Automatic Generation of Electromyogram Diagnosis Report: Task and Dataset

Anonymous ACL submission

Abstract

Report-writing of electromyogram can be problematic for under-experienced physicians and time-consuming for experienced physicians. In this paper, we explore to generate textual report from tabular diagnostic records of electromyogram. We construct the first dataset for this task and demonstrate results of some baseline approaches.

1 Introduction

Electromyography (EMG) refers to the muscle bioelectrical pattern recorded with an electromyograph (Ni et al., 2020). It is one of the major diagnostic tools for identifying and characterizing disorders of the motor unit (Daube, 2002). After the EMG examination, the physicians will get the records of the electrical signals and perform a two-step analysis. Firstly, they analyze the wave form and convert signals to tabular data with pre-defined format. Secondly, they interpret the tabular data to a diagnosis report (Boon et al., 2008).

Figure 1 shows an anonymized EMG diagnostic report. The reports consist of two sections, *Findings* and *Impression*. The *Findings* section lists the key diagnostic results revealed in the tabular data. As for the *Impression* section, it contains an anatomic or physiologic diagnosis but not a final clinical diagnosis. The *Impression* should be brief, yet clear and disclose as much information as possible. Intuitively, *Findings* can be seen as a summary of tabular information, while *Impression* needs to be inferred in conjunction with the physician's clinical experience (Katirji, 2002). In this paper, we focus on the task of automatic report generation from EMG tabular data.

There is already a considerable of work for medical report generation (Jing et al., 2019; Liu et al., 2019b; Zhang et al., 2020b)). However, they mainly focus on x-ray images. Here, we introduce a new dataset which contains anonymized

针极肌电图（Electromyogram, EMG）

Name 被检肌肉	Ins Act 插入电位	Spon Act 自发电位				MUPs		Recruit 募集相
		Fibs 纤颤	PSW 正锐	Fascics 束颤	Others 其他	Polyph 多相	Form 形态	
L Dors.Int.I 左 第一背侧骨间肌	Norm 正常	-	1+	-	-	-	>5mv MUP	干扰相
L Ext.Dig.Com 左 指总伸肌	Norm 正常	-	-	-	-	-	-	单-混相
...	...	...	...	...	...	...	...	...

神经传导测定 (Nerve Conduction Study)

Name 检查神经	Type 项目	Stim 刺激	Rec 记录	Lat 潜伏期	Amp 波幅	Dist 距离	CV 速度
L Median 左 正中神经	Motor 运动	Wrist 腕	APE 拇短展肌	4.5	7.0	53	
R Peroneal 右 腓总神经	F-wave F波	Ankle 踝	EDB 趾短伸肌	50.5			
L Ulnar 尺神经	Sensory 感觉	Dig III 中指	Wrist 腕	3.4	15	120	46.2
...	...	...	...	...	...	...	...

Findings:
EMG：被检肌未见明显肌源性或神经源性损害肌电改变。
NCV：左侧正中神经运动传导远端潜伏期正常上限，感觉神经传导速度轻度减慢；余运动和感觉神经传导速度和波幅正常范围。左侧正中运动神经F波潜伏期略延长。
Needle EMG revealed no denervation or reinnervation in the muscles examined. Nerve conduction studies revealed mildly prolonged motor distal latency and slowed sensory nerve conduction velocity of left Median Nerve.

Impression:
提示：左侧正中神经腕部轻度损害，CTS可考虑
These findings are consistent with a mild carpal tunnel syndrome on the left side.

Figure 1: An example of a EMG diagnostic report, upper region is the tabular information of the electrophysiological examination and lower region is the diagnostic report

tabular result of electrophysiological examination and corresponding diagnostic reports written by physicians and demonstrate a pipeline to generate diagnostic reports from tabular data of EMG examination. This is a the first attempt in this field.

Considering the heterogeneity of the *Findings* and *Impression*, we treat the generation of EMG diagnostic report as two tasks, we generate *Findings* and *Impression* separately from tabular information of the electrophysiological examination. Both tasks are formalized as table-to-text generation tasks. We trained neural-based models on these two tasks and tried to learn physicians' clinical experience in EMG diagnosis from a large number of real diagnostic reports.

2 Dataset and Task Description

2.1 Dataset

In this section, we introduce our new annotated dataset MIME (Medical Information Mart for Elec-

Measurement	Value
# of Samples	2,848
Vocab	549
Avg # of Records	266
Avg Length (*Findings*)	82
Avg Length (*Impression*)	29

Table 1: Dataset Statistics

tromyogram (Denny-Brown, 1949)), which includes anonymized tabular result of electrophysiological examination and corresponding diagnostic reports written by physicians (Wang et al., 2018). In an electrophysiological examination, the patient usually has multiple physical tests, including EMG, NCV, RNS, Blink, LET, SEP, MEP, SET, Inching, etc (Miura et al., 2020). To build this dataset, we kept diagnostic reports that contained only EMG and NCV tests (Judzewitsch et al., 1983) (around 85%), leaving more complex scenarios for future work. The final dataset consists of 2,848 EMG diagnostic reports of patients in Huashan Hospital Affiliated to Fudan University [1] in 2006, 2007, 2010 and 2013, and it's divided into 2278, 285, 285, as train, validation, test set respectively. Each report in our dataset consists of three parts:

- **Patient information** (such as *gender*, *height*, *age*)
- **Pathological examination results** (EMG & NCV test, in tabular form)
- **Diagnostic opinion** (*Findings* & *Impression*, the results of EMG and NCV test are summarized in *Findings* section, the diagnostic results are summarized in the *Impression* section).

To facilitate the evaluation of the generated *Findings* quality of the model. We extract corresponding quintuples (detection, location, project, target, state) for each sentence in the test set, and each quintuple describes the fact of a specific detection item. The first four items can uniquely locate a cell in the table, and the last item corresponds to the description of the unit state. We emphasize that such an evaluation scheme is most appropriate when evaluating generations that are primarily intended to summarize information. While *Impression* needs to be inferred in conjunction with the physician's clinical experience and there is very little overlap between *Impression* and tabular information. Designing evaluation metrics for *Impression* will be more difficult, and we will leave it for future work.

[1] https://www.huashan.org.cn/

Table 1 gives some basic statistics for our MIME dataset. The vocabulary size is 549, which indicates that the lexicon is very limited in our EMG diagnostic report setting. The average number of records in the table is 266, and the average length of *Findings* and *Impression* are 82 and 29, respectively.

2.2 Task Description

In this paper, we treat the generation of EMG diagnostic report as a table-to-text task. We generate both *Findings* and *Impression* from tabular information of the electrophysiological examination using a pre-trained language model (GPT2) (Radford et al., 2019; Zhao et al., 2019), and two other non-pretrained models as baselines. Details of these models are described in the next section.

3 Methods

In this section, we will first introduce some notations, and then we will describe how to generate EMG diagnostic reports using our models by specifying how are our models organized and how is input arranged.

3.1 Notations

Consider the following notations:

- We use $r_1, r_2, ..., r_n$ to denote a table $\mathbf{T}$, and for a regular table, each r represents a cell in the table and have a 2-tuples form that contains column name $r.k$ (key), and cell value $r.v$ (value).
- We use $y_1, y_2, ..., y_{|T|}$ denotes a piece of text Y, and each y is a token or a word.
- Our dataset consists of $(\mathbf{T}, Y)$ pairs and it's worth noting that although we have multiple tables or text segments, we can encode them in exactly the same way, therefore, for convenience, we use the tuple $(\mathbf{T}, Y)$ to represent input and output respectively.

3.2 Compared Models

3.2.1 Long Short Term Memory (LSTM)

Our model follows the standard encoder-decoder architecture (Bahdanau et al., 2014), where the encoder encodes the table into hidden representations and the decoder generates text conditioned on these representations.

The first layer of the network consists in learning two embedding matrices to embed the record keys and values. Each record embedding is computed by a linear projection on their concatenation. We use a bidirectional LSTM (Hochreiter and Schmidhuber,

Reference	Findings	EMG	左下肢部分肌在静息下见纤颤、正尖波，轻收缩部分肌 MUP 见偏宽大电位伴不规则波和多相电位增多，重收缩募集减少。 (Some muscles of the left lower extremity showed fibrillation and positive sharp waves at rest. In the light contraction part of the muscles, MUP showed partial broad potential with increased irregular waves and polyphasic potentials, and reduced recruitment of re-contractions.)
		NCV	左侧腓总神经 CMAP 波幅较对侧降低；余被检感觉和运动神经传导速度和波幅正常范围。左侧胫神经 H 反射未引出。 (The CMAP amplitude of the left common peroneal nerve was lower than that of the contralateral side; the rest of the tested sensory and motor nerve conduction velocity and amplitude were normal. Left tibial nerve H reflex not elicited.)
	Impression		神经源性损害肌电改变，累及左下肢部分肌，左侧 L5 根性损害可考虑。 (Neurogenic damage changes in myoelectricity, involving some muscles of the left lower limb, damage to the left L5 root can be considered.)
Transformer	Findings	EMG	左下肢部分肌见纤颤、正尖波；轻收缩左下肢部分肌见 MUP 偏宽大伴或不伴多相电位和不规则波增多；重收缩募集减少。 (Some muscles of the left lower extremity showed fibrillation and positive sharp waves; some muscles of the left lower extremity were slightly contracted, the MUP was widened with or without polyphasic potentials and irregular waves; the recruitment of re-contractions decreased.)
		NCV	左侧腓总神经运动传导 CMAP 波幅略降低，余神经传导速度和波幅正常范围。 (The CMAP amplitude of the motor conduction of the left common peroneal nerve is slightly reduced, and the conduction velocity and amplitude of the remaining nerve are in the normal range.)
	Impression		神经源性损害肌电改变，右 L5-S1 根性损害可考虑。 (Neurogenic damage changes in myoelectricity, right L5-S1 root damage can be considered.)
LSTM	Findings	EMG	右侧下肢部分肌见纤颤正尖波；轻收缩部分肌见 MUP 偏宽大或见巨大电位；重收缩募集减少。 (Some muscles of the right lower extremity showed positive sharp waves of fibrillation; some muscles with light contraction saw MUP widening or huge potentials; recruitment of heavy contractions was reduced.)
		NCV	右侧正中神经和尺神经运动传导 CMAP 波幅降低；余运动和感觉神经传导速度和波幅正常范围。运动神经 F 波潜期正常范围。 (The motor conduction CMAP amplitude of the right median nerve and ulnar nerve decreased; the remaining motor and sensory nerve conduction velocity and amplitude were normal. Normal range of motor nerve F wave latency.)
	Impression		神经源性损害肌电改变，累及右下肢部分肌。L5 根性损害可考虑。 (Neurogenic injury changes in electromyography, involving some muscles of the right lower limb. L5 root damage can be considered.)
GPT2	Findings	EMG	左下肢部分肌和左侧腓外肌见纤颤正尖波；轻收缩部分肌见 MUP 部分偏宽大伴或不伴多相电位和不规则波增多；重收缩募集略减少。 (Some muscles of the left lower extremity and the left extraperoneal muscles showed positive sharp waves of fibrillation; some muscles of light contraction showed that the MUP part was too wide with or without the increase of polyphasic potentials and irregular waves; the recruitment of heavy contractions was slightly reduced.)
		NCV	左侧腓总神经运动传导 CMAP 波幅降低；余运动和感觉神经传导速度和波幅正常范围。运动神经 F 波潜伏期正常范围或未引出。 (The motor conduction CMAP amplitude of the left common peroneal nerve decreased; the remaining motor and sensory nerve conduction velocity and amplitude were normal. The motor nerve F wave incubation period is normal range or not elicited.)
	Impression		1.左侧腓总神经损害，腓骨头-腓骨头上 1cm 处明显。2.慢性神经源性损害肌电改变，累及左下肢部分肌，L-4 根性损害可能。 (1. The damage to the left common peroneal nerve is obvious at 1cm above the head of the fibula. 2. Chronic neurogenic damage, electromyographic changes, involving part of the left lower limb muscles, L-4 root damage may be.)

Figure 2: Example of generation of *Findings* and *Impression* with gold reference

1997) on top of the cell embedding to obtain the table representation. After the table is represented as a sequence of vectors, a decoder based on LSTM (Hochreiter and Schmidhuber, 1997) is applied to generate text token by token.

3.2.2 Transformer

We linearize the table and feed the records into standard Transformer (Vaswani et al., 2017). The linearization of the table consists of a concatenation of row cells. And since each cell (i.e. record) is represented by the key and value, We concatenate them together and get the representation of the cell using a layer of MLP, which is same as the record embedding layer described above.

3.2.3 GPT2

We follow previous work on linearizing knowledge base as natural language (Liu et al., 2019a; Zhang et al., 2020a) to propose “table linearization”, which uses template to flatten the table $\mathbf{T}$ as a document $P_T = w_1, \cdots, w_{|T|}$ fed into pre-trained language models to generate statement Y , where we use w_i to denote the i-th word in the generated paragraph P_T and $|T|$ to denote the length of the paragraph (the word w_i is either a table entry or a functional word in the template). The original table $\mathbf{T}$ is transformed into a paragraph by horizontally scanning each cell in the table.

After table linearization, we directly feed the paragraph P_T as the input to the pre-trained GPT-2 model and generate the output sentence Y. We finetune the model on MIME by maximizing the likelihood of $p(Y|P_T;\beta)$, with β denoting the parameters of GPT-2 model (Radford et al., 2019; Zhao et al., 2019).

3.3 Text Generation

For the generation of *Findings* or *Impression* based on the table, we both use the three above-mentioned table-to-text models, the only difference is the out text.

4 Experiment & Result

We base our implementation on Huggingface’s Transformer (Wolf et al., 2019) for GPT-2 (Radford et al., 2019; Zhao et al., 2019) with word vocabulary of 20K. The batch size is 2. The model is finetuned using Adam optimizer (Kingma and Ba, 2017) with a learning rate of 1e-6.

Model	B-1	B-2	B-4	R-1	R-2	R-L	TC	TM	CS-acc
LSTM	58.9	54.9	48.6	80.0	66.2	76.0	35.0	29.0	60.4
Transformer	72.0	68.4	62.2	85.4	74.2	81.9	42.4	34.5	72.3
GPT2	**76.4**	**73.8**	**69.5**	**88.3**	**80.0**	**86.1**	**52.1**	**42.5**	**88.4**

Table 2: Overall performance of different models for *Findings* generation. The best result is marked in bold. The Prediction Accuracy of Cell State (CS-acc) represents the accuracy of the fifth state prediction for those accurate 4-tuples extracted by the model.

Model	B-1	B-2	B-4	R-1	R-2	R-L
LSTM	50.1	45.5	36.9	62.8	49.5	61.4
Transformer	53.0	48.6	39.4	65.5	53.1	64.4
GPT2	**59.4**	**55.9**	**48.6**	**70.6**	**60.2**	**69.7**

Table 3: Overall performance of different models for *Impression* generation.

4.1 Result and Analysis

We use **ROUGE** (Lin, 2004) and **BLEU** (Papineni et al., 2002) scores to evaluate our model. And we report BLEU-1, BLEU-2, BLEU-4 scores and the F_1 scores for unigram (ROUGE-1) and bigram (ROUGE-2) and longest common subsequence overlap (ROUGE-L).

We also propose two information retrieval (IR) based metrics. These metrics compare the gold and generated descriptions and measure to what extent the extracted facts are aligned or differ.First, we apply an information extraction (IE) system to extract quintuple in $Findings$. The value ranges of the first four items in the quintuple can be obtained directly from the tables of the training set. The last item is obtained from our manually labeled test set(only 12). For example, in the sentence *Tibial nerve H reflex latency upper limit of normal.*, an IE tool will extract the pair (Tibial nerve, -, H reflex, latency, normal). Second, we compute two metrics on the extracted information:

• **Tuple Coverage (TC)** estimates how well the generated description containing the gold description in terms of mentioned quintuple. Obviously, based on this simple entity extraction IE system, each item in the 5-tuple may contain multiple elements at the same time. When only the extracted quintuple contains the truly labeled quintuple, we call it tuple coverage. For example, quintuple (ulnar nerve/tibial nerve, -, H reflex, latency, normal) covers quintuple (tibial nerve, -, H reflex, latency, normal).

• **Tuple Matching (TM)** measures how well the system is able to generate text containing factual (i.e., correct) facts. If and only if the two tuples are exactly the same, we call it a match.

While **ROUGE** and **BLEU** is perhaps a reasonably effective way of evaluating text generation, we note that it primarily rewards fluent text generation, rather than generations that capture the most important information in the database which is extremely important for medical diagnosis. Our proposed IE system can be used as an approximation to solve this evaluation challenge. The result for *Findings* generation is in the Table 2 and the result for *Impression* generation is in the Table 3.

As is shown in the table, all models get relatively good textual overlap with reference text. And the pre-trained model achieves the best results on all metrics benefit from the rich language information contained in it. The extractive metrics provide further insight into the behavior of the models. We first note that on the gold documents $y_{1:|T|}$, the extractive model reaches 70.5 coverage and 50.9 match rate. Using the LSTM model, generation only has a tuple coverage (TC) of 35.0 indicating that 4-tuples are often generated incorrectly. The best pre-trained model improves this value to 52.1, a significant improvement and potentially the cause of the improved **ROUGE** and **BLEU** score, but still far below gold. It is worth noting that all the models seem to get a relatively high prediction accuracy for the fifth item on the accurately matched quadruples. This shows that in the $Findings$ generation task, it is more difficult to locate a specific position in the table than to describe its state after finding the precise location.

5 Conclusions

This paper explores the automatic generation of electromyogram diagnostic report. We formalize the generation as two tasks, namely, table-to-findings and findings-to-impressions. To evaluate the generation results, we introduce both token-level and fact-level evaluations. Results of some baselines on our self-constructed dataset are demonstrated.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Andrea J Boon, Kais I Alsharif, C Michel Harper, and Jay Smith. 2008. Ultrasound-guided needle emg of the diaphragm: technique description and case report. *Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine*, 38(6):1623–1626.

Jasper R Daube. 2002. Assessing the motor unit with needle electromyography. *CONTEMPORARY NEUROLOGY SERIES*, 66:293–323.

D Denny-Brown. 1949. Interpretation of the electromyogram. *Archives of Neurology & Psychiatry*, 61(2):99–128.

Heng Gong, Xiaocheng Feng, Bing Qin, and Ting Liu. 2019. Table-to-text generation with effective hierarchical encoder on three dimensions (row, column and time). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3143–3152, Hong Kong, China. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Parag Jain, Anirban Laha, Karthik Sankaranarayanan, Preksha Nema, Mitesh M. Khapra, and Shreyas Shetty. 2018. A mixed hierarchical attention based encoder-decoder approach for standard table summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 622–627, New Orleans, Louisiana. Association for Computational Linguistics.

Baoyu Jing, Zeya Wang, and Eric Xing. 2019. Show, describe and conclude: On exploiting the structure information of chest X-ray reports. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6570–6580, Florence, Italy. Association for Computational Linguistics.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia. Association for Computational Linguistics.

Roman G Judzewitsch, Jonathan B Jaspan, Kenneth S Polonsky, Clarice R Weinberg, Jeffrey B Halter, Eugen Halar, Michael A Pfeifer, Cynthia Vukadinovic, Lawrence Bernstein, Michael Schneider, et al. 1983. Aldose reductase inhibition improves nerve conduction velocity in diabetic patients. *New England Journal of Medicine*, 308(3):119–125.

Bashar Katirji. 2002. The clinical electromyography examination: An overview. *Neurologic clinics*, 20(2):291–303.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Angli Liu, Jingfei Du, and Veselin Stoyanov. 2019a. Knowledge-augmented language model and its application to unsupervised named-entity recognition. *arXiv preprint arXiv:1904.04458*.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019b. Clinically accurate chest x-ray report generation. *arXiv preprint arXiv:1904.02633*.

Yasuhide Miura, Yuhao Zhang, Curtis P Langlotz, and Dan Jurafsky. 2020. Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*.

Jianmo Ni, Chun-Nan Hsu, Amilcare Gentili, and Julian McAuley. 2020. Learning visual-semantic embeddings for reporting abnormal findings on chest x-rays. *arXiv preprint arXiv:2010.02467*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019a. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6908–6915.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019b. Data-to-text generation with entity modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2023–2035.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Jiaoyan Chen, Wei Zhang, and Huajun Chen. 2020a. Relation adversarial network for low resource knowledge graph completion. In *Proceedings of The Web Conference 2020*, pages 1–12.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020b. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online. Association for Computational Linguistics.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.

A Related Work

Data-to-text generation Wiseman, Shieber, and Rush (Wiseman et al., 2017) introduced a document-scale data-to-text dataset with relatively large table records and long reference texts and proposed extraction based evaluation metrics for automatically evaluating generation quality. More specifically, they introduced an information extraction module to evaluate content generation, and ordering of the data-to-document model. Puduppully, Dong and Lapata (Puduppully et al., 2019a) model a content-selection and-planning module separate from text generation, with the idea that introducing a direct signal, i.e. a loss on orderly selection of table records would improve generation performance. Gong, Feng, Qin, Bing and Liu. (Gong et al., 2019) presented a hierarchical encoder that learn records' representation along row and column and obtain row-level representation for subsequent decoding. Jain et al. (Jain et al., 2018) proposed a mixed hierarchical attention based encoder-decoder model to leverage the structural information in tables. Puduppully, Dong and Latapa (Puduppully et al., 2019b) propose an entity-centric architecture such that instead of treating entities as ordinary tokens, they create dynamically updated entity-specific representations and generates text using hierarchical attention on table and entity memory cell.

Automatic Medical Report generation Jing, Xie and Xing (Jing et al., 2018) proposed a co-attention mechanism to localize regions containing abnormalities and generate descriptive texts for them. Jing, Wang and Xing (Jing et al., 2019) proposed a multi-agent framework to exploit the structural features within report sections for generating Chest X-ray Reports where they have two agents for generating text about abnormal and normal results separately with the observation that the distribution between abnormality and normality is imbalanced and the wordings are quite different in text describing abnormal and normal results. Liu et al. (Liu et al., 2019b) proposed a generation model which hierarchically first chooses topics and then generates words from topics and they optimized the model for clinical correctness which a proposed clinically coherent reward via reinforment learning. Zhang, Merck, Tsai, Manning and Langlotz (Zhang et al., 2020b) leveraged an existing information extraction module to extract a zero-one vector of 14 dimension indicating the presence or absence of 14 clinical observations in chest radiology reports and apply reinforcement learning with a factual correctness reward to improve the factuality of generated reports.