	DET	DT	Definite=Def|PronType=Art	2	det
2	third	_	ADJ	JJ	Degree=Pos|NumType=Ord	3	nsubj_pass
3	run	_	VERB	VBN	Tense=Past|VerbForm=Part|Voice=Pass	0	root
4	be	_	AUX	VBD	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	3	aux
5	be	_	AUX	VBG	VerbForm=Ger	3	aux_pass
6	head	_	NOUN	NN	Number=Sing	3	obl
7	.	_	PUNCT	.	_	3	punct
8	by	_	ADP	IN	_	6	case
9	the	_	DET	DT	Definite=Def|PronType=Art	6	det
10	firm	_	NOUN	NN	Number=Sing	6	nmod
11	an	_	DET	DT	Definite=Ind|PronType=Art	10	det
12	investment	_	NOUN	NN	Number=Sing	10	compound
13	of	_	ADP	IN	_	10	case

Fig. 6.9: Shallow input version of the previous example (Track 1). (Image from [222]).

1	third	_	ADJ	_	Degree=Pos	2	A2
2	run	_	VERB	_	Tense=Past|Aspect=Progr	0	ROOT
3	head	_	NOUN	_	Number=Sing|Definiteness=Def	2	A1
4	firm	_	NOUN	_	Number=Sing|Definiteness=Indef	3	A2
5	investment	_	NOUN	_	Number=Sing	4	AM

Fig. 6.10: Deep input derived from the previous example (Track 2). (Image from [222]).

6.2.2 E2E

The **e2e** text generation dataset[2], described in [97, 98, 236], is limited to the restaurant domain and is intended to exhibit high syntactic complexity and different discourse phenomena. Figure 6.11 shows a sample data point, including a meaning representation for the restaurant domain and a single corresponding reference sentence. The domain ontology is shown in Figure 6.12. An example with multiple reference sentences is included in Figure 6.13. The e2e dataset includes 6,039 meaning representations and 51,426 reference sentences.

MR	name[The Wrestlers], priceRange[cheap], customerRating[low]
Reference	The wrestlers offers competitive prices, but isn't highly rated by customers.

Fig. 6.11 A meaning representation and the corresponding reference sentence. (Image from [97]).

One of the early approaches to the e2e task is the neural template induction approach of [343]. An example is shown in Figure 6.14. The templates are learned automatically and include alternative phrasings,

[2] http://www.macs.hw.ac.uk/InteractionLab/E2E/

Attribute	Data Type	Example value
name	verbatim string	*The Eagle, ...*
eatType	dictionary	*restaurant, pub, ...*
familyFriendly	boolean	*Yes / No*
priceRange	dictionary	*cheap, expensive, ...*
food	dictionary	*French, Italian, ...*
near	verbatim string	*Zizzi, Cafe Adriatic, ...*
area	dictionary	*riverside, city center, ...*
customerRating	dictionary	*1 of 5 (low), 4 of 5 (high), ...*

Fig. 6.12 Domain ontology for the e2e task. (Image from [97]).

```
name[The Eagle],
eatType[coffee shop],
food[French],
priceRange[moderate],
customerRating[3/5],
area[riverside],
kidsFriendly[yes],
near[Burger King]
```

Fig. 6.13: Example from the e2e dataset: "The three star coffee shop, The Eagle, gives families a mid-priced dining experience featuring a variety of wines and cheeses. Find The Eagle near Burger King."

e.g., "providing/serving/offering". Once generated, these templates are then filled in with the missing words for the specific example, e.g., as in Figure 6.15.

Source Entity: Cotto

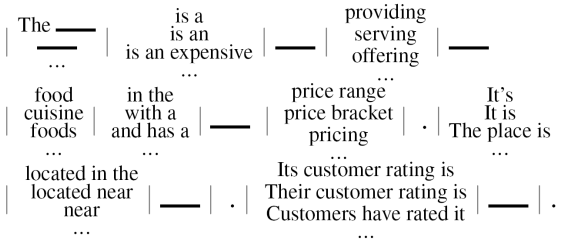
type[coffee shop], rating[3 out of 5], food[English], area[city centre], price[moderate], near[The Portland Arms]

System Generation:

Cotto is a coffee shop serving English food in the moderate price range. It is located near The Portland Arms. Its customer rating is 3 out of 5.

(a) An example from the e2e dataset.

Neural Template:



(b) An induced neural template.

Fig. 6.14: E2e examples. (Images from [343]).

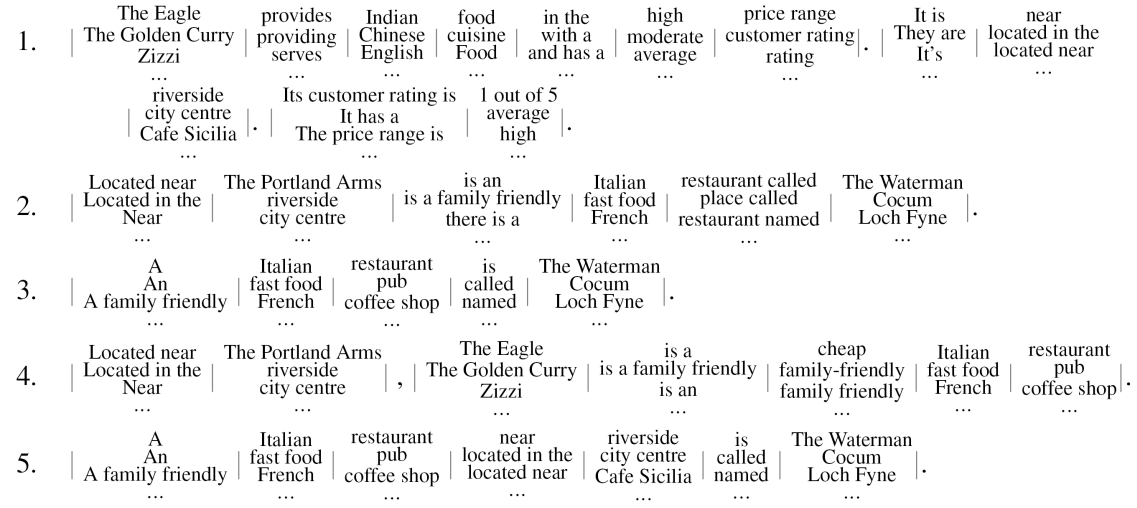


Fig. 6.15: Filled out templates. (Image from [343]).

6.2.3 WebNLG

The **WebNLG** challenge [113], introduced in 2017, is based on mapping RDF triples to plain text. The WebNLG dataset includes 25,298 (data,text) pairs and 9,674 sets of up to seven RDF triples extracted from DBpedia. The texts verbalize these sets of RDF triples. The examples are all domain-specific in 15 domains such as University, Building, Airport, City, Athlete, Politician, Astronaut, Artist, etc. In the example shown in Figure 6.16, three RDF triples about the same person are combined into a single sentence.

```
(JOHN E BLAHA BIRTHDATE 1942 08 26)
(JOHN E BLAHA BIRTHPLACE SAN ANTONIO)
(JOHN E BLAHA OCCUPATION FIGHTER PILOT)
```

Fig. 6.16: WebNLG example from [113] corresponding to the sentence "John E Blaha, born in San Antonio on 1942-08-26, worked as a fighter pilot.".

A follow up paper [40] expands[3] the corpus to German and also introduces some new tasks such as Discourse Ordering, Lexicalization (word choice), and Referring Expression (e.g., pronoun) Generation. An example of an input set with five RDF triples is shown in Figure 6.17. Figure 6.18 shows the mapping between the tags and entities for the template "AGENT-1 is located in BRIDGE-1 , PATIENT-1 and serves the city of PATIENT-2 . BRIDGE-1 is part of PATIENT-3 and PATIENT-4."[40].

A more recent[4] version of WebNLG expands the generation task to other languages such as Russian. An example with six triples is shown in Figure 6.19. This challenge includes three tasks: RDF-to-Text (Generation), Text-to-RDF, and Automatic Evaluation for WebNLG.

[3] https://github.com/ThiagoCF05/webnlg

[4] https://webnlg-challenge.loria.fr/challenge_2020

Subject	Predicate	Object
Appleton International Airport	location	Greenville, Wisconsin
Greenville, Wisconsin	isPartOf	Ellington, Wisconsin
Greenville, Wisconsin	isPartOf	Menasha (town), Wisconsin
Greenville, Wisconsin	country	United States
Appleton International Airport	cityServed	Appleton, Wisconsin

Fig. 6.17: Input set of five RDF triples that can be verbalized as "The Appleton International Airport is located in Greenville, Wisconsin, United States and serves the city of Appleton, Wisconsin. Greenville is part of the town of Menasha and Ellington, Wisconsin.". (Example from [40]).

Tag	Entity
AGENT-1	Appleton International Airport
BRIDGE-1	Greenville, Wisconsin
PATIENT-1	United States
PATIENT-2	Appleton, Wisconsin
PATIENT-3	Menasha (town), Wisconsin
PATIENT-4	Ellington, Wisconsin

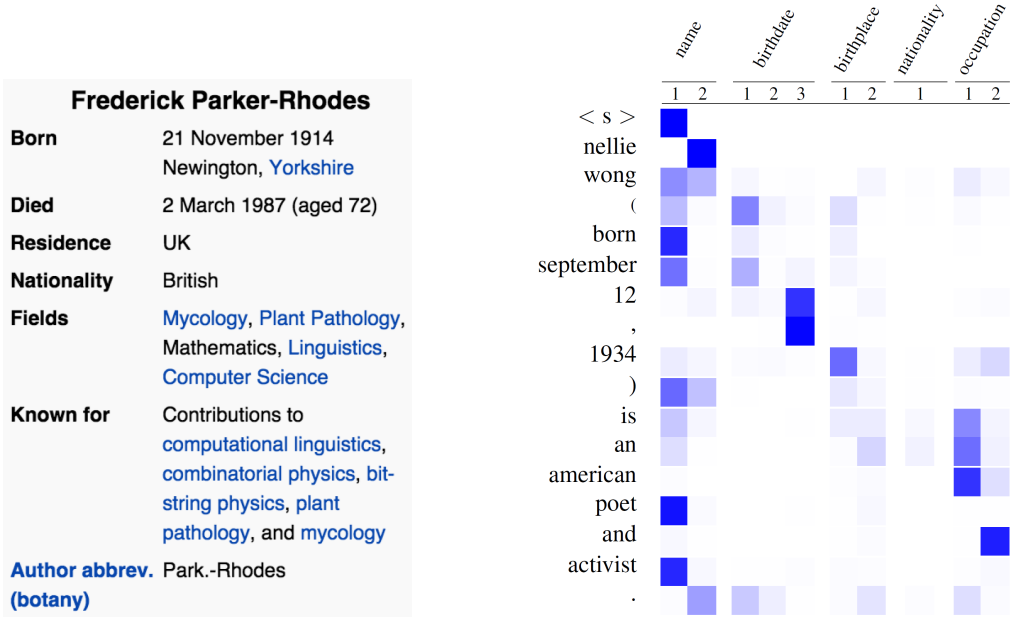
Fig. 6.18: Tags and entities for the delexicalized/wikified templates. (Example from [40]).

```
<entry category="Company" eid="Id3" shape="(X (X) (X) (X) (X) (X) (X))" shape_type="sibling" size="6">
      <modifiedtripleset>
        <mtriple>Chinabank | foundingDate | 1920-08-16</mtriple>
        <mtriple>Chinabank | numberOfLocations | 295</mtriple>
        <mtriple>Chinabank | foundationPlace | Manila</mtriple>
        <mtriple>Chinabank | type | Public_company</mtriple>
        <mtriple>Chinabank | foundationPlace | Insular_Government_of_the_Philippine_Islands</mtriple>
        <mtriple>Chinabank | location | Philippines</mtriple>
      </modifiedtripleset>
</entry>
```

Fig. 6.19: WebNLG example with six input triples. The output sentences are (1) "Chinabank was founded on August 16, 1920 in the Insular Government of the Philippine Islands In Manila. Located in the Philippines, it is a publicly traded company with 295 branches.", (2) "Chinabank was founded on August 16, 1920 in Manila, in the Insular Government of the Philippine Islands. It is a publicly traded company, located in the Philippines with 295 branches.", and (3) "Publicly traded Chinabank, founded in Manila, Philippines, at the time of the Insular Government of the Philippine Islands, on August 16, 1920, operates 295 banking centers.".

6.2.4 WikiBio

WikiBio was introduced by Lebret et al. [183] for the biography domain. It includes more than 700K sentences from Wikipedia biographies. Figure 6.20a shows an example from the WikiBio dataset. Figure 6.20b shows the attention scores from slot to word. Finally, Figure 6.21 displays three examples of the outputs of their table-conditioned neural language model (TableNLM).



(a) WikiBio example.

(b) Attention scores. =.

Fig. 6.20: Wikibio examples. (Images from [183]).

Some of the early work on WikiBio[5] is Liu et al. [207]. In this paper, the authors propose a structure-aware seq2seq architecture that uses a field-gating encoder which associates an LSTM unit with the corresponding field value. During decoding, dual attention to both the words and fields connects the generated description and the table. Figure 6.22 shows a sample Wikipedia infobox and the corresponding field representation. Figure 6.23 shows the structure-aware seq2seq architecture.

The follow up paper by Sha et al. [283] introduces an order-planning generation model for Wikipedia data (Figure 6.25). They use a self-adaptive gate (Figure 6.26) that takes into account both content-based and link-based attention (Figure 6.28). A sample output is shown in Figure 6.27.

[5] https://github.com/tyliupku/wiki2bio

Model	Generated Sentence
Reference	frederick parker-rhodes (21 march 1914 – 21 november 1987) was an english linguist, plant pathologist, computer scientist, mathematician, mystic, and mycologist.
Baseline (Template KN)	frederick parker-rhodes (born november 21 , 1914 – march 2 , 1987) was an english cricketer .
Table NLM +Local (field, start)	frederick parker-rhodes (21 november 1914 – 2 march 1987) was an australian rules footballer who played with carlton in the victorian football league (vfl) during the XXXXs and XXXXs .
+ Global (field)	frederick parker-rhodes (21 november 1914 – 2 march 1987) was an english mycology and plant pathology , mathematics at the university of uk .
+ Global (field, word)	frederick parker-rhodes (21 november 1914 – 2 march 1987) was a british computer scientist , best known for his contributions to computational linguistics .

Fig. 6.21: Reference and sample system outputs for the WikiBio input shown in Figure 6.20a (Image from [183]).

name	George Mikell
birthname	Jurgis Mikelaitis
birthdate	4 April 1929 (age 88)
birthplace	Bildeniai, Lithuania
nationality	Lithuanian, Australian
occupation	Actor, writer
years active	1957–present
known for	The Guns of Navarone The Great Escape

word	*Field embedding*
George	(name, 1, 2)
Mikell	(name, 2, 1)
Jurgis	(birthname, 1, 2)
Mikelaitis	(birthname, 2, 1)
4	(birthdate, 1, 7)
April	(birthdate, 2, 6)
...	...
The	(known for, 5, 3)
Great	(known for, 6, 2)
Escape	(known for, 7, 1)

Fig. 6.22 Wikipedia infobox and its field representation for George Mikell. (Image from [207]).

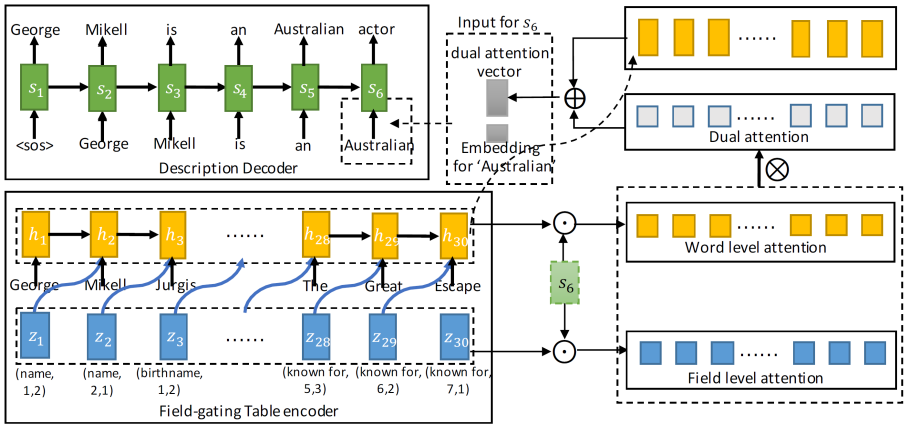


Fig. 6.23: Structure-aware seq2seq architecture for generating the description for George Mikell in Figure 6.22 (Image from [207]).

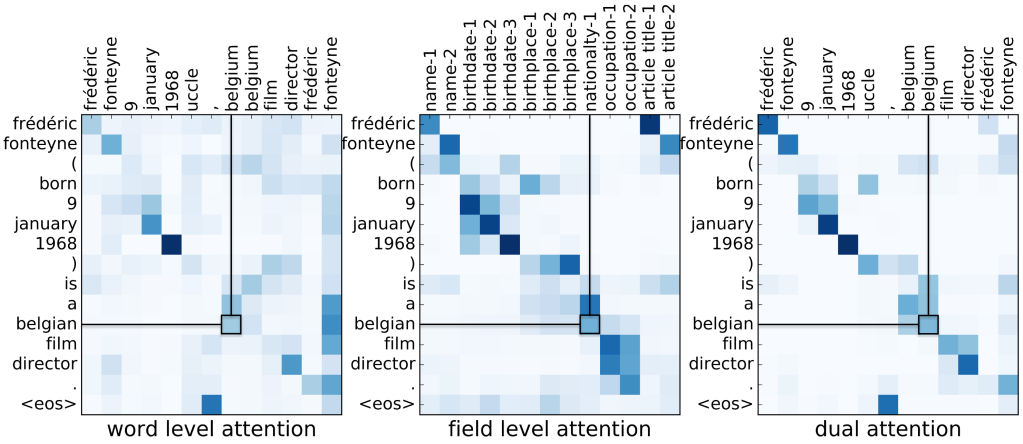


Fig. 6.24: Dual attention. (Image from [207]).

Table:

ID	Field	Content
1	Name	*Arthur Ignatius Conan Doyle*
2	Born	*22 May 1859 Edinburgh, Scotland*
3	Died	*7 July 1930 (aged 71) Crowborough, England*
4	Occupation	*Author, writer, physician*
5	Nationality	*British*
6	Alma mater	*University of Edinburgh Medical School*
7	Genre	*Detective fiction fantasy*
8	Notable work	*Stories of Sherlock Homes*

Text: Sir Arthur Ignatius Conan Doyle (22 May 1859 – 7 July 1930) was a British writer best known for his detective fiction featuring the character Sherlock Holmes.

Fig. 6.25 A matching pair of a Wikipedia infobox and its description. (Image from [283]).

6.2.5 RotoWire

The **RotoWire** dataset[6] introduced in [342], includes the summaries of basketball games aligned to the matching box-and-line scores. A sample data point is shown in Figure 6.29. An output of a system[7] for input records describing a basketball game is shown in Figure 6.30.

A related paper is Puduppully et al. [256]. Figure 6.31 shows the generation model with explicit content selection and planning. The basic idea is to first generate a content plan highlighting *what to say*, i.e., which information from the table should be mentioned, and *in which order* the content should be generated; then the text generation step generates a document by taking the content plan into account. Figure 6.32 shows a sample output based on template-based system and Figure 6.33 shows a sample output based on the proposed model. This work was followed by [257] and [258] to take into account of important content such as entities, events and their interactions as well as their high-level organization.

[6] https://github.com/harvardnlp/boxscore-data

[7] https://github.com/harvardnlp/data2text

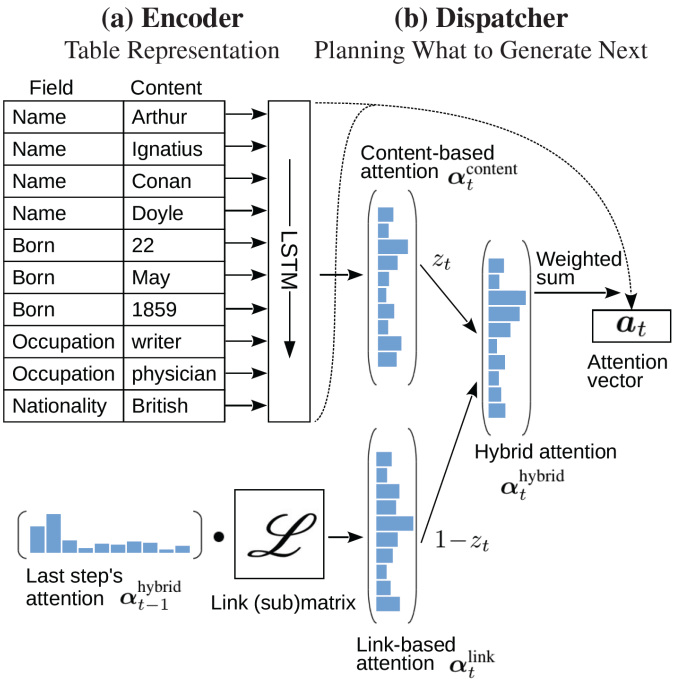


Fig. 6.26 The encoder and the dispatcher of the system. (Image from [283]).

Name	Emmett John Rice
Birth date	December 21, 1919
Birth place	Florence, South Carolina, United States
Death date	March 10, 2011 (aged 91)
Death place	Camas, Washington, United States
Nationality	American
Occupation	Governor of the Federal Reserve System, Economics Professor
Known for	Expert in the Monetary System of Developing Countries, Father to Susan E. Rice

Reference	emmett john rice (december 21 , 1919 – march 10 , 2011) was a former governor of the federal reserve system , a Cornell university economics professor , expert in the monetary systems of developing countries and the father of the current national security advisor to president barack obama , susan e . rice .
Content-based attention	emmett john rice (december 21 , 1919 – march 10 , 2011) was an economist , author , public official and the former american governor of the federal reserve system , the first african american UNK .
Hybrid attention	emmett john rice (december 21 , 1919 – march 10 , 2011) was an american economist , author , public official and the former governor of the federal reserve system , expert in the monetary systems of developing countries .

Fig. 6.27: Sample outputs for the given infobox. (Image from [283]).

6.2.6 WikiTableT

WikiTableT[8] was introduced by Chen et al. [51]. This large dataset with millions of instances is based on Wikipedia sections, the corresponding tables and matching metadata. It covers a diverse range of topics and generation tasks. An example is shown in Figure 6.34.

[8] https://github.com/mingdachen/WikiTableT

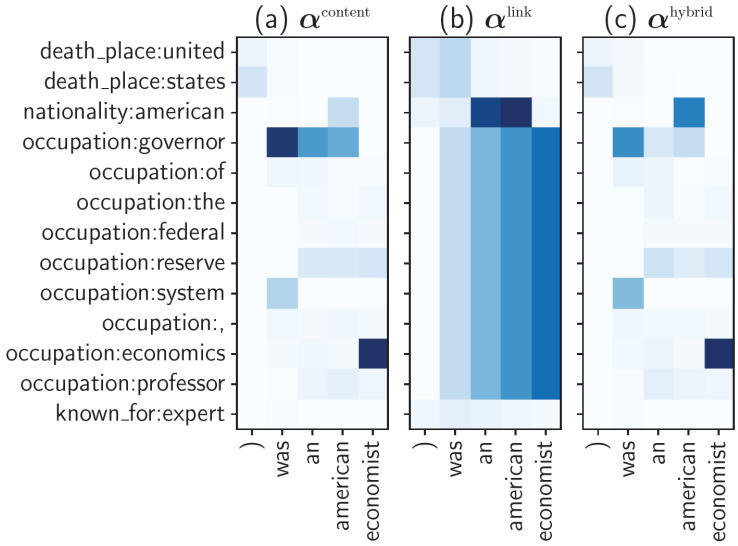


Fig. 6.28 Attention map between words and fields. (a) Content-based attention. (b) Link-based attention. (c) Hybrid attention (Image from [283]).

TEAM	WIN	LOSS	PTS	FG_PCT	RB	AS ...
Heat	11	12	103	49	47	27
Hawks	7	15	95	43	33	20

PLAYER	AS	RB	PT	FG	FGA	CITY ...
Tyler Johnson	5	2	27	8	16	Miami
Dwight Howard	4	17	23	9	11	Atlanta
Paul Millsap	2	9	21	8	12	Atlanta
Goran Dragic	4	2	21	8	17	Miami
Wayne Ellington	2	3	19	7	15	Miami
Dennis Schroder	7	4	17	8	15	Atlanta
Rodney McGruder	5	5	11	3	8	Miami
Thabo Sefolosha	5	5	10	5	11	Atlanta
Kyle Korver	5	3	9	3	9	Atlanta
...						

(a) Rotowire Table.

The Atlanta Hawks defeated the Miami Heat , 103 - 95 , at Philips Arena on Wednesday . Atlanta was in desperate need of a win and they were able to take care of a shorthanded Miami team here . Defense was key for the Hawks , as they held the Heat to 42 percent shooting and forced them to commit 16 turnovers . Atlanta also dominated in the paint , winning the rebounding battle , 47 - 34 , and outscoring them in the paint 58 - 26.The Hawks shot 49 percent from the field and assisted on 27 of their 43 made baskets . This was a near wire - to - wire win for the Hawks , as Miami held just one lead in the first five minutes . Miami (7 - 15) are as beat - up as anyone right now and it 's taking a toll on the heavily used starters . Hassan Whiteside really struggled in this game , as he amassed eight points , 12 rebounds and one blocks on 4 - of - 12 shooting ...

(b) The corresponding summary.

Fig. 6.29: RotoWire examples. (Images from [342]).

The Utah Jazz (38 - 26) defeated the Houston Rockets (38 - 26) 117 - 91 on Wednesday at Energy Solutions Arena in Salt Lake City . The Jazz got out to a quick start in this one , out - scoring the Rockets 31 - 15 in the first quarter alone . Along with the quick start , the Rockets were the superior shooters in this game , going 54 percent from the field and 43 percent from the three - point line , while the Jazz went 38 percent from the floor and a meager 19 percent from deep . The Rockets were able to out - rebound the Rockets 49 - 49 , giving them just enough of an advantage to secure the victory in front of their home crowd . The Jazz were led by the duo of Derrick Favors and James Harden . Favors went 2 - for - 6 from the field and 0 - for - 1 from the three - point line to score a game - high of 15 points , while also adding four rebounds and four assists ....

Fig. 6.30 A generated document on RotoWire data (Correctly mapped data are shown in blue and erroneous mappings are shown in red (Image from [342]).

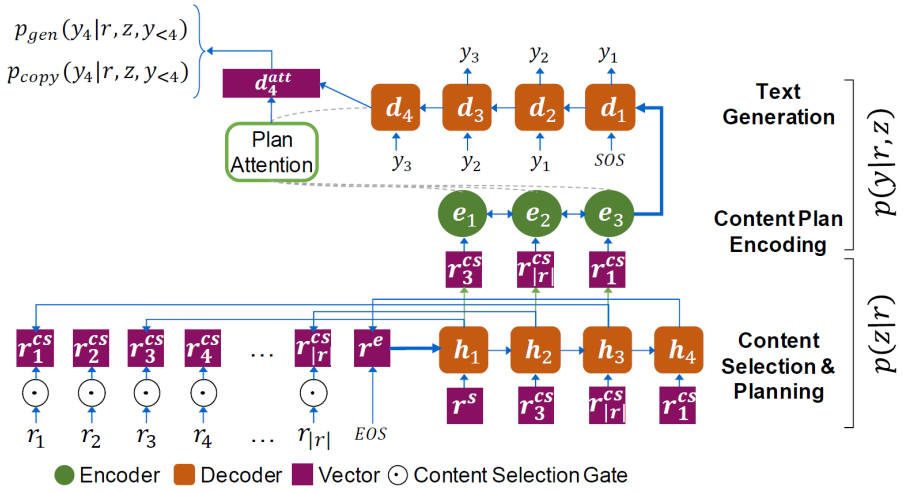


Fig. 6.31: Generation model. (Image from [256]).

The Washington Wizards (8-13) defeated the Denver Nuggets (8-15) 92-85. Bradley Beal scored 26 points (8-15 FG, 4-7 3PT, 6-6 FT) to go with 3 rebounds. Nikola Jokic scored 17 points (6-10 FG, 0-0 3PT, 5-7 FT) to go with 11 rebounds. Markieff Morris scored 15 points (5-12 FG, 0-0 3PT, 5-5 FT) to go with 3 rebounds. John Wall scored 15 points (5-14 FG, 0-4 3PT, 5-6 FT) to go with 7 rebounds. Danilo Gallinari scored 14 points (3-11 FG, 1-8 3PT, 7-9 FT) to go with 4 rebounds. Jusuf Nurkic scored 13 points (6-6 FG, 0-0 3PT, 1-2 FT) to go with 7 rebounds. The Washington Wizards' next game will be at home against the Dallas Mavericks, while the Denver Nuggets will travel to play the Bulls.

Fig. 6.32: Sample output based on template. (Image from [256]).

The Washington Wizards defeated the visiting Denver Nuggets 92-85 at Verizon Center on Monday. The Wizards (8-13) came into this game winners of five of their last eight games, but the Wizards (8-15) jumped out to a 10-point lead at the end of the first quarter. Bradley Beal led the way for the Wizards with a game-high 26 points on 8-of-15 shooting from the field. John Wall shot 5-of-14 from the field on his way to 15 points, to go along with seven rebounds, five assists, three steals and one block. Jusuf Nurkic chipped in 13 points, seven rebounds and one assist. Jameer Nelson filled out the stat sheet with 10 points, eight assists, four rebounds and two steals in 36 minutes. As a team, it was a forgettable shooting night for the Nuggets, as the team shot just 46 percent from the field. Next up, the Nuggets play the second game of a back-to-back when they host the Denver Nuggets on Wednesday, while the Wizards host the Portland Trail Blazers on Friday.

Fig. 6.33: Sample output based on the generation model in Figure 6.31. (Image from [256]).

Journey to the Center of the Earth (also called Jules Verne's Journey to the Center of the Earth) is a 1959 American science fiction adventure film in color by De Luxe, distributed by 20th Century Fox. The film, produced by Charles Brackett and directed by Henry Levin, stars James Mason, Pat Boone, and Arlene Dahl. Bernard Herrmann wrote the film score, and the film's storyline was adapted by Charles Brackett from the 1864 novel of the same name by Jules Verne.

Section Data

Attribute	Value
musical composition	20th Century Fox
PERSON	Jules Verne
dependence syndrome	alcoholic
film genre	adventure film
business	Deluxe Entertainment Services Group, Inc.
based on	A Journey to the Center of the Earth

Article Data

Attribute	Value
instance of	film
director	Henry Levin
composer	Bernard Herrmann
released	1959, 12, 16
genre	science fiction film
genre	fantasy film
starring	James Mason, Pat Boone, Arlene Dahl

Title Data

Document title	Journey to the Center of the Earth (1959 film)	Section title	Introduction

Fig. 6.34: An example from WikiTableT. (Image from [51]).

6.3 Domain Independent Table-to-Text

In recent years, the focus in data-to-text research has shifted significantly toward domain-independent dataset. In this section, we will discuss the recent work in this area.

6.3.1 ToTTo

ToTTo[9] is an open-domain table-to-text dataset with 120,000 training examples for controlled generation from Wikipedia tables. The annotators were instructed to generate an one-sentence description of a region of a given table. Figure 6.35 illustrates the annotation process which consists of (1) showing an original Wikipedia text and table and asking an annotator to highlight the cells mentioned in text, (2) deleting phrases from original text not supported by the highlighed cells, (3) replacing pronouns with named entities from the table, and (4) polishing the produced text by another annotator. Another example from the dataset is shown in Figure 6.36, and some statistics about the dataset, compared to the data-to-text datasets that were available at the time, is given in Figure 6.37.

6.3.2 DART

DART[10] is very similar to ToTTo. It is also based on selected areas from a table and the matching human-written verbalizations of these areas. DART was introduced in [229]. Some statistics about the dataset in comparison with other related datasets are shown in Figure 6.38. Data is collected from both human annotations and automatic annotations. For the former, as shown in Figure 6.39, the annotation is done by (1) collecting parent-child relations between columns (top panel), (2) constructing an ontology from the annotations and selecting the nodes that correspond to the highlighted cells (middle panel), and (3) extracting triples. For the latter, WikiSQL questions are mapped to declarative sentences and the cells are highlighted based on the provided answers and/or SQL queries. This mapping turns text-to-data annotations

[9] https://github.com/google-research-datasets/totto

[10] https://github.com/Yale-LILY/dart

Table Title: Gabriele Becker
Section Title: International Competitions
Table Description: None

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	4x100 m relay	44.60
1994	World Junior Championships	Lisbon, Portugal	12th (semis)	100 m	11.66 (wind: +1.3 m/s)
			2nd	4x100 m relay	44.78
1995	World Championships	Gothenburg, Sweden	7th (q-finals)	100 m	11.54
			3rd	4x100 m relay	43.01

Original Text: 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.
Text after Deletion: she at the 1995 World Championships in both individually and in the relay.
Text After Decontextualization: Gabriele Becker competed at the 1995 World Championships in both individually and in the relay.
Final Text: Gabriele Becker competed at the 1995 World Championships both individually and in the relay.

Fig. 6.35: A datapoint from ToTTo showing the annotation process with the text summarizing the highlighted cells. (Image from [243]).

Table Title: Montpellier
Section Title: Climate
Table Description: *None*

Climate data for Montpellier (1981–2010 averages)													
Month	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Year**
Record high °C (°F)	21.2 (70.2)	22.5 (72.5)	27.4 (81.3)	30.4 (86.7)	35.1 (95.2)	37.2 (99.0)	37.5 (99.5)	36.8 (98.2)	36.3 (97.3)	31.8 (89.2)	27.1 (80.8)	22.0 (71.6)	37.5 (99.5)
Average high °C (°F)	11.6 (52.9)	12.8 (55.0)	15.9 (60.6)	18.2 (64.8)	22.0 (71.6)	26.4 (79.5)	29.3 (84.7)	28.9 (84.0)	25.0 (77.0)	20.5 (68.9)	15.3 (59.5)	12.2 (54.0)	19.9 (67.8)
Daily mean °C (°F)	7.2 (45.0)	8.1 (46.6)	10.9 (51.6)	13.5 (56.3)	17.3 (63.1)	21.2 (70.2)	24.1 (75.4)	23.7 (74.7)	20.0 (68.0)	16.2 (61.2)	11.1 (52.0)	8.0 (46.4)	15.1 (59.2)
Average low °C (°F)	2.8 (37.0)	3.3 (37.9)	5.9 (42.6)	8.7 (47.7)	12.5 (54.5)	16.0 (60.8)	18.9 (66.0)	18.5 (65.3)	15.0 (59.0)	11.9 (53.4)	6.8 (44.2)	3.7 (38.7)	10.4 (50.7)
Record low °C (°F)	−15 (5)	−17.8 (0.0)	−9.6 (14.7)	−1.7 (28.9)	0.6 (33.1)	5.4 (41.7)	8.4 (47.1)	8.2 (46.8)	3.8 (38.8)	−0.7 (30.7)	−5 (23)	−12.4 (9.7)	−17.8 (0.0)
Average precipitation mm (inches)	55.6 (2.19)	51.8 (2.04)	34.3 (1.35)	55.5 (2.19)	42.7 (1.68)	27.8 (1.09)	16.4 (0.65)	34.4 (1.35)	80.3 (3.16)	96.8 (3.81)	66.8 (2.63)	66.7 (2.63)	629.1 (24.77)
Average precipitation days	5.5	4.4	4.7	5.7	4.9	3.6	2.4	3.6	4.6	6.8	6.1	5.6	57.8
Average snowy days	0.6	0.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.7	2.4
Average relative humidity (%)	75	73	68	68	70	66	63	66	72	77	75	76	70.8
Mean monthly sunshine hours	142.9	168.1	220.9	227.0	263.9	312.4	339.7	298.0	241.5	168.6	148.8	136.5	2,668.2
Source #1: Météo France													
Source #2: Infoclimat.fr (humidity and snowy days, 1961–1990)													

Target sentence: Extreme temperatures of Montpellier have ranged from −17.8 °C recorded in February and up to 37.5 °C (99.5 °F) in July.

Fig. 6.36: A datapoint from ToTTo, with "interesting reference language". (Image from [243]).

Dataset	Train Size	Domain	Target Quality	Target Source	Content Selection
Wikibio (Lebret et al., 2016)	583K	Biographies	Noisy	Wikipedia	Not specified
Rotowire (Wiseman et al., 2017)	4.9K	Basketball	Noisy	Rotowire	Not specified
WebNLG (Gardent et al., 2017b)	25.3K	15 DBPedia categories	Clean	Annotator Generated	Fully specified
E2E (Novikova et al., 2017)	50.6K	Restaurants	Clean	Annotator Generated	Partially specified
LogicNLG (Chen et al., 2020)	28.5K	Wikipedia (open-domain)	Clean	Annotator Generated	Columns via entity linking
ToTTo	**120K**	**Wikipedia (open-domain)**	**Clean**	**Wikipedia (Annotator Revised)**	**Annotator highlighted**

Fig. 6.37: Comparing data-to-text datasets. (Image from [243]).

into a data-to-text dataset. An example from DART that uses the table title is shown in Figure 6.40, and one more example is shown in Figure 6.41.

	Input Unit	Examples	Vocab Size	Words per SR	Sents per SR	Tables
WikiTableText	Row	13,318	—	13.9	1.0	4,962
LogicNLG	Table	37,015	122K	13.8	1.0	7,392
ToTTo	Highlighted Cells	136,161	136K	17.4	1.0	83,141
DART	Triple Set	82,191	33.2K	21.6	1.5	5,623

Fig. 6.38: Data-to-text dataset statistics. (Image from [229]).

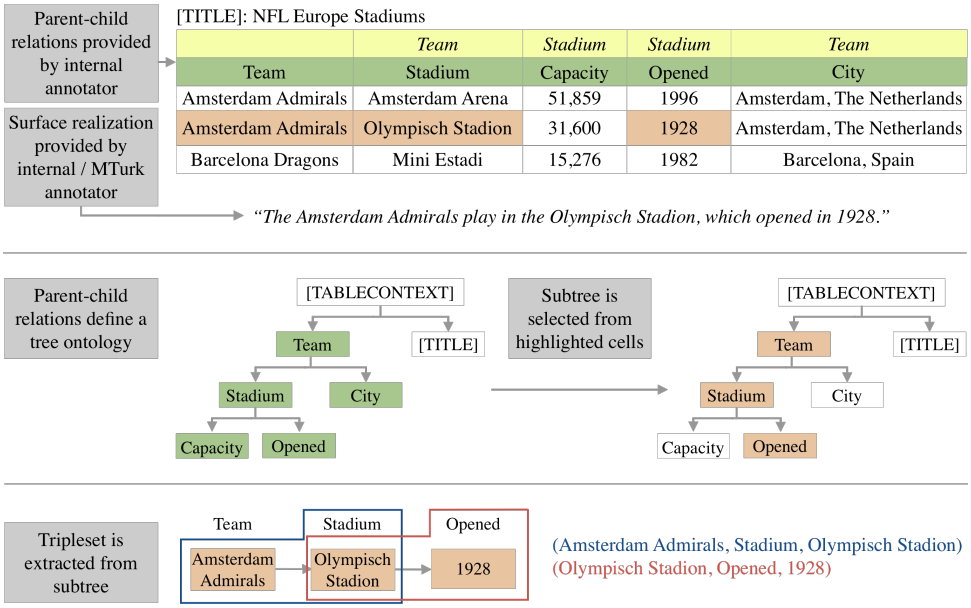


Fig. 6.39: Human annotation interface for DART. (Image from [229]).

6.3.3 FeTaQA

FetaQA was introduced by Nan et al. [228]. As the name suggests, this is a table question-answering dataset with a focus on complex reasoning than simple schema comprehension. The dataset includes 10K Wikipedia-based tuples that include a table, a question, a free-form answer, and the supporting table cells. The data is sampled from ToTTo with some constraints on table size and the number of highlighted rows. Answers were selected from table-grounded sentences in the dataset and the questions were collected from

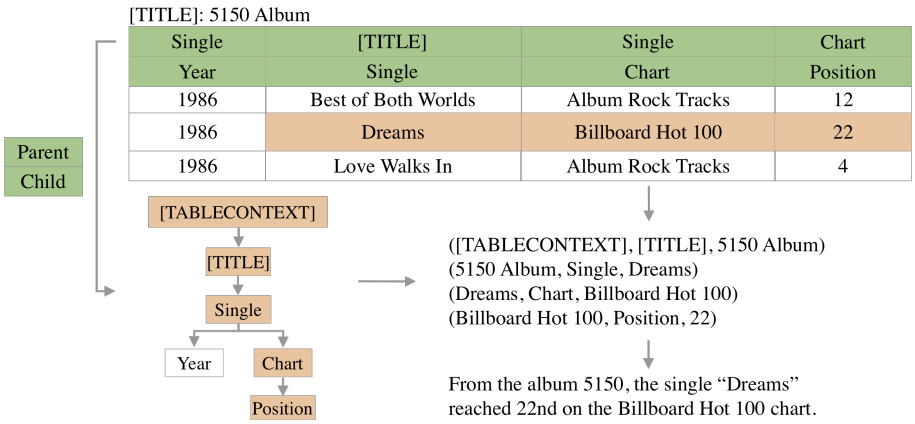


Fig. 6.40: An example from DART that uses the table title. (Image from [229]).

Input triples:
¡H¿ Andrew Rayel ¡R¿ associated Band/associated Musical Artist ¡T¿ Christian Burns
¡H¿ Andrew Rayel ¡R¿ associated Band/associated Musical Artist ¡T¿ Jonathan Mendelsohn
Reference:
andrew rayel , is associated with musical artist jonathan mendelsohn and christian burns .
train on WebNLG - BART-base output:
christian mendelsohn and andrew rayel are both associated with the same band , christian burns .
train on DART - BART-base output:
andrew rayel is associated with christian burns and jonathan mendelsohn .

Fig. 6.41: A sample DART data point. (Image from [229]).

human annotators. The annotators were asked to write questions that were answered by the ToTTo sentence. Examples from FeTaQA are shown in Figure 6.42. The FeTaQA annotation interface is shown in Figure 6.43.

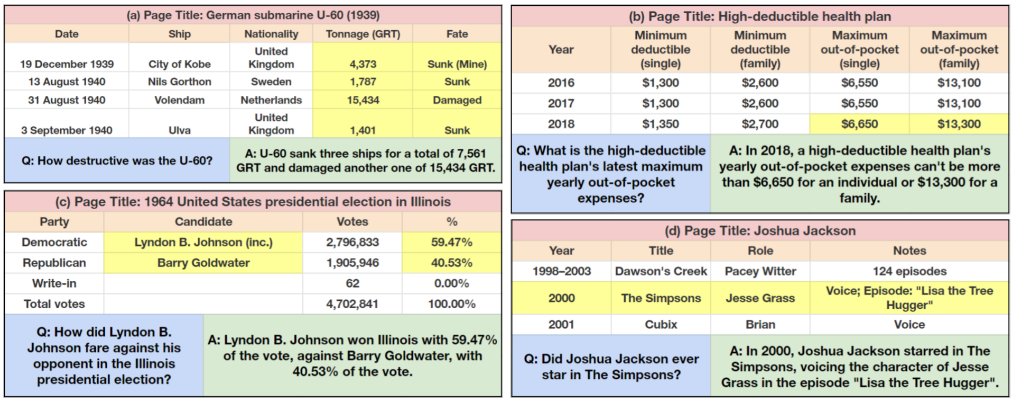


Fig. 6.42: FetaQA examples. (Image from [228]).

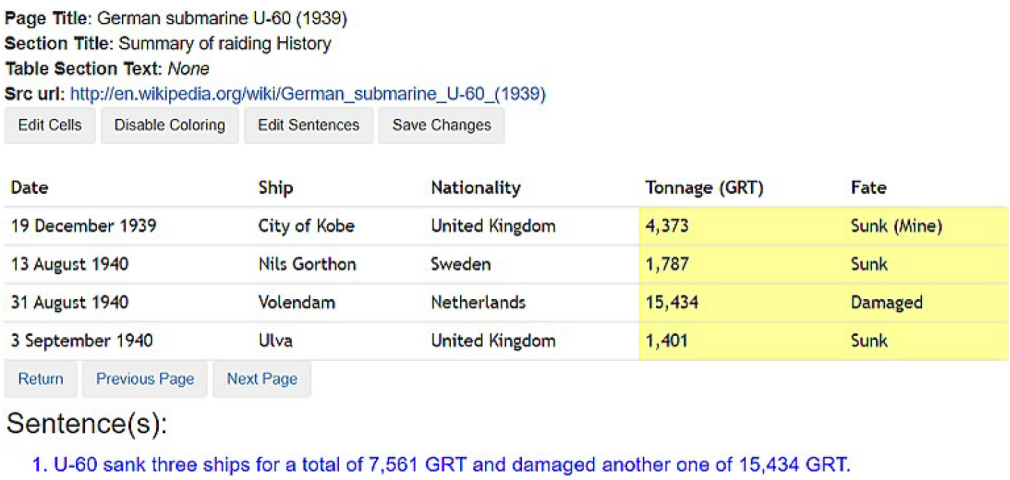


Fig. 6.43: FetaQA Annotation interface. (Image from [228]).

Two benchmark models are suggested for the task, as shown in Figure 6.44: (1) a pipeline model that first parses a question and the table that has the answer into some denotations before generating an answer using a data-to-text model; and (2) an encoder-decoder model based on large pretrained models.

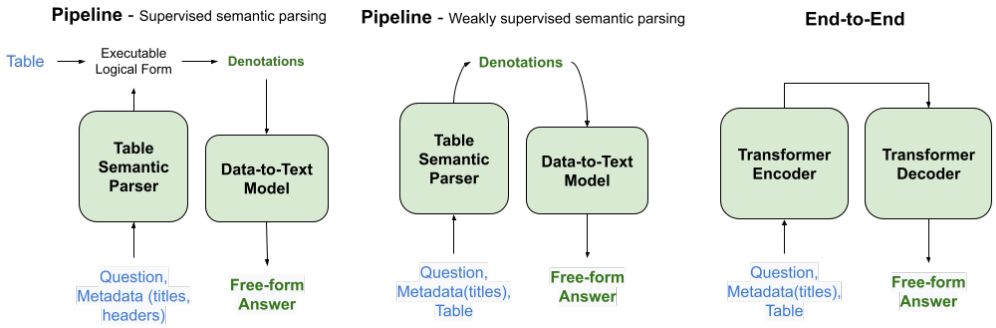


Fig. 6.44: FetaQA model diagrams. (Image from [228]).

6.3.4 TabFact

TabFact[11] [56] introduces fact verification for tables. The dataset includes 16,000 Wikipedia tables and 118,000 manually annotated statements in natural language, labeled as either entailed or refuted. The authors describe two models, Table-BERT and Latent Program Algorithm (LPA). The former is used to

[11] https://github.com/wenhuchen/Table-Fact-Checking

linearize and encode the tables and statements into continuous vectors. Then LPA parses the statements and executes the programs against the tables. An example is shown in Figure 6.45.

United States House of Representatives Elections, 1972

District	Incumbent	Party	Result	Candidates
California 3	John E. Moss	democratic	re-elected	John E. Moss (d) 69.9% John Rakus (r) 30.1%
California 5	Phillip Burton	democratic	re-elected	Phillip Burton (d) 81.8% Edlo E. Powell (r) 18.2%
California 8	George Paul Miller	democratic	lost renomination democratic hold	Pete Stark (d) 52.9% Lew M. Warden , Jr. (r) 47.1%
California 14	Jerome R. Waldie	republican	re-elected	Jerome R. Waldie (d) 77.6% Floyd E. Sims (r) 22.4%
California 15	John J. Mcfall	republican	re-elected	John J. Mcfall (d) unopposed

Entailed Statement

1. John E. Moss and Phillip Burton are both re-elected in the house of representative election.
2. John J. Mcfall is unopposed during the re-election.
3. There are three different incumbents from democratic.

Refuted Statement

1. John E. Moss and George Paul Miller are both re-elected in the house of representative election.
2. John J. Mcfall failed to be re-elected though being unopposed.
3. There are five candidates in total, two of them are democrats and three of them are republicans.

Fig. 6.45: TabFact example. (Image from [56]).

6.3.5 LogicNLG

LogicNLG was introduced in [54]. This innovative project goes beyond surface-level description of records and focuses on generating statements that can be logically entailed from the data in open-domain semi-structured tables. A sample data point is shown in Figure 6.46. Compared with other similar datasets, as shown in Figure 6.47, LogicNLG offers rich logical inference over tables. Also, the diversity of the schema introduces some challenges for rule-based models . Sample outputs for a given table can be seen in Figure 6.48. Finally, Figure 6.49 illustrates the evaluation of semantic parsing in the paper.

Medal Table from Tournament

Nation	Gold Medal	Silver Medal	Bronze Medal	Sports
Canada	3	1	2	Ice Hockey
Mexico	2	3	1	Baseball
Colombia	1	3	0	Roller Skating

Surface-level Generation

Sentence: Canada has got 3 gold medals in the tournament.
Sentence: Mexico got 3 silver medals and 1 bronze medal.

Logical Natural Language Generation

Sentence: Canada obtained 1 more gold medal than Mexico.
Sentence: Canada obtained the most gold medals in the game.

Fig. 6.46: Sample from LogicNLG. (Image from [54]).

	Vocab	Examples	Vocab/Sent	Tables	Domain	Source	Inference	Schema
WEATHERGOV	394	22.1K	0.01	22.1K	Weather	Crawled	No	Known
WikiBIO	400K	728K	0.54	728K	Biography	Crawled	No	Limited
ROTOWIRE	11.3K	4.9K	0.72	4.9K	NBA	Annotated	Few	Known
LOGICNLG	122K	37.0K	**3.31**	7.3K	**Open**	Annotated	**Rich**	**Unlimited**

Fig. 6.47: Statistics about LogicNLG and related datasets. (Image from [54]).

player	country	year (s) won	total	to par
larry nelson	united states	1981 , 1987	152	+ 8
jack nicklaus	united states	1963 , 1971 , 1973 1975 , 1980	152	+ 8
lee trevino	united states	1974 , 1984	152	+ 8
hubert green	united states	1985	153	+ 9
lanny wadkins	united states	1977	155	+ 11
dave stockton	united states	1970 , 1976	157	+ 13

larry nelson , jack nicklaus , and lee trevino all shot 8 strokes over par
larry nelson , lee trevino , and dave stockton each won two pga championships in the 1970s - 1980s
jack nicklaus had more pga championship wins than larry nelson and lee tevino combined
dave stockton shot five strokes worse than larry nelson , jack nicklaus , and lee trevino
three golfers shot worse than 8 strokes over par

Fig. 6.48: Statements generated as part of LogicNLG. (Image from [54]).

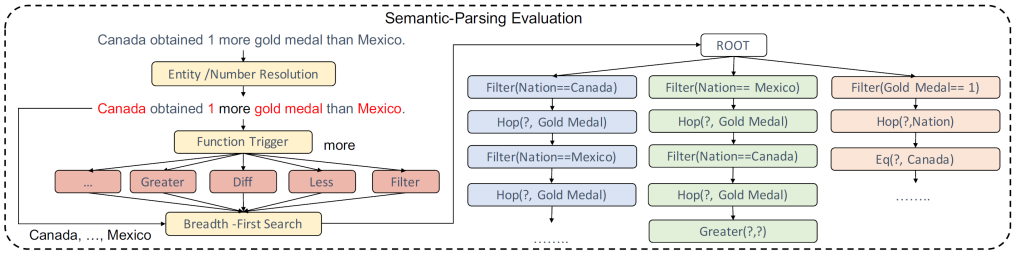


Fig. 6.49: Semantic Parsing Evaluation for LogicNLG. (Image from [54]).

6.3.6 Logic2Text

Logic2Text[12] [59] addresses the problem of generating descriptions of tables that require inference across records. They formulate such a task as generation from logical forms. The dataset includes 10,753 descriptions, paired with the corresponding logical forms. Figure 6.50 provides an illustration.

[12] https://github.com/czyssrs/Logic2Text

table caption: opec

country	region	joined opec	population (july 2012)	area (km square)
algeria	africa	1969	37367226	2381740
angola	africa	2007	18056072	1246700
iraq	middle east	1960	31129225	437072
libya	africa	1962	5613380	1759540
nigeria	africa	1971	170123740	923768
...	...	...	...	...

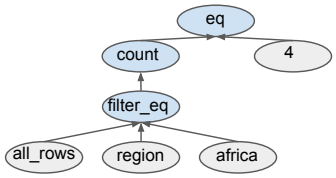
Surface-level NLG

Description: angola, from the region africa, joined opec in 2007, with an population of 18056072 in 2012.

Description: algeria, from the region africa, joined opec in 1969, with an population of 37367226 in 2012.

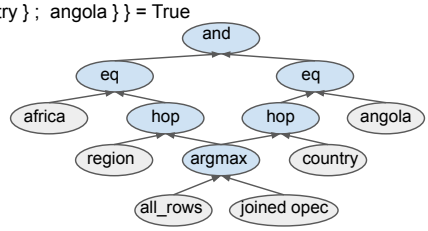
Logical-level NLG with logical forms (our dataset)

logical form: eq { count { filter_eq { all_rows ; region ; africa } } ; 4 } = True



Description: In 2012 in opec, there were 4 member countries from africa.

logical form: and { eq { hop { argmax { all_rows ; joined opec } ; region } ; africa } ; eq { hop { argmax { all_rows ; joined opec } ; country } ; angola } } = True



Description: In 2012 in opec, angola, from africa, was the latest country to join.

Fig. 6.50 A sample pair of a table and a description. The bottom part of the figure shows outputs that are based on the logical form. (Image from [59]).

6.3.7 GEM

The **GEM** benchmark [115] is a more recent task that includes many datasets, such as the aforementioned WebNLG, DART, and ToTTo, as well as the summarization tasks MLSum [277], XSum [230], and WikiLingua [180]. Figure 6.51 lists these datasets. A multilingual version, GEM v.2, was recently released [116] and it covers 40 datasets in 51 languages.

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 Jurčíček, 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.	*en/es/ru/tr/vi	*175k	Article

Fig. 6.51: Datasets included in GEM v.1. (Image from [115]).

6.4 Pretraining for Tables

When dealing with hybrid data (e.g., tables and text), it is important to learn joint representations of different modalities (see Figure 6.52 for the structure of a Wikipedia table). Many recent language models for tables have focused on just that. These transformer-based tabular models (TaLMs), usually built on top of pre-trained models such as BERT and T5, not only learn a join representation of content and structure but also inherit some of the semantic and text understanding features of the underlying pre-trained models.

Many of TaLMs have adopted an encoder architecture to learn a contextual representation of tables. These include TaPas [139], TaBERT [363], TURL [88], GraPPa [366], TABBIE [148], TUTA [334], TableFormer [360] and others. Some TaLMs have used an encoder-decoder architecture to better support text generation tasks such as table-to-text. These include KGPT [55], RPT [307], TaPEx [206], UnifiedSKG [349], STTP [350], etc. TableGPT [121] adopts a decoder architecture and is fine-tuned on GPT-2.

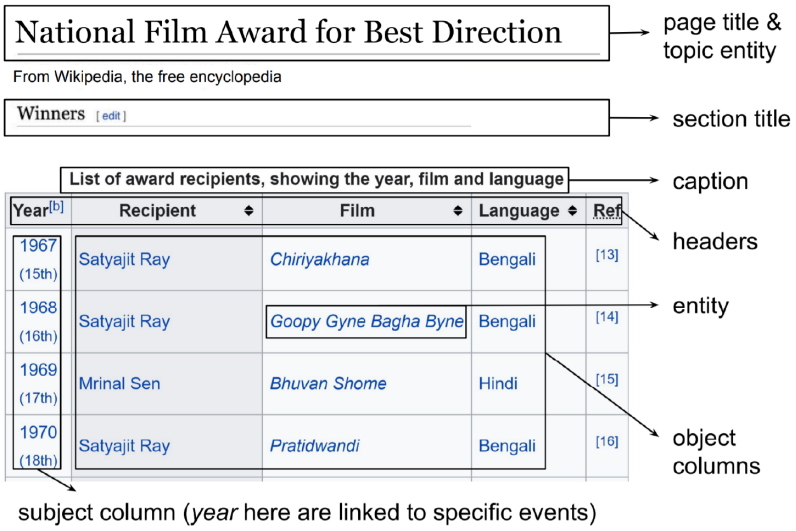


Fig. 6.52: Wikipedia Table Structure. (Image from [88]).

We will look at models for pre-training from tables and text such as TaPas [139], TaBERT [363], GraPPa [366], and others. See [93] for a partial list of such model architectures.

6.4.1 TURL

TURL[13] [88] introduces a pre-training and finetuning approach to relational tables. The pre-training stage is used to learn deep contextualized representations of the tables. Different table components are separately encoded and are fused together. A structure-aware transformer encoder is used to model the table rows and columns. A Masked Entity Recovery objective is used to pre-train the system, allowing each element to attend to other table elements on the same row or column. The system is evaluated on six table understanding tasks, including cell filling, row population, entity linking, column type annotation and relation extraction. The TURL architecture is shown in Figure 6.53.

6.4.2 TUTA

TUTA [334] describe a unified pre-training architecture for tables. In order to understand a table, the paper introduces a tree-based structure used to describe the spatial and hierarchical information in tables, followed by tree-based attention and position embeddings. TUTA [14] is evaluated on five datasets for cell type classification and table type classification. An example from TUTA is shown in Figure 6.54. The TUTA architecture is shown in Figure 6.55.

[13] https://github.com/sunlab-osu/TURL

[14] https://github.com/microsoft/TUTA_table_understanding/

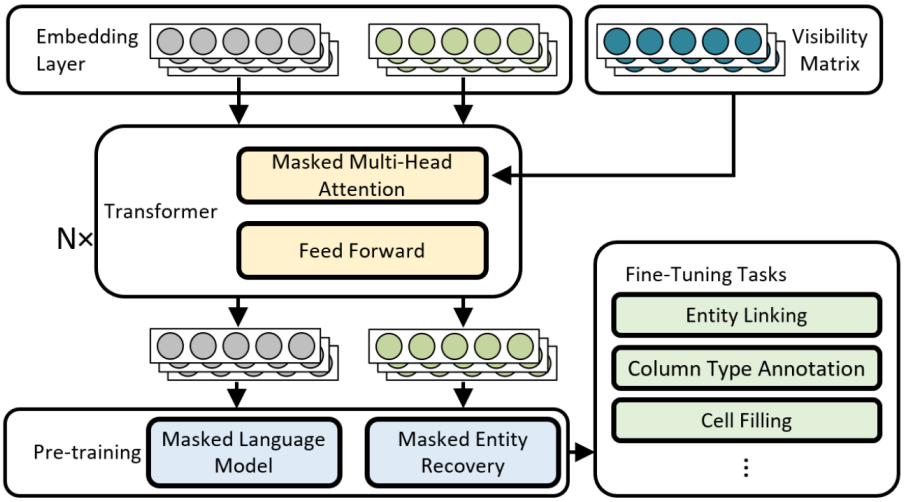


Fig. 6.53: Architecture of TURL. (Image from [88]).

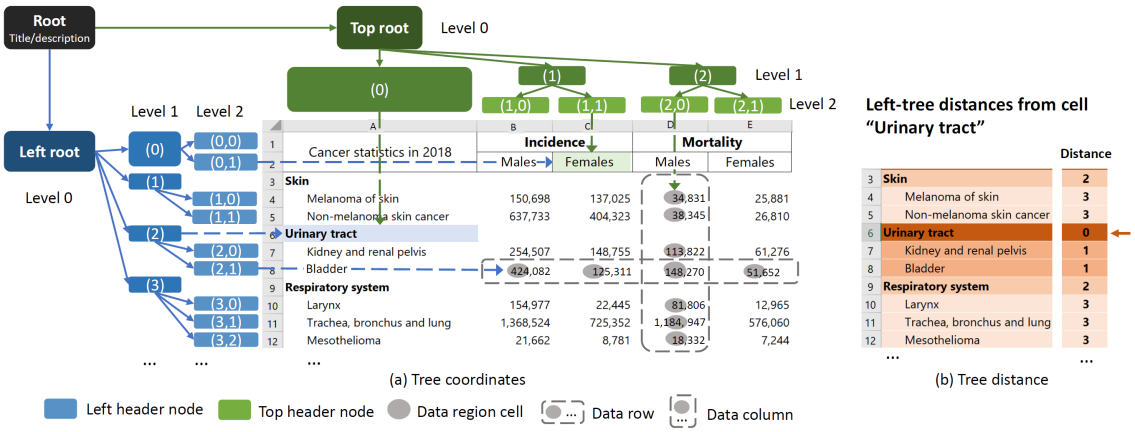


Fig. 6.54: Tree coordinates and tree distance for tables, as used in TUTA. (Image from [334]).

6.4.3 Tapas

TAPAS [139] is based on the idea of adding table layout information to the pretrained model. It modifies the BERT architecture so that it can be pretrained on tables and the matching text segments. It also uses positional embeddings to encode the table structure. It employs weak supervision in order to use the tables without the need of intermediate logical forms. The denotations are predicted by a selection of table cells. An example of the TaPaS model is shown in Figure 6.56. Figure 6.57 shows an internal representation, while Figure 6.58 shows a sample table from TaPaS and its corresponding questions.

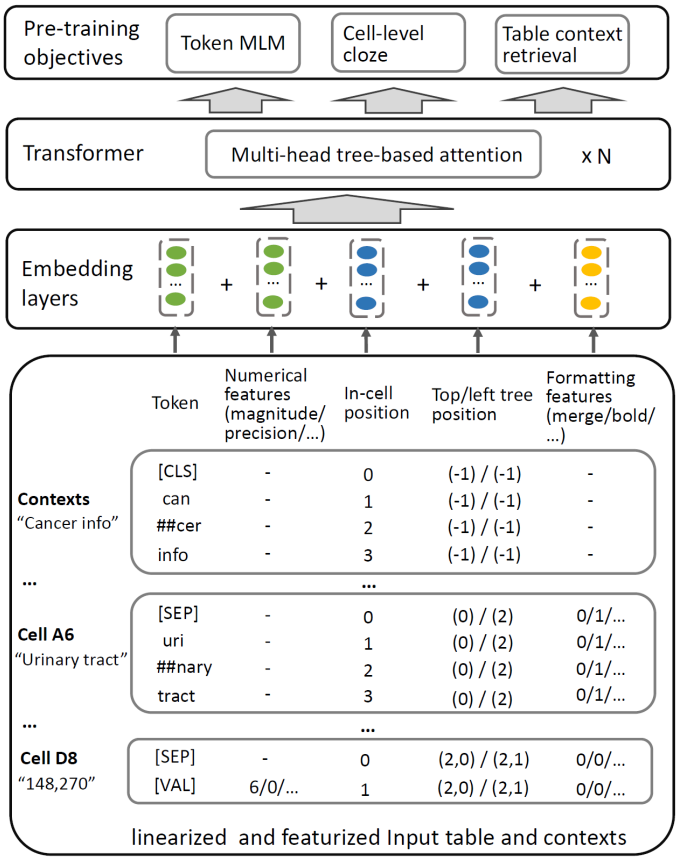


Fig. 6.55 The architecture of TUTA. (Image from [334]).

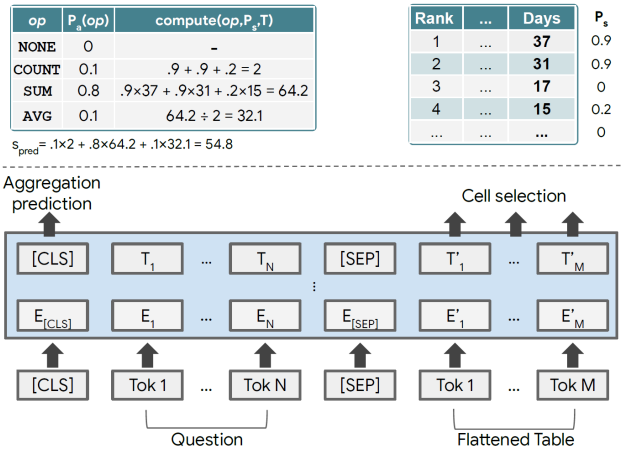


Fig. 6.56 TaPaS model. (Image from [139]).

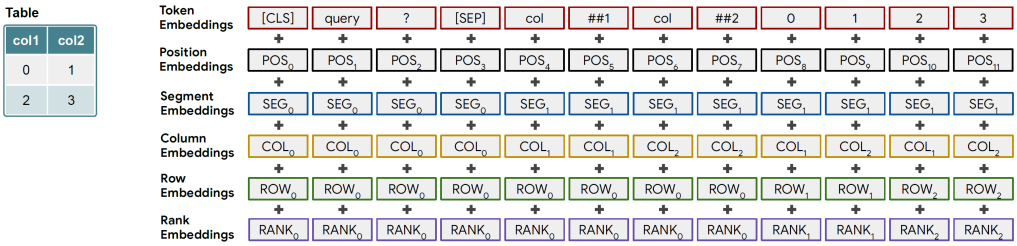


Fig. 6.57: Encoding of the question "query?" in TaPaS. (Image from [139]).

Table

Rank	Name	No. of reigns	Combined days
1	Lou Thesz	3	3,749
2	Ric Flair	8	3,103
3	Harley Race	7	1,799
4	Dory Funk Jr.	1	1,563
5	Dan Severn	2	1,559
6	Gene Kiniski	1	1,131

Example questions

#	Question	Answer	Example Type
1	*Which wrestler had the most number of reigns?*	Ric Flair	Cell selection
2	*Average time as champion for top 2 wrestlers?*	AVG(3749,3103)=3426	Scalar answer
3	*How many world champions are there with only one reign?*	COUNT(Dory Funk Jr., Gene Kiniski)=2	Ambiguous answer
4	*What is the number of reigns for* Harley Race?	7	Ambiguous answer
5	*Which of the following wrestlers were ranked in the bottom 3?*	{Dory Funk Jr., Dan Severn, Gene Kiniski}	Cell selection
	Out of these, who had more than one reign?	Dan Severn	Cell selection

Fig. 6.58: A table from TaPaS and some of the corresponding questions. (Image from [139]).

6.4.4 TaBERT

The semantic parser used by **TaBERT** [363] was built on top of Bert and then trained on 26 million linearized tables and their contexts in English. It achieved state of the art on the WikiTableQuestions dataset and also did very well on the Spider dataset. Figure 6.59 shows how TaBert represents utterances and tables.

6.4.5 Grappa

Grappa [366] uses a Synchronous Context-Free Grammar (SCFG) to generate synthetic question-SQL pairs. It can learn compositional inductive bias through techniques commonly used for semantic parsing, namely masked language modeling (MLM) on joint table+language datasets. Combined with semantic parsers, Grappa achieves SOTA on four tasks: {full supervised, weakly supervised} x {WikiSQL, WikiTableQuestions}. Figure 6.60 shows the Grappa pre-training approach, and Figure 6.61 shows some of the non-terminals and production rules using the SCFG for generating synthetic training data for Grappa.

Unlike TaBERT and TaPas, which are trained on millions of web tables aligned with noisy nearby content, Grappa is trained on synthetic data plus a smaller but cleaner set of existing text+table datasets. The four table-based semantic parsing and question answering datasets used in Grappa are shown in Figure 6.62. The top two are fully supervised, while the other two are weakly-supervised. Figure 6.63 shows examples of the inputs and annotations for the four semantic parsing tasks.

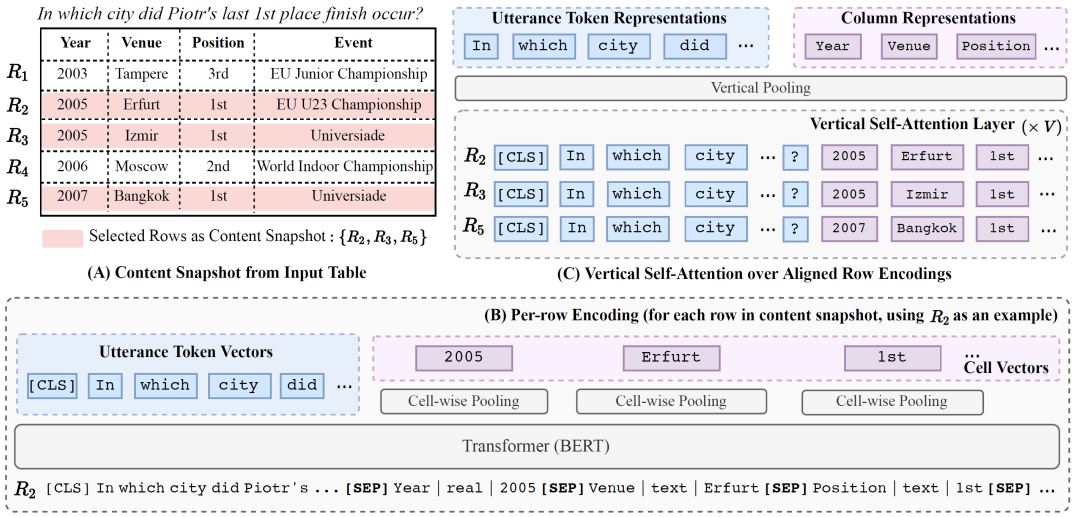


Fig. 6.59: TaBert representations of utterances and tables from WikiTableQuestions. (Image from [363]).

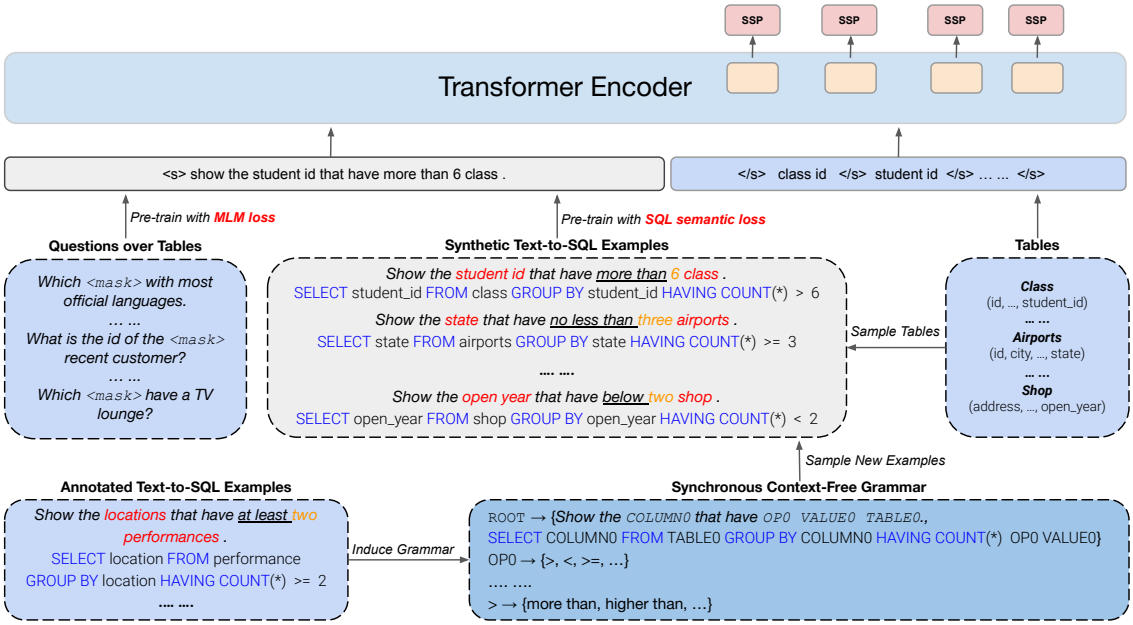


Fig. 6.60: Grappa pre-training approach. (Image from [366]).

Non-terminals	Production rules
TABLE $\rightarrow t_i$ COLUMN $\rightarrow c_i$ VALUE $\rightarrow v_i$ AGG $\rightarrow$ ⟨ MAX, MIN, COUNT, AVG, SUM⟩ OP $\rightarrow$ ⟨ =, ≤, ≠, ... , LIKE, BETWEEN ⟩ SC $\rightarrow$ ⟨ ASC, DESC ⟩ MAX $\rightarrow$ ⟨"maximum", "the largest"...⟩ ≤ $\rightarrow$ ⟨"no more than", "no above"...⟩ ...	1. ROOT $\rightarrow$ ⟨"For each COLUMN0 , return how many times TABLE0 with COLUMN1 OP0 VALUE0 ?", `SELECT COLUMN0 , COUNT ( * ) WHERE COLUMN1 OP0 VALUE0 GROUP BY COLUMN0` ⟩ 2. ROOT $\rightarrow$ ⟨"What are the COLUMN0 and COLUMN1 of the TABLE0 whose COLUMN2 is OP0 AGG0 COLUMN2 ?", `SELECT COLUMN0 , COLUMN1 WHERE COLUMN2 OP0 ( SELECT AGG0 ( COLUMN2 ) )` ⟩

Fig. 6.61: Examples of non-terminals and production rules in the SCFG for Grappa. (Image from [366]).

Task & Dataset	# Examples	Resource	Annotation	Cross-domain
SPIDER Yu et al. (2018b)	10,181	database	SQL	✓
Fully-sup. WIKISQL Zhong et al. (2017)	80,654	single table	SQL	✓
WIKITABLEQUESTIONS Pasupat & Liang (2015)	2,2033	single table	answer	✓
Weakly-sup. WIKISQL Zhong et al. (2017)	80,654	single table	answer	✓

Fig. 6.62: The four table-based semantic parsing and question answering datasets used in Grappa. (Image from [366]).

Task	Question	Table/Database	Annotation
SPIDER	Find the first and last names of the students who are living in the dorms that have a TV Lounge as an amenity.	database with 5 tables e.g.`student`, `dorm_amenity`, ...	SELECT T1.FNAME, T1.LNAME FROM STUDENT AS T1 JOIN LIVES_IN AS T2 ON T1.STUID=T2.STUID WHERE T2.DORMID IN (SELECT T3.DORMID FROM HAS_AMENITY AS T3 JOIN DORM_AMENITY AS T4 ON T3.AMENID=T4.AMENID WHERE T4.AMENITY_NAME= 'TV LOUNGE')
Fully-sup. WIKISQL	How many CFL teams are from York College?	a table with 5 columns e.g. `player`, `position`, ...	SELECT COUNT CFL TEAM FROM CFLDRAFT WHERE COLLEGE = 'YORK'
WIKITABLEQUESTIONS	In what city did Piotr's last 1st place finish occur?	a table with 6 columns e.g. `year`, `event`, ...	"Bangkok, Thailand"
Weakly-sup. WIKISQL	How many CFL teams are from York College?	a table with 5 columns e.g. `player`, `position`,...	2

Fig. 6.63: Examples from the four semantic parsing tasks. (Image from [366]).

6.4.6 Tabbie

Tabbie[148] holds a different goal from previous work such as TaPas, TaBERT, and Grappa. It is intended to predict corrupted cell values and therefore is not trained on text. Tabbie's corrupted cell prediction objective is based on ELECTRA. It uses two transformers, one for rows and another one for columns, as can be seen in Figure 6.64. Tabbie is evaluated on three tasks that measure the level of semantic understanding of tables (in the absence of associated text), namely column population, row population, and column type prediction.

6.4.7 Other recent papers

TableFormer was introduced by [360]. It represents a structure-aware encoding architecture that transforms table structure information into attention. This structure is invariant to both row and column order. It performed well on the SQA, WikiTableQuestions, and TabFact datasets, and did especially well in settings in which the rows or columns were shuffled. The architecture of TableFormer is shown in Figure 6.66.

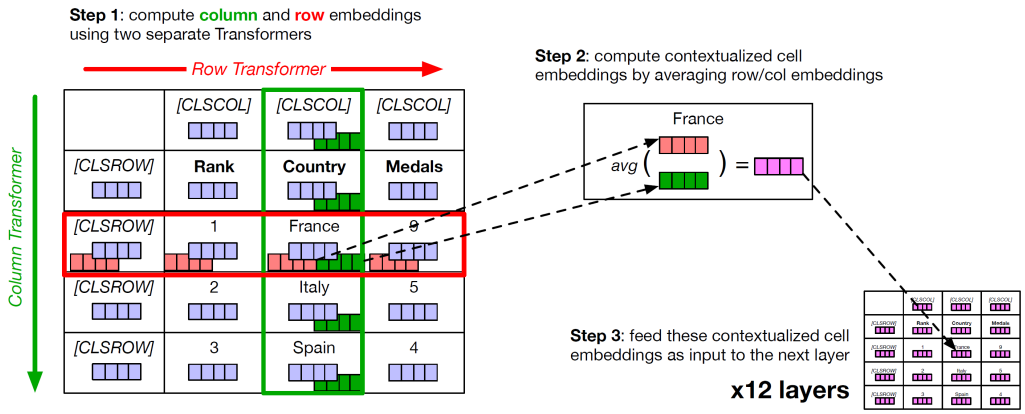


Fig. 6.64: Row and column transformers in Tabbie. (Image from [148]).

Nation	Gold	Silver	Bronze
Great Britain	2	1	2
Spain	**1**	**2**	**0**
Ukraine	**0**	**2**	**0**

Question: Which nation received 2 silver medals?
Gold Answer: Spain, Ukraine
TAPAS: Spain
TABLEFORMER: Spain, Ukraine
TABLEFORMER w/o a proposed structural bias: Spain

Fig. 6.65: An example that shows how TableFormer does better than TAPAS. (Image from [360]).

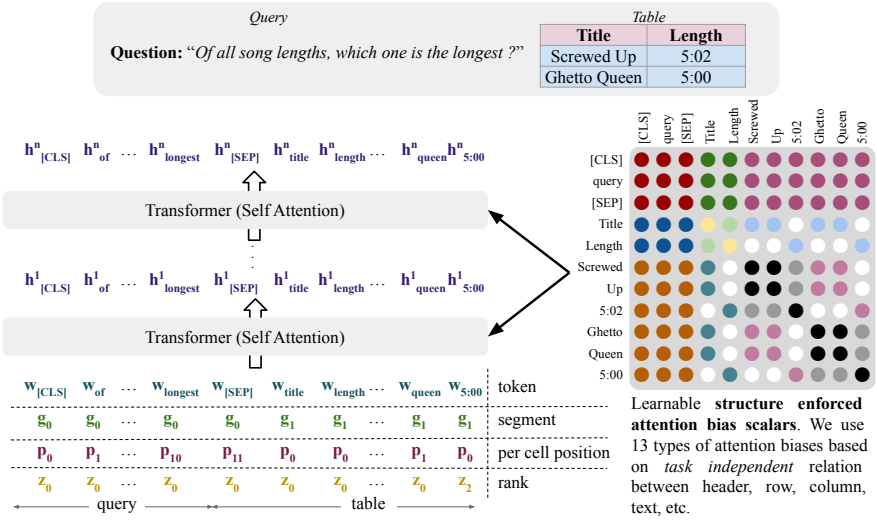


Fig. 6.66: TableFormer architecture. (Image from [360]).

As a follow up to TaPaS, Herzig et al. [138] study open domain question answering over tables via dense retrieval. They use a retriever trained to deal with tabular data. Due to the lack of datasets in this domain, they create a new task based on a subset of the Natural Questions [179] dataset.

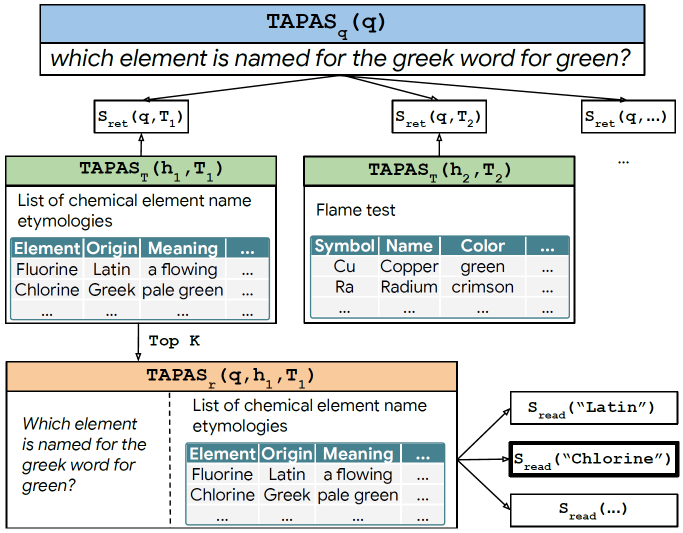


Fig. 6.67 A dense table retriever. (Image from [138]).

TaPEx[15] [206] trains a neural SQL program executor using a synthetic corpus that includes SQL queries and their execution outputs. The method achieves strong results on WikiSQL, WikiTableQuestions, SQA, and TabFact. An example of TaPEx is shown in Figure 6.68.

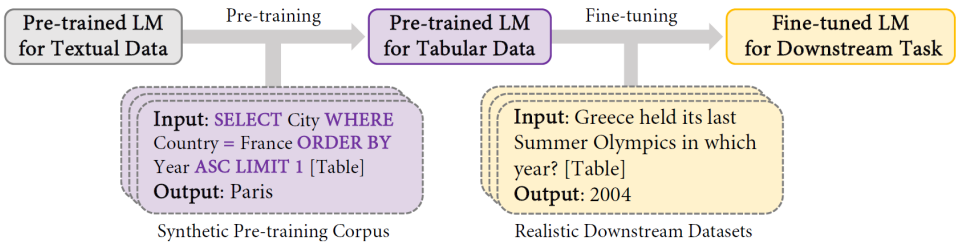


Fig. 6.68: TaPEx Example. (Image from [206]).

Another paper that examines pretraining and finetuning strategies for data-to-text tasks is [162]. The authors show that text-to-text pretraining as in the T5 model [264] can outperform end-to-end neural architectures specifically designed for data-to-text generation. They obtain strong results on out-of-domain datasets. The authors conduct their experiments on three datasets, with the sizes shown in Figure 6.69. Each dataset has a different type of structured data (e.g., tables in ToTTO, meaning representation in MultiWoz and triples in WebNLG), and examples of each are shown in Figure 6.70 (WebNLG), and Figure 6.71 (ToTTo).

[15] https://github.com/microsoft/Table-Pretraining

Dataset	Train	Dev	Test
WebNLG	18.1K	2.2k	4.9k
ToTTo	120K	7.7k	7.7k
Multiwoz	56.8K	7.3k	7.3k

Fig. 6.69 Dataset sizes. (Image from [162]).

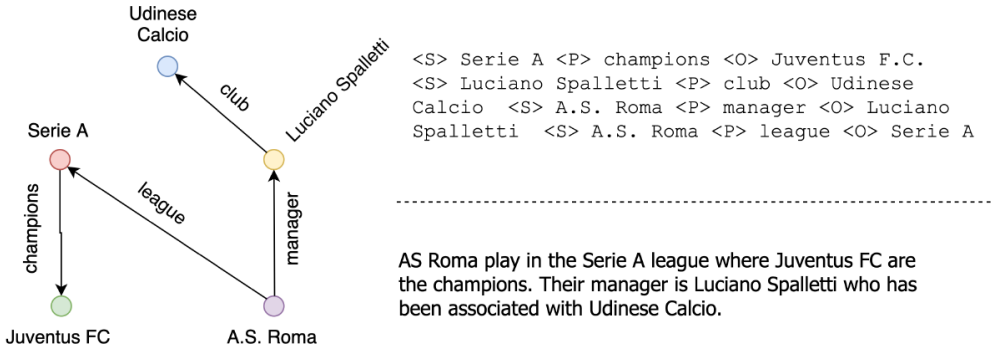


Fig. 6.70: WebNLG example. (Image from [162]).

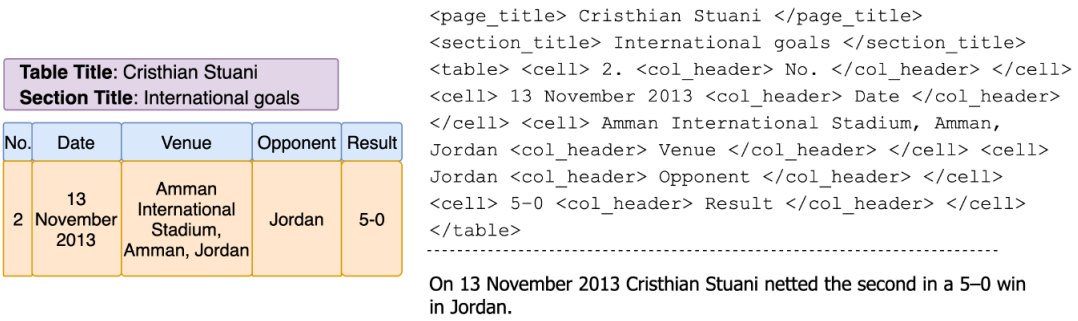


Fig. 6.71: ToTTo example. (Image from [162]).

A recent paper[16] [335] claims that models do not really have to be designed with tables in mind. They look at the task of table retrieval and demonstrate that a simple, text-based model can achieve results comparable to these of table-specific models. They compare a generic dense passage retriever, which when fine-tuned on linearized tables, performs favorably compared to a dense table retriever on the NQ-table dataset [138], a table-focused subset of the Natural Questions dataset [179].

HybridQA[17] [57] is a large-scale QA dataset that involves reasoning on heterogeneous information, both Wikipedia tables and free-form text linked to the entities in tables. Figure 6.72 and Figure 6.73 show examples of HybridQA datapoints.

Another recent QA dataset is **OTT-QA**[18] [53]. The acronym stands for Open Table-and-Text Question Answering. Answering such questions requires multi-hop inferences across tables and text. The methods

[16] https://github.com/zorazrw/nqt-retrieval
[17] https://github.com/wenhuchen/HybridQA
[18] https://github.com/wenhuchen/OTT-QA

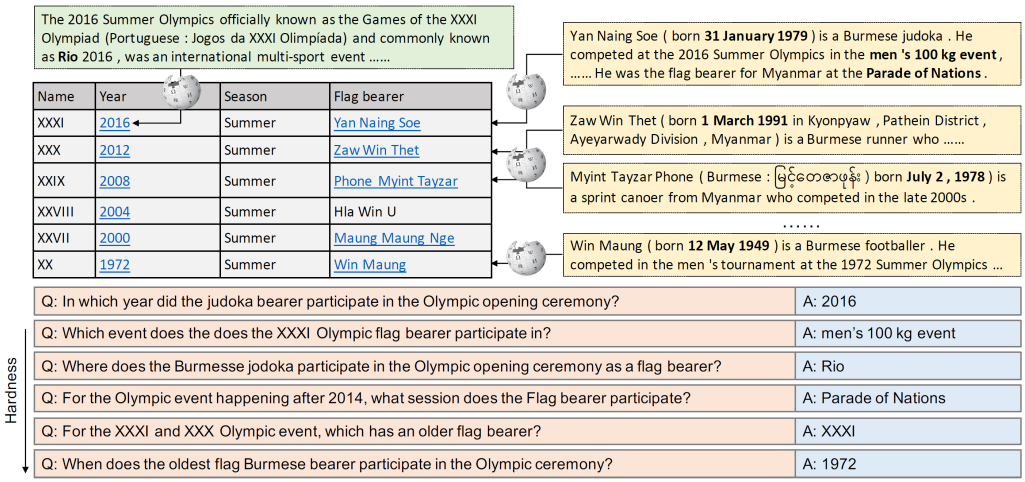


Fig. 6.72: Annotated question answering pairs in HybridQA. (Image from [57]).

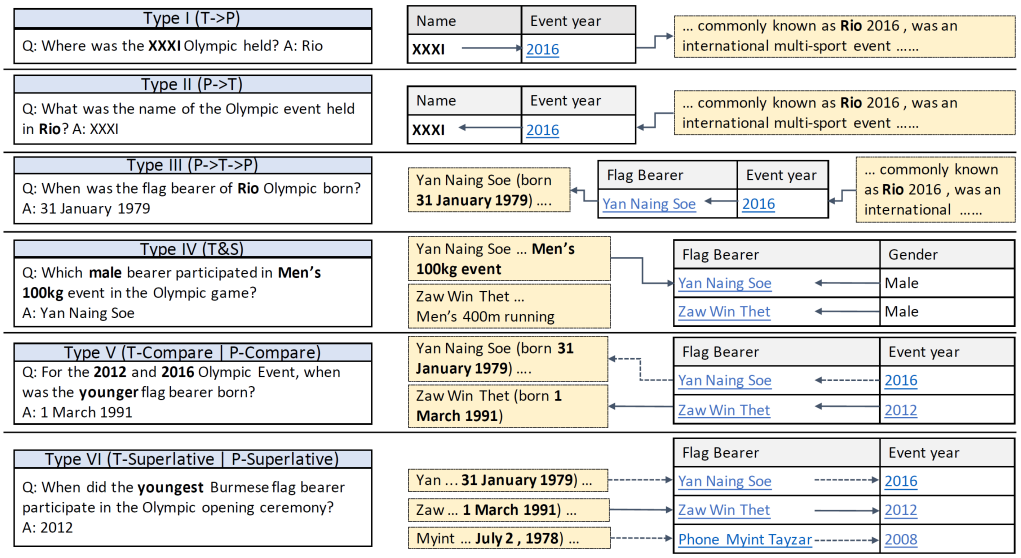


Fig. 6.73: Multi-hop questions in HybridQA. (Image from [57]).

used in the paper include "arly fusion" that combines related table and text unites into a single block, and "cross-block reader" which applies global-local sparse attention. The best results are obtained when the two methods are used in combination. An example is included in Figure 6.74. Some of the matching questions are shown in Figure 6.75.

Turning Tables [364] is a recent paper that adds reasoning to semi-structured tables. They describe a method for the automatic generation of question-paragraph pairs. These pairs require multiple reasoning skills such as fact composition and number comparison. Figure 6.76 shows an example.

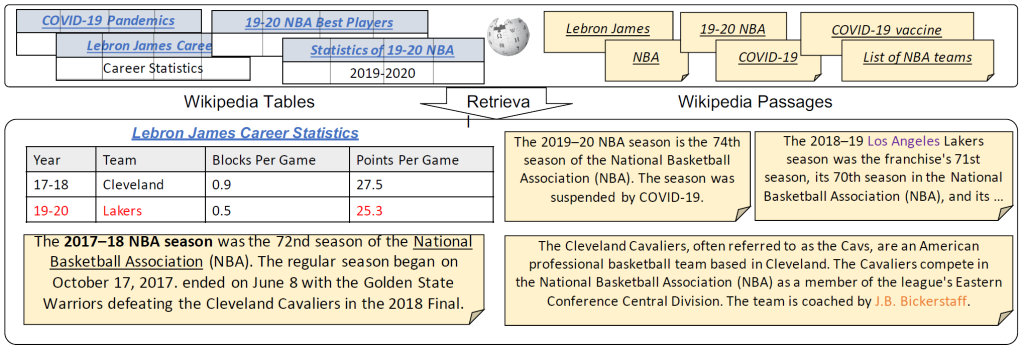


Fig. 6.74: Multi-hop reasoning over two candidate pools. (Image from [53]).

Q1: How many points per game did Lebron James get in the COVID-19 NBA Season? A1: COVID-19 -> 19-20 Season -> 25.3

Q2: Who is the coach of the team that Lebron James played in to achieve his highest score in his career? A2: 27.5 -> Cleveland -> J. B. Bickerstaff

Q3:What suspends the NBA season during which Lebron James has an average points per game of 25.3? A3: 25.3 -> 19-20 Season -> COVID-19

Q4:For the season suspended by COVID-19 and the season defeated by Warriors in Final, which season has Lebron obtained more points? A4: 25.3 < 27.5 -> 17-18

Fig. 6.75: More sample questions. (Image from [53]).

Round	Date	Opponent	Venue	Result	Attendance
R3	31 October 1990	Portsmouth	H	0–0	16,699
R3R	6 November 1990	Portsmouth	A	3–2	16,085
R4	28 November 1990	Oxford United	A	2–1	9,789
QF	16 January 1991	Tottenham Hotspur	H	0–0	34,178
QFR	23 January 1991	Tottenham Hotspur	A	3–0	33,861
SF 1st Leg	24 February 1991	Sheffield Wednesday	H	0–2	34,074
SF 2nd Leg	27 February 1991	Sheffield Wednesday	A	1–3	34,669

Composition: q: *What was the Result when the Round was R4?* **c:** The Date when the Round was R4 was 28 November 1990. The Result when the Date was 28... **a: 2-1**

Comparison q: *Which Round had a higher Attendance: QF or QFR?* **c:** The Attendance when the Round was QF was 34,178. The Attendnace when the Round was QFR... **a: QF**

Date Difference: q: *The Opponent was Portsmouth how much time before the Opponent was Sheffield Wednesday?* **c:** The Date when the Opponent... **a: 3 months and 18 days**

Fig. 6.76: An example of a table and automatically generated question-context answer triples. (Image from [364]).

HiTab[19] [61] moves beyond QA and NLG for flat tables and focuses on questions over hierarchical tables (e.g., Figure 6.77). Figure 6.78 shows a HiTab example.

	A	B	C	D	E	F	G
1	**TABLE 3.** Primary source and mechanism of support for full-time master's and doctoral students in science and engineering: 2017						
2	Source and mechanism	All full-time graduate students		Master's		Doctoral	
3		Total	Percent	All	Percent	All	Percent
4	**All full-time**	**433,916**	**100.0**	**209,221**	**100.0**	**224,695**	**100.0**
5	Self-support	161,641	37.3	139,373	66.6	22,268	9.9
6	All sources of support	272,275	62.7	69,848	33.4	202,427	90.1
7	Federal	65,999	15.2	10,736	5.1	55,263	24.6
8	Department of Agricu	2,361	0.5	938	0.4	1,423	0.6
9	Department of Defens	8,089	1.9	2,568	1.2	5,521	2.5
16	Other	9,098	2.1	3,462	1.7	5,636	2.5
17	Institutional	182,135	42.0	52,319	25.0	129,816	57.8
18	Other U.S. source	19,432	4.5	5,136	2.5	14,296	6.4
19	Foreign	4,709	1.1	1,657	0.8	3,052	1.4
20	All mechanisms of support	272,275	62.7	69,848	33.4	202,427	90.1
21	Fellowships	39,368	9.1	5,687	2.7	33,681	15.0
22	Traineeships	10,945	2.5	1,497	0.7	9,448	4.2
23	Research assistantships	103,586	23.9	19,702	9.4	83,884	37.3
24	Teaching assistantships	84,499	19.5	22,171	10.6	62,328	27.7
25	Other mechanisms	33,877	7.8	20,791	9.9	13,086	5.8

Target text:
For doctoral students, the proportion of support from research assistantships is 10 points higher than that from teaching assistantships.

Highlighted cells:
From entity alignment: Doctoral, percent, research assistantships, teaching assistantships. From quantity alignment: 37.3, 27.7

Operators:
DIFF

Input sequence after sub table selection and serialization:
[SEP] source and mechanism [SEP] doctoral [SEP] percent [SEP] all mechanisms of support [SEP] research assistantships [SEP] 37.3 [SEP] teaching assistantships [SEP] 27.7 [SEP] DIFF [SEP] 9.6

Fig. 6.77 Sample hierarchical table from HiTab. (Image from [61]).

R2D2 [227] is another recent paper. It focuses on unfaithful data-to-text generation by training a system as both generator and faithfulness discriminator. The resulting R2D2 system achieves state-of-the-art results on FeTaQA, LogicNLG, and ToTTo. An example can be seen in Figure 6.79.

The last work that we want to mention is **NeuralDB**[20] [313, 314]. The idea here is to encode all the data into plain text instead of a relational schema. This is a novel approach to answering database queries over plain text representations of tabular data. Their approach scales to databases containing thousands of facts. An example is shown in Figure 6.80.

[19] https://github.com/microsoft/HiTab

[20] https://github.com/facebookresearch/NeuralDB

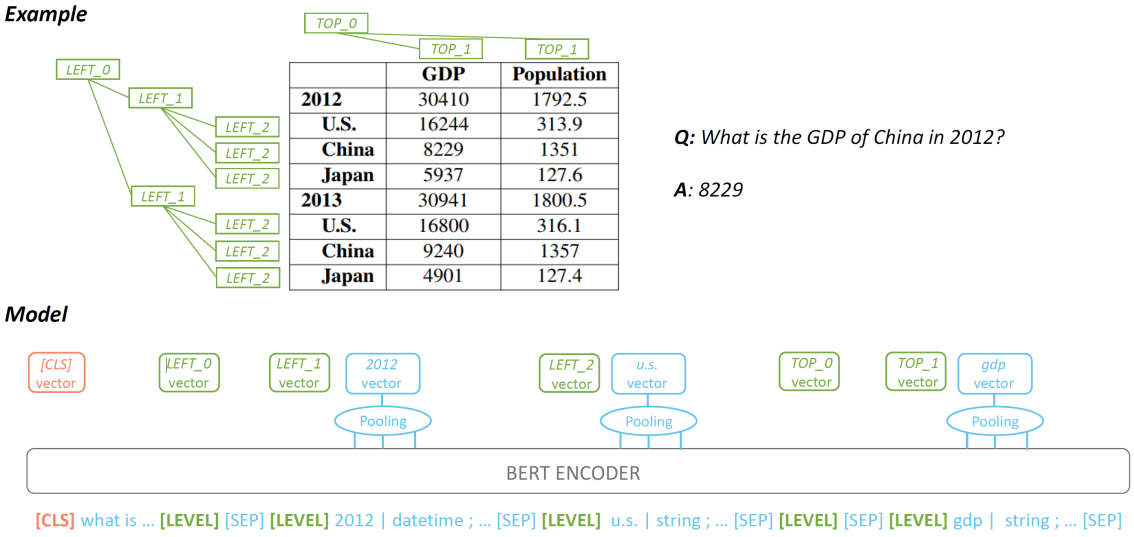


Fig. 6.78: HiTab example. (Image from [61]).

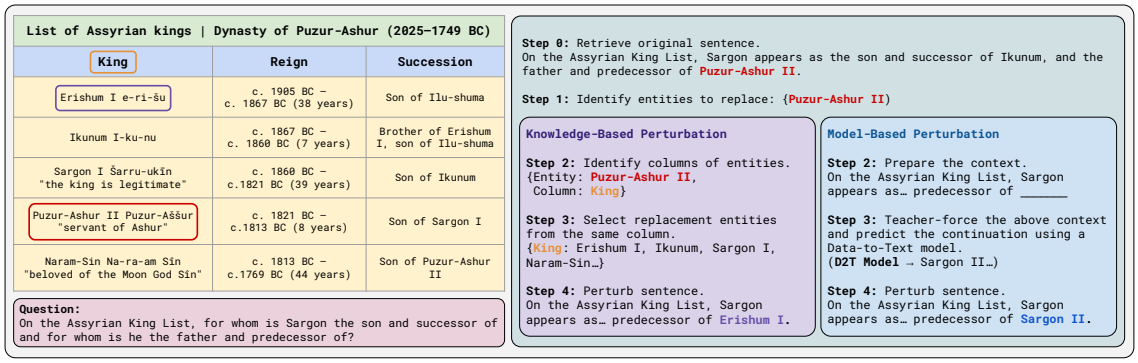


Fig. 6.79: Example from R2D2. (Image from [227]).

6.5 Summary

In this chapter we looked at the generation of natural language text from structured data. This is a very active research area and new relevant papers get published on a weekly basis. We will be following closely the literature on these topics and consider making the appropriate updates to this book.

For more information on text generation, we refer the reader to a few recent surveys on this topic: [41], [92], [102], [112], [114], [149], [158], [231], and [376]. For table pretraining, we refer the reader to two recent surveys [26] and [93].

Facts: (4 of 50 shown)

Nicholas lives in Washington D.C. with Sheryl.
Sheryl is Nicholas's spouse.
Teuvo was born in 1912 in Ruskala.
In 1978, Sheryl's mother gave birth to her in Huntsville.

Queries:

Does Nicholas's spouse live in Washington D.C.?
`(Boolean Join)` ⟶ `TRUE`

Who is Sheryl's husband?
`(Lookup)` ⟶ Nicholas

Who is the oldest person in the database?
`(Max)` ⟶ Teuvo

Who is Sheryl's mother?
`(Lookup)` ⟶ `NULL`

Fig. 6.80: A database encoded in plain text by NeuralDB. (Image from [314]).