Content Type Profiling of Data-to-Text Generation Datasets

Ashish Upadhyay and **Stewart Massie**
School of Computing
Robert Gordon University
Aberdeen, UK
{a.upadhyay,s.massie}@rgu.ac.uk

Abstract

Data-to-Text Generation (D2T) problems can be considered as a stream of time-stamped events with a text summary being produced for each. The problem becomes more challenging when event summaries contain complex insights derived from multiple records either within an event, or across several events from the event stream. It is important to understand the different types of content present in the summary to help us better define the system requirements so that we can build better systems. In this paper, we propose a novel typology of content types, that we use to classify the contents of event summaries. Using the typology, a profile of a dataset is generated as the distribution of the aggregated content types which captures the specific characteristics of the dataset and gives a measure of the complexity present in the problem. Through extensive experiments on different D2T datasets we demonstrate that neural generative systems specifically struggle to generate contents of complex types, highlighting the need for improved D2T techniques.

1 Introduction

An ecologically valid task requires the automated systems to resemble the real-world scenario as closely as possible in its output (de Vries et al., 2020). Accordingly, a Data-to-Text Generation (D2T) system needs to convey important insights extracted from the data in the textual summaries (Reiter, 2007; Gatt and Krahmer, 2018). Most D2T problems can be seen as a stream of time-stamped events with a textual summary of each event presenting the insights. An event is the time-period of interest for which the textual summary is written. For example, in sports reporting - a game played between two teams can be an event; whereas in weather forecasting - the time period and location for which the forecast is written can be considered one full event. The summaries can contain different types of facts with information sometimes coming from multiple events in the stream. In most cases, the facts in an event summary are verbatim of input records. Other times, these facts are derived from multiple records of either the same or multiple events in the stream. As an example, we show a part of baseball summary with multiple types of content in Figure 1.

The Bucks (10-10) *handled* **the Heat (9-9) 109-85** on **Friday night in Milwaukee**. ***It was the second victory over Miami for the Bucks this season after emerging victorious in Miami 91-84 on Nov. 16***. *Milwaukee fell behind early but clawed back into the game in the second quarter and held a four-point advantage at half. The Bucks were led by an unlikely face in Kendall Marshall*, who scored a ***season-high*** **20 points (7-8 FG, 4-5 3Pt) in 24 minutes**.

Figure 1: Part of a basketball summary showing different types of content. Information in "**bold**" such as Bucks' points in the game (109) can be directly copied from input data, while the ones in "*italics*" needs to be derived from multiple records such as the fact - Kendall Marshall leading the Bucks team. Finally, the information in "***bold & italics***" such as 'this was Miami's second victory against Bucks' are derived using records from multiple events in the stream.

A distribution of different content types in the summaries of a dataset can be used to generate its profile that can capture the specific characteristics of the dataset and provide a measure of complexity present in the problem. The dataset profile can help us better define the D2T system requirements, such as: the type of system to build - is a complex domain-specific system required or can a general system be effective; or any information gap (Thomson et al., 2020b) that needs to be bridged at the data level. Generally a problem's complexity is identified with the evaluation of systems built for

Proceedings of the 29th International Conference on Computational Linguistics, pages 5770–5782
October 12–17, 2022.

the task. There are several methods to evaluate the D2T systems, mostly by measuring the factual accuracy of generated texts (Thomson and Reiter, 2020; Wiseman et al., 2017; Garneau and Lamontagne, 2021; Kasner et al., 2021) or lexical similarity of generated texts with reference texts (Papineni et al., 2002; Lin, 2004; Zhang et al., 2020b). These, whilst being promising are reactive measures where a full cycle of system development is needed to evaluate both the dataset and D2T system together. Whereas dataset profiling can be a proactive measure to gain important insights about the task before even starting system development. There are other utilities of dataset profiling method as well, most notably, it can be used as a measure of dataset's complexity in datasheets for dataset (Gebru et al., 2021).

There are other type of datasets in D2T as well such as E2E (Dušek et al., 2020) or WebNLG (Colin et al., 2016) that do not follow the event based time-series setting. These datasets mostly focus on improving general ability of generation models such as transcribing a set of records, possibly in a domain-agonist setting. In this work, we do not address such datasets and rather focus on those which follow a time-series structure, where summaries may contain facts derived from across several event records, and also may require content selection on input data. In this paper, we propose a typology of different content types in D2T summaries based on the source of their information in the event stream. Our key contributions are as follows [1]:

- we propose a novel typology of content types in D2T summaries;
- we use the proposed typology to profile datasets and understand their characteristics;
- we demonstrate the challenge facing generation systems in producing complex contents.

The rest of the paper is organised as follows: in Section 2 we formally define the D2T task and discuss our proposed content type typology. We then go on to describe our experimental set up in Section 3, and discuss the results in Section 4. Some related works are discussed in Section 5, before concluding the paper in Section 6.

[1] code, data, and results are at https://github.com/ashishu007/Content-Type-Profiling

2 Methodology

The idea proposed in this paper is that we can create a profile capturing the specific characteristics of a dataset by looking at the distribution of content types present in its summaries. The first step towards achieving this is to formally define the main concepts (Section 2.1). Then, in second step, the content typology is defined by identifying the different types of facts that can be included in an event summary (Section 2.2).

2.1 Formalisation of D2T Generation Task

We start with formalising the concepts in a time-stamped D2T dataset. A data instance in such **D2T dataset** ($\mathcal{DB}$) is an **event** ($\mathcal{E}_i$) with a **data structure** ($\mathcal{D}_i$) for which a **textual summary** ($\mathcal{S}_i$) is written summarising the insights and information of the event. A data structure consists of multiple **entities** ($\mathcal{O}$) that are the objects involved in the event, and each entity is described by multiple **features** ($\mathcal{F}$) which are the attributes of those entities; such that:

$$\mathcal{DB} = [\mathcal{E}_{i-e}, \cdots, \mathcal{E}_i, \cdots, \mathcal{E}_{i+e}]$$
$$\mathcal{E}_i = \{\mathcal{D}_i, \mathcal{S}_i\}$$
$$\mathcal{D}_i = \{\mathcal{O}_1, \mathcal{O}_2, \cdots, \mathcal{O}_o\}$$
$$\mathcal{O}_o = \{\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_f\}$$

When building a text generation system g for a D2T task, the summary of an event is the function of current event as well as other events in stream:

$$\mathcal{S}_i = g(\mathcal{D}_i, \mathcal{DB})$$

A value $\mathcal{R}$ will be recorded for each feature $\mathcal{F}$ of an entity $\mathcal{O}$ which is considered a record in the data structure. So, an input data structure flattened into a sequence of records (mostly for training a neural generation system) will be:

$$\mathcal{D}_i = \{(\mathcal{R}_{1,1}, \mathcal{R}_{1,2}, \cdots, \mathcal{R}_{1,f}), \cdots (\mathcal{R}_{2,1}, \mathcal{R}_{2,2}, \cdots, \mathcal{R}_{2,f}), \cdots (\mathcal{R}_{o,1}, \mathcal{R}_{o,2}, \cdots, \mathcal{R}_{o,f})\}$$

2.2 Content Type Typology

The textual summary in a D2T dataset may contain facts derived from different sources. To begin with, a fact can be derived from either the same event or from the records of different events. For example, a basketball game summary generally mentions different stats scored by players in the game. Such

information is explicitly present in the input data of the game and can be directly copied to the output summary. Most times, summaries also mentions the average stats recorded by a player in past few games. To derive such facts, the generation system needs to consider the records from previous games as well. So based on the event source, a fact can be categorised as either **intra-event** (derived from the current event's records) or **inter-event** (derived from across-event records).

The intra-event category can be further granulated to identify the difficulty of generating a fact within the same event. Again, taking an example from a basketball summary, among many things, the summary could either mention some specific stat (points or rebounds) scored by a player in the game, or mention if the player has scored a double-double [2]. The information of specific stat of players is explicitly present in the input data, which can be directly copied to the output summary. While the information that the player scored a double-double is not explicitly present in the input data, which needs to be derived from several records of the player. So, within intra-event, there can be two different types of facts: **basic**, that can be just copied directly from the input data; and **complex**, that needs to be derived from the multiple records of the same event.

Considering the following notations: each summary $\mathcal{S}_i$ is a combination of multiple **sentences** $\mathcal{T}$, which will contain at-least one or more **fact** $\mathcal{L}$.

$$\mathcal{S}_i = \{\mathcal{T}_1, \mathcal{T}_2, \cdots, \mathcal{T}_j\}$$
$$\mathcal{T}_j = \{\mathcal{L}_1, \mathcal{L}_2, \cdots, \mathcal{L}_k\}$$

Thus, a Content Type typology of three classes is proposed based on the types of facts an event summary can contain:

Intra-Event Basic ($\mathcal{B}$): a fact that is copied from the input record set of the same event.

$$\mathcal{B} \iff \mathcal{L}_k = \mathcal{R}_{i,o,f}$$

Intra-Event Complex ($\mathcal{C}$): a fact that is derived from multiple records of the same event.

$$\mathcal{C} \iff \mathcal{L}_k = \mathcal{R}_{i,o,f} \otimes \mathcal{R}_{i,o\pm l,f\pm m} \otimes \cdots$$

Inter-Event ($\mathcal{A}$): a fact that is either copied or derived from the records of multiple events.

$$\mathcal{A} \iff \mathcal{L}_k = \mathcal{R}_{i,r} \otimes \mathcal{R}_{i-n,o\pm l,f\pm m} \otimes \cdots$$

[2] https://en.wikipedia.org/wiki/Double-double

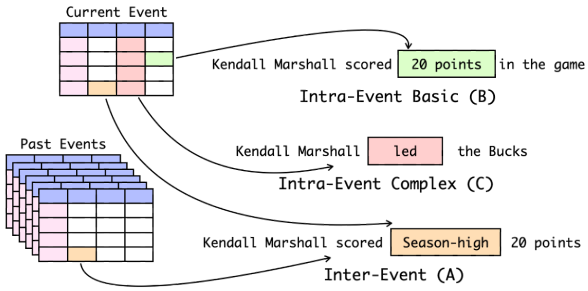

Figure 2: Different content types in a D2T summary. The Intra-Event Basic content is taken from only one record, while the Intra-Event Complex content is derived from multiple records of the same event. Finally, the Inter-Event content is derived from the records of multiple events.

where, i, j, & k denote an event, an object, and a feature respectively. l, m, n are positive integers and $\otimes$ is an operation that requires inference between more than one records. Thus, a sentence $\mathcal{T}$ can be assigned into one or more content type if it contains at-least one fact of that type.

Taking the example from Figure 1, the last sentence in the summary: "The *Bucks* were *led* by an unlikely face in *Kendall Marshall*, who scored a *season-high 20 points* (*7-8 FG, 4-5 3Pt*) in *24 minutes*" has multiple facts. To calculate the fact that *Kendall Marshall led the Bucks*, the generation system should be able to analyse the records of all players in Bucks team, which is **Intra-Event Complex** ($\mathcal{C}$) type of fact. Then the fact that he scored *season-high 20 points*, will only be calculated by analysing the records of all games in which Kendall Marshall played, which is **Inter-Event** ($\mathcal{A}$) type of fact. And finally, the shot-breakdown (*7-8 FG, 4-5 3Pt*) and *number of minutes* he played are the facts that can be directly copied from the input, which are **Intra-Event Basic** ($\mathcal{B}$) type of facts. Thus this sentence will be classified into all three content types ($\mathcal{B}$, $\mathcal{C}$, $\mathcal{A}$). Whereas the second sentence (***It was the second victory ...***) will be classified as inter-event type ($\mathcal{A}$) as it only contains information from multiple games (also see Figure 2).

It is also possible to extend the Inter-Event category into Basic and Complex categories. But, it is left for the future work mainly because of two reasons: first, both inter-event basic and inter-event complex will pose similar challenge to a generation system, which is, having access to data from other events during run-time; and second, the occurrence of inter-event basic facts will be rare, as often records from multiple events is used to derive

a new fact rather than being used as a single fact.

2.3 Building Content Type Classifier

With the Content Type typology defined for a D2T task, the next step is to use this typology in generating the dataset profile. However, the datasets can be large with sometimes more than 20k event summaries, each containing 15-20 sentences. Thus manually annotating these summaries to generate the dataset profile will be difficult. In this work, a **Multi-Label Classifier** is used for generating the dataset profile by classifying the sentences of event summaries into their content types. More specifically, a multi-label classifier function f is learned to map a sentence $\mathcal{T}$ to its content types y as:

$$y_j = f(\mathcal{T}_j)$$

where $y_j \subseteq \mathcal{Y}$, and $\mathcal{Y} = \{\mathcal{B}, \mathcal{C}, \mathcal{A}\}$

More detail on building the content type classifier is given in Appendix A. A pictorial representation of the process of creating a dataset's profile is shown in Figure 3.

3 Experimental Setup

The experiments are performed in three phases. In first phase, the aim is to understand the characteristics of human authored summaries, for which, dataset profiles are generated based on human authored summaries using the proposed content type typology (Section 4.1). The second phase aims to demonstrate the challenge state-of-the-art generation systems face in attempting to generate complex (inter-event and intra-event complex) content (Section 4.2). This is evaluated by comparing the errors made by the generation systems for each content types. Finally, the proposed methodology is used to understand the concept-drift issue in a problem domain which can help in building better systems capable of handling such domain-specific issues (Section 4.3).

3.1 Datasets

Four datasets from different domains are used for profiling: **MLB** (Puduppully et al., 2019b) and **SportSett** (Thomson et al., 2020a) datasets from sports domain; **SumTime** (Sripada et al., 2003) dataset from weather forecasting; and **Obituary** (Upadhyay et al., 2020) from Obituary generation domain. These datasets contain events' structured data on the input side parallelly aligned with human-authored textual summaries of each event.

- **MLB** dataset contains stats from MLB games aligned with their summaries written by human authors. Each sample in the dataset contains the box score and play-by-play record of the of the game on the input side, which is aligned with a textual summary of around 20 sentences long. In this dataset, games are considered as the events; players, teams and the plays as entities of the event; and different stat-types as the features of an entity. The dataset is split for train/valid/test sets, containing 22821/1739/1744 samples each.

- **SportSett** dataset contains box- & line- scores from NBA (basketball) games aligned with human-written summaries describing the respective game. The games are from season 2014 to 2018, out of which 2014, 2015 & 2016 are used as train split, 2017 for valid split, and 2018 for test split. The number of samples in train/valid/test sets is 4745/1228/1229 respectively. Similar to MLB, each game in the dataset is an event; the players teams from the game are entities; and stat-types are the features.

- **SumTime** dataset contains human written weather forecasts for a day written for oil and gas offshore engineers in Aberdeen, UK. The forecasts are usually written from two types of NWP data: wave data; and mmo data (please refer to Sripada et al. (2003) for a detailed discussion on the data organisation). The representation of SumTime on different dimensions is as follows: each sample in the dataset covers a forecast during 12-hours time-period (AM forecast or PM forecast). The time period for which the forecast is written is considered an event; the different entities are the elements described in the forecast such as wind or wave; and the hours of the day for which the readings are taken for those elements are the features. The total number of samples in train/valid/test sets are 793/99/100.

- **Obituary** dataset contains a sample of 850 obituaries aligned with their personal information. In this dataset, an Obituary (or a death) is the event; where the deceased person is

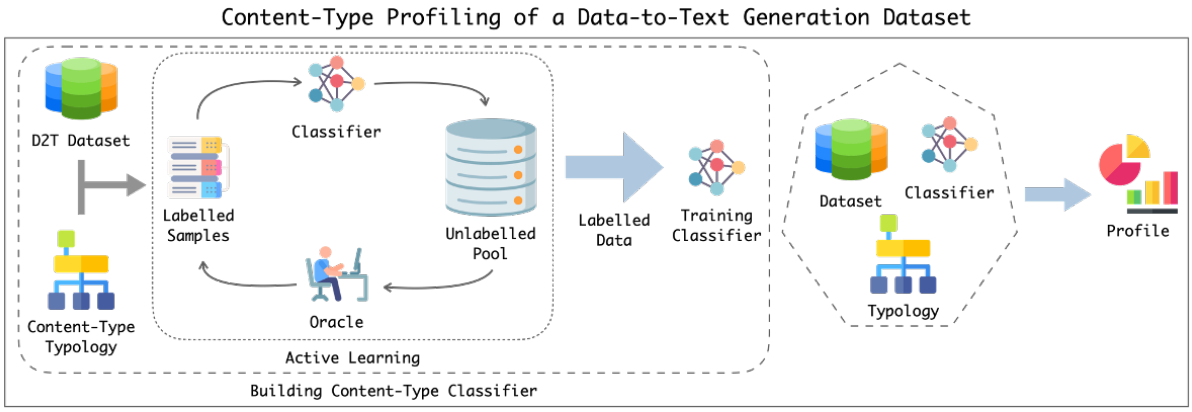

Figure 3: Generating the Content Type Profile of a dataset

the entity; and information related to their personal life and funeral are the different features. The dataset contains 800/20/20 obituaries in train/valid/test set respectively.

3.2 Generation Systems

Phase two of the experiments analyses the ability of state-of-the-art generation systems in producing different types of content. For this, several state-of-the-art generation systems are used to produce summaries on the held-out test-set of datasets mentioned above. For MLB and SportSett, two benchmark neural systems from literature are used: first, the macro-planning model (**Plan**) from Puduppully and Lapata (2021); and second, the entity-based model (**Ent**) from Puduppully et al. (2019b) are used on both datasets. In addition, the hierarchical transformer model (**Hir**) from Rebuffel et al. (2020) is used as a third model for SportSett. For Obituary and SumTime, we are not aware of any existing neural benchmarks, therefore we develop our own generation systems by fine-tuing T5-base (**T5**) from Raffel et al. (2020), BART-base (**BART**) from Lewis et al. (2020), and Pegasus (**Peg**) from Zhang et al. (2020a) on each dataset respectively. Different metric evaluation scores of these developed systems are shown in Appendix B.

3.3 Generation Systems' Accuracy Evaluation

The generation systems' ability of generating complex content is evaluated by measuring the accuracy error-rate of generation systems within each content type category. We calculate the **error rate** of generation systems by manually annotating 10 randomly selected summaries from each system following a pre-established gold standard annotation scheme of D2T systems evaluation developed by Thomson and Reiter (2020). Within each summary, all the generated claims (whether correct or incorrect) within each category are identified and then the error rate is calculated as the ratio of *total incorrect claims to total claims generated*. The error-rate ratio of each content type category is calculated separately. The length of summaries vary from 15-20 sentences each summary in MLB and SportSett to 4-5 sentences in SumTime and Obituary. The annotations are done by the authors themselves and are available on the GitHub repository. The distribution of sentences across different content types from all the evaluated system generated summaries is shown in Table 1.

4 Results and Discussions

In this section, we use the Content Type classifiers build for different datasets using method described in Section 2.2 for generating their profiles. The performance of best classifier for each dataset with its Macro-F1 score and the number of samples used for training is shown in Appendix A. As discussed in the previous section, the dataset profiles will be used for: first, analysing the human-authored summaries from different datasets (Section 4.1); second, analysing the system generated summaries from several state-of-the-art neural generative systems (Section 4.2); and third, characterising the concept-drift issue in SportSett dataset (Section 4.3).

4.1 Analysing Human Authored Summaries

The content type distribution found in human authored summaries from different datasets is shown in Figure 4. On the x-axis, the different content type categories are shown, while the y-axis displays the percentage of sentences belonging to that cate-

Dataset	MLB		SportSett			SumTime			Obituary		
System	Ent	Plan	Ent	Plan	Hir	T5	BART	Peg	T5	BART	Peg
Intra-Event Basic ($\mathcal{B}$)	83	94	71	84	82	20	20	20	35	30	33
Intra-Event Complex ($\mathcal{C}$)	119	193	71	53	84	10	10	10	11	10	10
Inter-Event ($\mathcal{A}$)	55	47	44	33	46	0	0	0	0	0	0

Table 1: Number of sentences from different categories manually annotated for error-rate evaluation

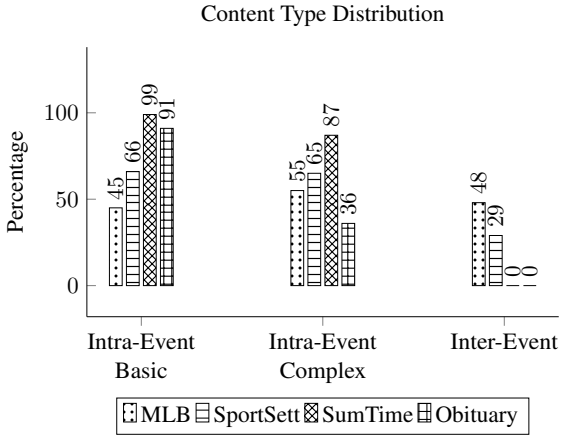

Figure 4: Profile of various datasets based on their human authored summaries

gory. It is noted here that one sentence can be assigned more than one category since a sentence can contain multiple facts of different categories. We can see that MLB has the highest amount of inter-event sentences (48%) with 55% intra-event complex, and 45% intra-event basic sentences. SportSett has 29% inter-event sentences with 65% intra-event complex and 63% intra-event basic sentences. In SumTime, although there are no inter-event sentences, 87% sentences are intra-event complex with 99% of them also being intra-event basic. Obituary has 91% intra-event basic sentences as well as the least percentage of intra-event complex sentences (36%). Obituary also doesn't have any inter-event type sentence in the summaries.

These numbers suggest that humans written summaries do not contain just information copied from input data. Rather, they are full of complex insights derived from multiple records, and possibly multiple events in the stream. This demonstrates that while designing a D2T system, two requirements are necessary: first, in most D2T tasks, an event cannot be considered independent, as it's solution might depend on the data from multiple events in the stream; and second, a system developed for D2T task should be capable of performing complex analytical operations in order to derive implicit information from the given records. Ignoring these requirements will lead to building systems capable of generating only the easier contents and missing the interesting complex insights. Similar issues are observed in other language generation tasks as well where systems try to generate less complex content in order to be safe (Feng et al., 2021; Du and Black, 2019).

4.2 Analysing System Generated Summaries

After analysing the dataset profiles generated using human-authored summaries, we investigate if current state-of-the-art generation systems can produce content with similar profile of human-authored summaries. We show the content type distribution in the system generated along with human reference test-set summaries from different datasets in Figure 5. It can be observed that systems on SumTime and Obituary are able to generate similar amount of intra-event basic & complex sentences as in human written summaries, however, with inaccuracies (will be discussed later in this section). These two datasets do not have any inter-event content in human reference summaries, and thus no such content in system generations as well. MLB and SportSett generated texts have different content type distribution compared to their human reference summaries. If we look at the the generations of Plan system in both datasets, it has the lowest inter-event sentences in MLB and lowest inter-event & intra-event complex sentences in SportSett. This can be attributed to the macro-planning design of the system which restricts the system for producing only information explicitly available in input data. Another pattern can be observed in SportSett, where the system generated summaries have relatively more inter-event sentences than human written summaries, which we explore in the next section.

We also show the error-rate across categories of different systems from different datasets (along with mean and standard deviation of error-rates across systems) in Table 2. The error-rates demon-

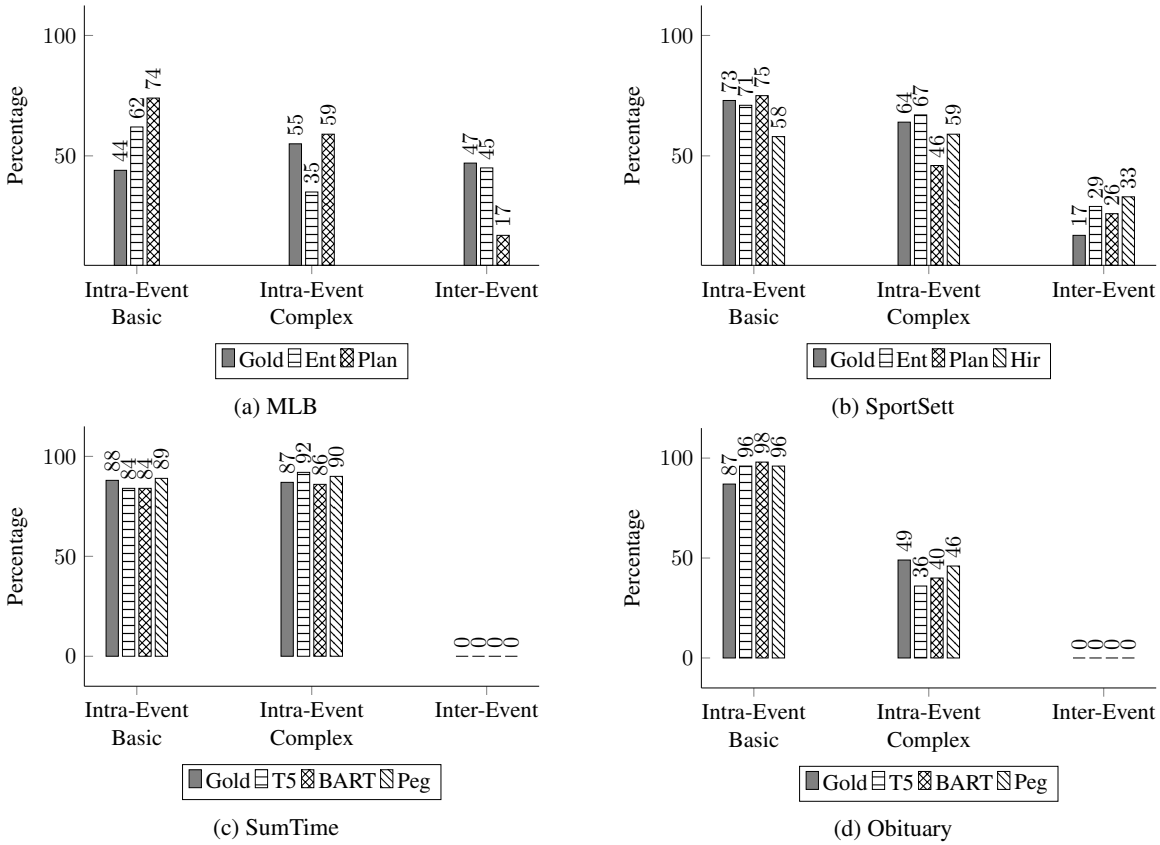

Figure 5: Profile of various datasets based on summaries generated from different generation systems

strate that generation systems struggle to produce the contents of complex types. Generation systems across all datasets have low error-rates in intra-event basic category while higher error-rates in intra-event complex and highest error-rates in the inter-event categories.

Mean error-rate in **intra-event basic** category ranges from 5.5% in Obituary, around 12.5% in MLB and SportSett, to over 38% in SumTime. SumTime has the higest error in this category, which may be due to the lack of training data and a highly domain-specific problem that requires identification of multiple relationships to generate summaries. The systems used for SumTime are not custom designed for the task as with MLB and SportSett, and employs general NLG models, thus having higher error-rate in SumTime than other datasets. Overall, error-rate across all datasets in intra-event basic category is lower than intra-event complex or inter-event (where present) categories.

In the case of **intra-event complex**, almost all the datasets have around 40-50% mean error-rate. This shows systems struggle to learn the domain specific relationships required to derive complex information from the supplied data. Only **Ent** system in MLB and **Plan** system in SportSett have notably lower error-rate. However, these systems are generating comparatively lesser intra-event complex content compared to other systems in order to improve the accuracy (reducing error-rate) by producing less complex content which is easier to generate. All the **inter-event** facts in MLB are incorrect (100% error-rate) while the error-rate in SportSett for inter-event category is also very high. This is not surprising as the input to these systems doesn't take data from across the event stream into account during run-time. SportSett has actually produced some accurate inter-event facts by producing standard phrases learnt from the training data (e.g. "team X has won four out of last five games") that turns out to be correct sometimes. These results clearly demonstrate the difficulty generation systems have in producing content of complex types. Similar observations have been made in literature as well showing the struggle of current state-of-the-art generation systems in producing content that needs to be derived from multiple records or sources (Thomson and Reiter, 2021; Thomson et al., 2020b). The main difference between these works and our proposed content type profiling approach is that in

Systems	Intra-Event Basic ($\mathcal{B}$)	Intra-Event Complex ($\mathcal{C}$)	Inter-Event ($\mathcal{A}$)
		MLB	
Ent	13.98	38.6	100
Plan	10.84	45.27	100
Total	12.4 $\pm$ 1.5	41.9 $\pm$ 3.3	100 $\pm$ 0
		SportSett	
Ent	13.64	61.29	86.96
Plan	6.72	27.54	77.66
Hir	16.72	51.79	91.59
Total	12.3 $\pm$ 4.1	46.8 $\pm$ 14.2	85.4 $\pm$ 5.7
		SumTime	
T5	38.1	50.57	-
BART	39.19	51.85	-
Peg	35.95	48.45	-
Total	37.7 $\pm$ 1.3	50.2 $\pm$ 1.4	-
		Obituary	
T5	2.74	41.67	-
BART	4.86	52.27	-
Peg	9.09	48.39	-
Total	5.5 $\pm$ 2.6	47.4 $\pm$ 4.3	-

Table 2: System-wise error-rates of generation systems developed for various datasets categorised by content types (lower is better; $\downarrow$)

previous works, the insights are drawn by evaluating the summaries generated from systems whereas with our method many such insights can be drawn proactively before going into system development.

4.3 Concept Drift in SportSett

We further apply our proposed content type profiling methodology to capture the concept drift issue in SportSett dataset. This dataset contains NBA games from season 2014 to 2018 and follows a seasonal partition to generate the train/test/valid splits. Seasons 2014, 2015 and 2016 are used for train set while 2017 and 2018 are used for validation and test sets respectively (please refer to Thomson et al. (2020a) for more details). In Figure 6, the content type distribution of summaries by year is shown. The summaries from earlier years contain greater amount of inter-event sentences while comparatively little in later years. Even with intra-event sentences, there are lesser intra-event basic sentences in summaries from earlier years than later ones, indicating that summaries in the training set are more complex than in the validation and test sets. This discrepancy explains the observed distribution of summaries generated from the different systems as shown in Figure 5b. We can see the system generations have more inter-event

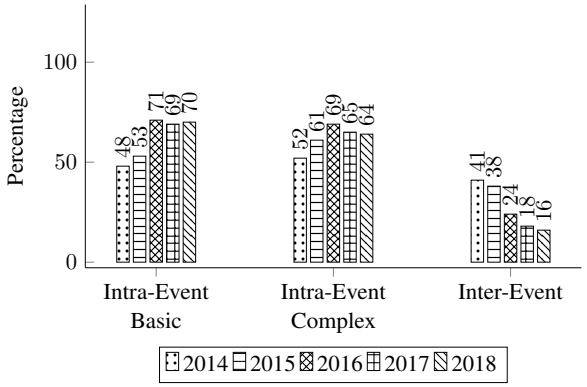

Figure 6: SportSett dataset profile by NBA seasons

sentences because the training data has more inter-event sentences. Concept drift in this D2T problem is captured by our dataset profiling method. This concept-drift can also be explained with the change in authors in different years writing the summaries. In Figure 7, we also show two authors who wrote summaries in different years: 'Auth1' in 2014-15; and 'Auth2' in 2017-18. It is clear that different authors from different years have different distribution which may explain the concept-drift problem in the dataset.

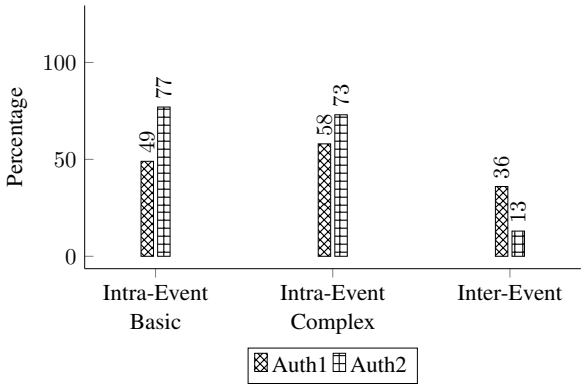

Figure 7: SportSett dataset profile by authors

5 Related Works

Evaluation of texts produced from generation systems is widely used to identify the complexity of a dataset and improve the systems afterwards. There are two main approaches taken for evaluations: automated metrics, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), chrF++ (Popović, 2017), borrowed from machine translation research, or RG (Wiseman et al., 2017), specific to the D2T task; and human evaluations, where users are asked to rate the generations on a Likert scale (Dušek et al., 2020; Chen et al., 2020; Puduppully et al., 2019a). While automated metrics are easier to use, they often correlate poorly with the human evaluation (Reiter, 2018). Human evaluation is considered to be the gold standard for NLG evaluations, however Likert scale based evaluation of single sentences doesn't give a lot of information about the quality of generation. Recent works have proposed to take a more task-oriented evaluation of generated texts that give more insights into the kind of error generation systems make (Thomson and Reiter, 2020, 2021; Kasner et al., 2021; Garneau and Lamontagne, 2021).

There have been few notable works that try to improve the dataset in order to build better generation systems. SportSett dataset by Thomson et al. (2020a) presented an improved resource with better modelling of the dataset to increase the overlap between input data and output text. Gong et al. (2019) and Thomson et al. (2020b) acknowledge the event stream behaviour of D2T tasks by incorporating additional information (both within-event as well as across-event) to improve the quality of generations. In this work, we do not yet try to solve the problem of handling complex or across-event content in summaries. Rather our aim is to profile a dataset through the typology of content types which can be used to identify the complexity of the dataset.

6 Conclusion

In this paper, we presented a typology of different content types in D2T summaries. The proposed typology is used to profile multiple datasets, which captures their characteristics and provide a measure of complexity present in the datasets. Extensive experimentation is performed to demonstrate the challenge facing generation systems in producing complex types of content. We further use the profiling method to identify the concept drift problem in a dataset. Through this work, we argue that a dataset's content type profile can help us define system requirements for building better generation systems.

In future, we plan to employ insights gained from this work in building better D2T generation systems capable of producing accurate complex content for event summaries. This will require a system to be able to: understand the domain-specific rules, for deriving implicit information from the input data: and able to operate in huge search space, to select important content from vast possibility of across-event information. The dataset profiling method will help in identifying any information gap between input data and output summary and characterise the domain-specific rules in different categories based on their information source.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. We also like to thank the members of AI & Reasoning Reading Group at Robert Gordon University, Aberdeen and CLAN Reading Group at University of Aberdeen for their valuable feedback on the work. We also thank Anjana Wijekoon for helpful discussions and proofreading the paper before submission.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020. KGPT: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648, Online. Association for Computational Linguistics.

Emilie Colin, Claire Gardent, Yassine M'rabet, Shashi Narayan, and Laura Perez-Beltrachini. 2016. The WebNLG challenge: Generating text from DBPedia data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 163–167, Edinburgh, UK. Association for Computational Linguistics.

Harm de Vries, Dzmitry Bahdanau, and Christopher D. Manning. 2020. Towards ecologically valid research on language user interfaces. *CoRR*, abs/2007.14435.

Wenchao Du and Alan W Black. 2019. Boosting dialog response generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 38–43, Florence, Italy. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge. *Computer Speech & Language*, 59:123–156.

Steven Y. Feng, Jessica Huynh, Chaitanya Prasad Narisetty, Eduard Hovy, and Varun Gangal. 2021. SAPPHIRE: Approaches for enhanced concept-to-text generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 212–225, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Nicolas Garneau and Luc Lamontagne. 2021. Shared task in evaluating accuracy: Leveraging pre-annotations in the validation process. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 266–270, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

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.

Zdeněk Kasner, Simon Mille, and Ondřej Dušek. 2021. Text-in-context: Token-level error detection for table-to-text generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 259–265, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019a. Data-to-text generation with content selection and planning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6908–6915. AAAI Press.

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, Florence, Italy. Association for Computational Linguistics.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. A hierarchical model for data-to-text generation. In *European Conference on Information Retrieval*, pages 65–80. Springer.

Ehud Reiter. 2007. An architecture for data-to-text systems. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 97–104, Saarbrücken, Germany. DFKI GmbH.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Somayajulu Sripada, Ehud Reiter, and Ian Davy. 2003. Sumtime-mousam: Configurable marine weather forecast generator. *Expert Update*, 6(3):4–10.

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

Craig Thomson and Ehud Reiter. 2021. Generation challenges: Results of the accuracy evaluation shared task. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 240–248, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Somayajulu Sripada. 2020a. SportSett:basketball - a robust and maintainable data-set for natural language generation. In *Proceedings of the Workshop on Intelligent Information Processing and Natural Language Generation*, pages 32–40, Santiago de Compostela, Spain. Association for Computational Lingustics.

Craig Thomson, Zhijie Zhao, and Somayajulu Sripada. 2020b. Studying the impact of filling information gaps on the output quality of neural data-to-text. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 35–40, Dublin, Ireland. Association for Computational Linguistics.

Ashish Upadhyay, Stewart Massie, and Sean Clogher. 2020. Case-based approach to automated natural language generation for obituaries. In *International Conference on Case-Based Reasoning*, pages 279–294. Springer.

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

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Building a Content Type Classifier for Each Dataset

To build the content type classifier for a dataset, some training (ranging from 200 to 600 samples) and testing (250 samples) samples are created by manually annotating the sentences from the train and valid set of that D2T dataset's summaries. Several learning methods (Random Forest, SVM, Logistic Regression) with several features (TF, TF-IDF, dense embeddings from Roberta model from Liu et al. 2019 (RobEmb)) along with a fine-tuned Roberta model are used for building multiple classifiers. A query-by-committee (active learning) approach is used to build the clssifiers' training dataset. The classifier with best Macro-F1 score on held-out test-set of a given dataset is used as the Content Type classifier of that dataset. The performance of Content Type classifiers used for each dataset is shown in Table 3. The table shows the Macro-F1 score of each classifier and the number of samples used for training the classifier. All the trained classifiers with their training and testing samples is shared on the GitHub repo.

B Evaluation Scores of Neural Generation Systems Developed for Different Datasets

Table 4 shows the automated metric scores of different neural systems build for each dataset. The following metrics are used: BLEU (Papineni et al., 2002); ROUGE-L (Lin, 2004); METEOR (Banerjee and Lavie, 2005); chrF++ (Popović, 2017); and BERT-Score (Zhang et al., 2020b). For all metrics, higher score is better (↑).

We also show the results from Extractive Evaluation (EE) metrics (Wiseman et al., 2017) on systems developed for MLB and SportSett datasets in Table 5. These metrics are: Relation Generation (RG); Content Selection (CS); and Content Ordering (CO). The EE scores for systems build on MLB dataset are taken from Puduppully and Lapata (2021) while the EE scores for systems build on SportSett dataset are calculated using Information Extraction system described by Thomson et al. (2020b).

Dataset	MLB	SportSett	SumTime	Obituary
Classifier	RobFT	RobFT	SVM w/ RobEmb	SVM w/ RobEmb
# Train Samples	600	600	200	200
Macro F1	85.64	91.0	98.66	98.46

Table 3: Content Type classifiers performance on various datasets

Systems	BLEU	ROUGE-L	METEOR	chrF++	BERTScore
MLB					
Ent	11.51	22.08	27	32	85.03
Plan	13.99	21.71	32	39	84.6
SportSett					
Ent	18.19	26.19	33	42	86.95
Plan	17.6	26.29	32	40	86.45
Hir	12.18	22.65	33	41	85.74
SumTime					
T5	24.67	52.92	38	47	89.66
BART	18.77	47.61	33	46	88.82
Peg	23.54	51.06	39	48	89.68
Obituary					
T5	47.3	65.38	64	67	94.04
BART	50.88	65.99	66	69	94.29
Peg	45.03	66.4	61	65	93.73

Table 4: Automated metric scores of different systems developed for various datasets (↑, higher is better)

Systems	RG	CS-Precision	CS-Recall	CO
MLB				
Ent	81.1	40.9	49.5	20.7
Plan	94.4	40.8	54.9	21.8
SportSett				
Ent	72.77	45.35	38.12	19.13
Plan	86.48	53.95	33.09	14.84
Hir	73.77	45.42	30.15	10.59

Table 5: Extractive Evaluation results of systems developed for MLB and SportSett dataset (↑, higher is better)