Guided Beam Search to Improve Generalization in Low-Resource Data-to-Text Generation

Nicolas Garneau
Department of Computer Science
University of Copenhagen, Denmark
and Université Laval, Canada

Luc Lamontagne
Department of Computer Science
Université Laval

Abstract

In this paper, we introduce a new beam search algorithm that improves the generalization of neural generators to unseen examples, especially in low-resource data-to-text settings. Our algorithm aims to reduce the number of omissions and hallucinations during the decoding process. For this purpose, it relies on two regression models to explicitly characterize factual errors. We explain how to create a new dataset to train these models given an original training set of less than a thousand data points. We apply our approach in the low-resource, legal setting using the French *Plum2Text* dataset, as well as in English using *WebNLG*. We observe in our experiment that this combination improves the faithfulness of pre-trained neural text generators using both human and automatic evaluation. Moreover, our approach offers a level of interpretability by predicting the number of omissions and hallucinations present in a given generation with respect to the input data. Finally, we visualize our algorithm's exploration of the hypothesis space at different steps during the decoding process.

1 Introduction

Data-to-text generation is commonly referred to as the task of verbalizing a structured input also known as a table of values. The table may contain several types of values such as text, numbers, categories, etc. In our study, we are specifically interested in improving the faithfulness of neural data-to-text generators. The relevance of their generations can be evaluated with respect to the coverage of the input table, i.e. to what extent the model omits values from the table. Moreover, neural text generators unfortunately have the tendency to hallucinate facts from the training set. Hence, generations can also be evaluated based on the number of hallucinated facts produced by the model (Dušek et al., 2018; Ji et al., 2022). The tendency of neural data-to-text generators to omit values and/or hallucinate facts can be exacerbated in low-resource settings. The models may overfit the training set, and bring generalization to unseen data points at stake.

In this paper, we propose to improve the faithfulness of data-to-text neural generators by reducing the number of hallucinations and omissions during the generation process, without having to re-train the generation models. This perspective has many incentives, especially since models are becoming larger and larger, thus harder to train (Brown et al., 2020; Hoffmann et al., 2022). To this end, we propose a modified version of the beam search algorithm specifically for the data-to-text setting. Meister et al. (2020) studied the behavior of the beam search algorithm under a regularized framework, showing that beam search enforces uniform information density. That is, "*it produces text with evenly distributed surprisal, a feature that human readers tend to prefer*". Inspired by this regularization framework, we introduce two characterization models that will guide the decoding algorithm by promoting generated beams containing fewer hallucinations and omissions.

The characterization of omissions and hallucinations is crucial in the legal setting. Hence, we apply our new decoding algorithm and analyze its benefits on the task of verbalizing criminal docket files using the *Plum2Text* dataset (Beauchemin et al., 2020; Garneau et al., 2021b). Using automatic and manual evaluation, we show that our algorithm improves generalization in a low-resource setting, especially on unseen data points. We also show that our approach generalizes to other datasets, such as WebNLG (Castro Ferreira et al., 2020). In the next section, we introduce related work regarding the mitigation of omissions and hallucinations for neural text generators. We then introduce the main contribution of this paper in Section 3, a new decoding algorithm for the data-to-text setting. We present the experiments and analysis in Section 4. We assess the generalization of our approach in Sec-

Proceedings of the 16th International Natural Language Generation Conference, pages 1–14
September 11–15, 2023. ©2023 Association for Computational Linguistics

tion 5 by applying it on WebNLG (Castro Ferreira et al., 2020). We conclude with our observations in Section 6. The models, datasets, generations and human evaluations are made publicly available[1].

2 Related Work

In this section, we study current mitigation techniques of omissions and hallucinations to improve neural generators' performance, limiting ourselves to the data-to-text setting[2]. These techniques *may* require changing the architecture and are enforced either during training or during inference. We classify mitigation techniques as being "invasive" or "non-invasive" to the generator. Invasive techniques require fine-tuning, adding a new objective function, or modifying the inner architecture of the generator.

In this paper, we focus on non-invasive techniques, which consider the generator as a black box and act either on the input or during the decoding process. These techniques are appealing for the fact that they do not require re-training the generator on the original training dataset. For example, Shin et al. (2020) proposed AutoPrompt, a model that learns how to create prompts for various sets of tasks. They basically search for "trigger" tokens using the gradient from the downstream task. According to their results, AutoPrompt outperformed fine-tuning methods in cases where the training dataset is small (i.e. 100-1000 samples). This method, however, does not necessarily mitigate the omission and hallucinations in the data-to-text setting. Similarly, Prefix-Tuning (Li and Liang, 2021) proposed a lightweight alternative to fine-tuning for natural language generation tasks, which keeps language model parameters frozen, but optimizes a small continuous task-specific vector, called the prefix. Then again, their method does not specifically mitigate omissions and hallucinations.

Ghazvininejad et al. (2017) proposed *Hafez*, a method weighing the current beam state based on a set of feature functions that take as input a target word, and sometimes the beam state (e.g. to check for repetitions). These feature functions, in our case, could be used to force the generation of proper charges, decisions, and pleading, for example. They added two terms to the standard Beam Search algorithm, given a current beam state and a predicted word;

$$score(b_i, w) = score(b_{i-1}) + \log Gen(w) + \sum_j \alpha_j * f_j(w); \forall w \in V_{suc} \quad (1)$$

where $score(b)$ is the score of the current beam state, $\log Gen(w)$ is the output logit of the generator, $f(*)$ are functions that scores word w weighted by α_i, and V_{suc} is a predefined vocabulary. Similarily, *Mention Flags* (Wang et al., 2021) tries to identify the presence of tokens in the hypothesis given a set of flags. Both methods face the same problem since they operate on surface tokens.

Anderson et al. (2017) also proposed to constrain the beam search algorithm operating at the lexical level using a finite-state machine that enforces the use of a specific vocabulary in the image captioning setting. However, their method does not scale well when the input is composed of sentences, since we don't know apriori the vocabulary we want to constrain. Balakrishnan et al. (2019) proposed a constrained decoding technique that leverages tree-structured meaning representations to control the semantic correctness of the generated text. While not explicitly characterizing omissions and hallucinations, their approach improved the faithfulness of the generative models. The prior work closest to ours is RANKGEN (Krishna et al., 2022), a ranking model that can be incorporated into the beam search scoring function during the decoding process. However, their method is designed for open-ended generation and does not yet scale to methods having a constrained output such as data-to-text, summarization, and machine translation.

Guerreiro et al. (2023) introduced DEHALLUCINATOR, a model that flags hypotheses once they are fully generated so that they can be overwritten. Our model differs from their approach since we are guiding the exploration of the tree during decoding. Finally, (Vijayakumar et al., 2016) introduced Diverse Beam Search, an algorithm that promotes diverse generations amongst groups of beams but does not strictly reward or penalize beams for specific properties. To the best of our knowledge, no method in the literature proposes a way that can be adapted without major changes to handle both omissions and hallucinations at the semantic level during the decoding step. Moreover, none of the methods can explicitly estimate the number of hallucinations and omissions in the hypotheses. We

[1] https://drive.google.com/drive/folders/1gdgWXr7GXoww2dC83xcH3TXgwyB3ov4W?usp=sharing

[2] One can see the mitigation as the optimization of the precision (reduce hallucination) and the recall (reduce omissions) of a given data-to-text generator.

thus wish to fill this gap by proposing a guided beam search algorithm to create more faithful neural data-to-text generations.

3 Guided Decoding by Predicting Omissions and Hallucinations

In this section, we introduce a new decoding algorithm that is designed to mitigate and explicitly characterize omissions and hallucinations for data-to-text generation. To this end, we create two predictive models: one predicting the number of omitted values from the table, m_o, and the other predicting the number of hallucinations, m_h. These models will thus weigh the current beam score to promote generated sequences with few, or hopefully no omissions or hallucinations, enforcing semantically accurate generations.

3.1 Characterization Models

The proposed models are designed to take as input the table's values, as well as the current generated sequence, and output a real value as the following;

$$o_i = m_o(V_i, s_i) \tag{2}$$

$$h_i = m_h(V_i, s_i) \tag{3}$$

where o_i is the predicted number of omissions, h_i is the predicted number of hallucinations. m_o is the omission model, m_h is the hallucination model, V_i is the set of table of values, and s_i is the current generated sequence. To obtain these models, we need to train them using a dataset that has as input the table, the generated sequence as well as their true labels, i.e. the number of omissions and hallucinations in the sequence. We further detail in the next section how we obtain such datasets from the original training set using *Plum2Text* as an example (Plum2Text's training set contains around 1K examples).

3.2 Training Data

We hereby propose to build one training dataset for each model, $\mathcal{O}$ and $\mathcal{H}$, based on the overlapping table values across the original training examples. It is important to note that each actual training example is used in both $\mathcal{O}$ and $\mathcal{H}$, labeled with zero omission and zero hallucination respectively. We create the other training examples as follows;

1. We randomly select two training instances (V_i, r_i), and (V_j, r_j) where r_i and r_j are reference texts of both examples
2. The set of omitted values O_i for r_j with respect to V_i correspond to the set difference between V_i and V_j
3. Similarly, the set of hallucinated values H_i for r_i with respect to V_j correspond to the set difference between V_j and V_i.

We formally describe the dataset creation in Algorithm 1 and we illustrate in Figure 1 the construction of a training example, created from two original examples taken from the *Plum2Text* dataset.

Algorithm 1 Creating Datasets $\mathcal{O}$ and $\mathcal{H}$

```
$\mathcal{O} \leftarrow \{\}$                                   ▷ set of omissions
$\mathcal{H} \leftarrow \{\}$                                   ▷ set of hallucinations
for $(V_i, r_i)$, $(V_j, r_j)$ in the training set do
    $O_i \leftarrow V_i \setminus V_j$            ▷ set diff. between $V_i$ and $V_j$
    $H_i \leftarrow V_j \setminus V_i$            ▷ set diff. between $V_j$ and $V_i$
    $\mathcal{O} \leftarrow \mathcal{O} \cup \{(V_i, r_j), |O_i|\}$
    $\mathcal{H} \leftarrow \mathcal{H} \cup \{(V_i, r_j), |H_i|\}$
end for
return $\mathcal{O}, \mathcal{H}$
```

Using *Plum2Text*, the omissions dataset $\mathcal{O}$ consists of 12,460 examples using an 80%–20% split resulting in train and test sets of 9,968 and 2,492 examples respectively. The hallucination dataset $\mathcal{H}$ consists of 30,473 examples also using an 80%–20% split resulting in train and test sets of 24,378 and 6,095 examples respectively. With respect to the training architecture, we used the multilingual version of BERT (Devlin et al., 2019) of 178M parameters available in the HuggingFace library[3]. We used the mean squared error loss and AdamW (Loshchilov and Hutter, 2019) as the optimizer with a learning rate of 0.001. We used a batch size of 10 on a GeForce 2080Ti Nvidia graphic card. To automatically evaluate the architectures, we considered several metrics: mean squared error (MSE), root mean square error (RMSE), mean average error, $\mathcal{R}^2$, and accuracy defined as follows;

$$a = \begin{cases} 1 \text{ if } p - t < 0.5 \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

where p is the prediction and t is the true value. As we can see in Table 1, both models achieve high performance across all metrics on the test set.

[3] We used the multilingual BERT (Devlin et al., 2019) because it provides a version with a pre-trained classification head, whereas CamemBERT (Martin et al., 2020) does not.

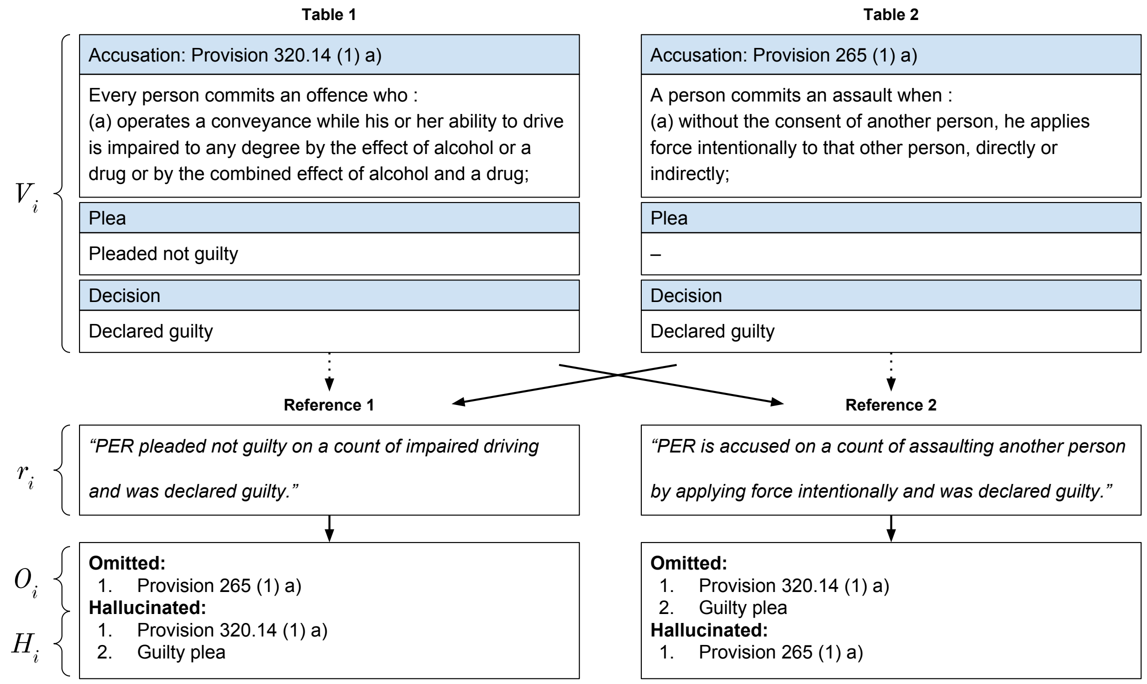


Table 1
Table 2
V_i
Accusation: Provision 320.14 (1) a)
Every person commits an offence who :
(a) operates a conveyance while his or her ability to drive is impaired to any degree by the effect of alcohol or a drug or by the combined effect of alcohol and a drug;
Plea
Pleaded not guilty
Decision
Declared guilty
Accusation: Provision 265 (1) a)
A person commits an assault when :
(a) without the consent of another person, he applies force intentionally to that other person, directly or indirectly;
Plea
–
Decision
Declared guilty
Reference 1
Reference 2
r_i
"PER pleaded not guilty on a count of impaired driving and was declared guilty."
"PER is accused on a count of assaulting another person by applying force intentionally and was declared guilty."
O_i
H_i
Omitted:
1. Provision 265 (1) a)
Hallucinated:
1. Provision 320.14 (1) a)
2. Guilty plea
Omitted:
1. Provision 320.14 (1) a)
2. Guilty plea
Hallucinated:
1. Provision 265 (1) a)

Figure 1: Given two training instances from *Plum2Text*, each with their respective table and reference, we pair the table from the first example with the reference of the second one and vice versa. This creates in total four training instances, two in each dataset $\mathcal{O}$ and $\mathcal{H}$. The first two "omitted" training instances are V_j, r_i paired with the omitted value "Provision 265 (1) a)", and V_i, r_j paired with the 2 omitted values "Provision 320.14 (1) a)" and "Guilty plea". The same procedure applies for the creation of the hallucinated training instances.

We also show the distribution of predicted vs actual values in Figure 2 using confusion matrices. The regression model on the omissions tends to underestimate the number of omissions in a given generation. The regression model on the hallucinations seems more balanced except for the cases where there are one or two hallucinations, underestimating them.

	Models	
Metric	**Omission**	**Hallucination**
MSE	0.05	0.05
RMSE	0.23	0.22
MAE	0.10	0.08
$\mathcal{R}^2$	0.99	0.99
Accuracy	0.96	0.97

Table 1: Performance of both omission and hallucination models on *Plum2Text* w.r.t the mean squared error (MSE), the root mean squared error (RMSE), the mean average error (MAE), $\mathcal{R}^2$, and accuracy.

3.3 Guided Decoding for Omission and Hallucination Mitigation

In order to mitigate omissions and hallucinations, we propose the following weighted beam search score b_i;

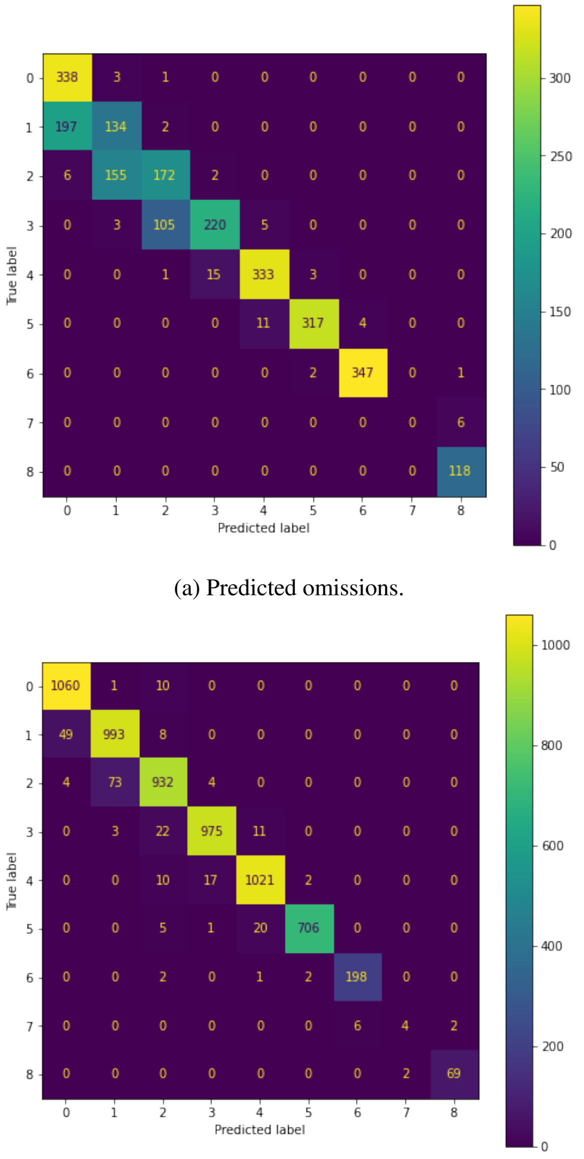
$$b_i = score(b_{i-1}) + \log(Gen(w_i)) + \phi_i \qquad (5)$$

where $score(b_{i-1})$ is the previous beam's score, $\log(Gen(w_i))$ is the score for word w_i provided by the generator, and ϕ is the following function based on the omission and hallucination scores o_i and h_i obtained from the characterization models:

$$\phi_i = \omega \cdot (v_i - o_i) - \gamma \cdot h_i \qquad (6)$$

where ω and γ are parameters to weigh the omissions and hallucinations respectively. While the hallucinations h_i are treated as a penalty on a beam score, the omissions are treated as a reward: v_i corresponds to the actual number of values in the table, whereas o_i is the number of detected omissions. If zero omissions are detected, the current beam will get a reward of $\omega \cdot v_i$.

In our experiments, ω and γ are initialized to 1 and we perform a grid search over a set of values between 0.0 and 5.0 to find the optimal ones depending on the use case. The number of beams b



(a) Predicted omissions.

(b) Predicted hallucinations.

Figure 2: Confusion matrices of the predicted omissions and hallucinations by the regression models on the *Plum2Text* test set.

parameterizes the original beam search algorithm. During the generation process, omission rewards and hallucination penalties are cumulated at each step. Regardless of the values of ω and γ, we apply a final processing step to fully reevaluate the ranking of the candidates w.r.t the generator's final log-likelihood and the omission/hallucination models using values of 1 for both ω and γ. This is motivated by the fact that the models, trained on full sentences, may provide more accurate predictions and thus result in a better candidate ranking.

4 Experiments

In our experiment, we use *CriminelBART*, a generative model introduced by Garneau et al. (2021a). We only analyze the vanilla and guided versions of *CriminelBART* since other methods proposed in the literature do not explicitly mitigate omissions and hallucinations. We trained *CriminelBART* on the train set of *Plum2Text*, and we begin by automatically evaluating different versions of the weighted beam search using a grid search over the hyper-parameters previously introduced. We then manually evaluate the performance of our new algorithm in Section 4.2. To assess the generalization performance of our algorithm, we added examples with 37 new provisions from the Criminal Code of Canada having no or very few occurrences in the original training set. Furthermore, we qualitatively analyze the behavior of our algorithm in Section 4.3.

4.1 Guided Decoding

In order to find the best generation model using the weighted decoding algorithm aforementioned, we performed a grid search exploration with the following hyper-parameters:

1. ω, the weight for omission detection.
2. γ, the weight for hallucination detection.
3. β, the number of beams.

Parameters	**Values**
Omission – ω	0.0, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0
Hallucination – γ	0.0, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0
Beam size – β	5, 10, 15

Table 2: Hyper-parameters search on the omission and hallucination weight (ω and γ respectively) and the beam size β.

Table 2 provides the values tried for each hyper-parameter. Among the 147 combinations, the best model uses weights of 0.2 for both omissions and hallucinations and a beam size of 15.

Evaluation results are presented in Table 3 for both the best-performing model using guided decoding and the original version of *CriminelBART*. We considered BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BertScore (Zhang et al., 2020) dubbed as BScore, NLI (Dušek and Kasner, 2020), and RANK (Garneau and Lamontagne, 2021). It has been shown that RANK highly correlates with human judgment (CITE), so we used this metric in the cross-validation step to select the best hyperparameters, ω, γ, and β, for each model. The results indicate that guided *CriminelBART* outperforms the original *CriminelBART* on 6 automatic evaluation metrics out of 9. The guided version of *CriminelBART* obtains similar performance with respect to BLEU-1, METEOR, and NLI.

The guided version of *CriminelBART* using the post-processing step described in the previous section obtains similar performance but we observe an interesting two-point gain on the RANK metric, improving from 0.76 to 0.78, over the original version limited to 0.72. It is important to note that RANK tends to have the highest correlation score with respect to human evaluation. Overall, we can conclude that the guided version of *CriminelBART* obtains better performance than the original one by up to 6 points with respect to the RANK metric. We also note that the number of predicted hallucinations and omissions also considerably decrease, going from 0.28 and 0.24 to 0.11 and 0.11 respectively. In the next section, we manually evaluate the generations.

4.2 Human Evaluation

In this section, we further analyze the generalization performance of both models by considering 45 table values that are either not in the training set or appear rarely. We hired three annotators that followed the same evaluation procedure introduced by Garneau et al. (2022) to manually assess the performance of both models. For our application, these table values correspond to legal provisions from the Criminal Code of Canada (CCC). From these 45 provisions, we added 37 new ones that we selected by skimming through the whole CCC. These are listed in Appendix A. We list down in Table 4 the whole set of provisions considered in this manual evaluation. We decided to not manually evaluate examples where other provisions were found often in the training set because both models are having a similar performance for these frequent cases.

We manually evaluated the generations of both the original version of *CriminelBART* and the model using guided beam search. We recruited three evaluators from a Faculty of Law that assigned a score between 1 to 10, 1 corresponding to a generation completely off-track, and 10 being a perfect generation. We used Krippendorff's alpha coefficient (Krippendorff, 2004) to analyze the inter-annotator agreement which is 0.69. We can see from Table 4 that the guided version of *CriminelBART* achieves better generalization performance on unseen provisions with an average score of 7.4, compared to the original version with a score of 3.9. That is, guided *CriminelBART* produces generations that verbalize the good provision with some hallucinations and/or omissions, whereas the original version mostly generates on-theme or off-track descriptions. It seems like the hallucination and omission models enable better exploration of the generation tree than regular beam search using maximum log-likelihood estimation. This can lead to better generations when using a higher number of beams (Meister et al., 2020). We specifically discuss and illustrate this phenomenon in Section 4.3.

Comparing *CriminelBART* and Guided *CriminelBART*, we found that for 10 out of 45 generations, the original version of *CriminelBART* generated commonly seen provisions such as 320.14 (driving under the influence), 266 (assault and battery), or 151 (sexual interference). We provide an example in Table 5 where the guided *CriminelBART* generated the good provision, but the original version generated unrelated content with respect to the input. There is one particular case where the original version produced a better generation which is on provision 345, "Stopping mail with intent" (see Table 4). Indeed, the guided version of *CriminelBART* produced a generation not capturing the act of stealing **mail**, while the original version did. In every other case, the original version attempted at generating meaningful content as being "on-theme", but most of the time the guided version was able to generate the right provision, with some factual errors, having a score above 5.

4.3 Beam Search Analysis

To better understand the behavior of our approach, we analyze the different steps in the beam search algorithm of both models for one generation involving table value "provision 431": *Attack on-premises, residence or transport of internationally protected person*. We illustrate in Figure 3 the paths taken by the two versions of beam search. The starting point, where the algorithms respectively branched on their own, is illustrated in blue.

			BLEU									Rates	
ω	γ	β	1	2	3	4	ROUGE	METEOR	BScore	NLI	RANK	Hal.	Om.
0.0	0.0	5	0.73	0.58	0.47	0.41	0.42	0.38	0.78	0.34	0.72	0.28	0.24
0.2	0.2	15	0.73	**0.59**	**0.48**	**0.43**	**0.44**	0.38	**0.79**	0.34	**0.76**	0.13	0.11
Post processing			0.73	**0.58**	**0.48**	**0.42**	**0.43**	0.37	**0.79**	0.34	**0.78**	**0.11**	**0.11**

Table 3: Automatic evaluation results of the best performing original *CriminelBART* ($\omega = 0.0$, $\gamma = 0.0$, $\beta = 5$), the best-performing model using the weighted beam search algorithm ($\omega = 0.2$, $\gamma = 0.2$, $\beta = 15$), and that same model using the post-processing finalization step.

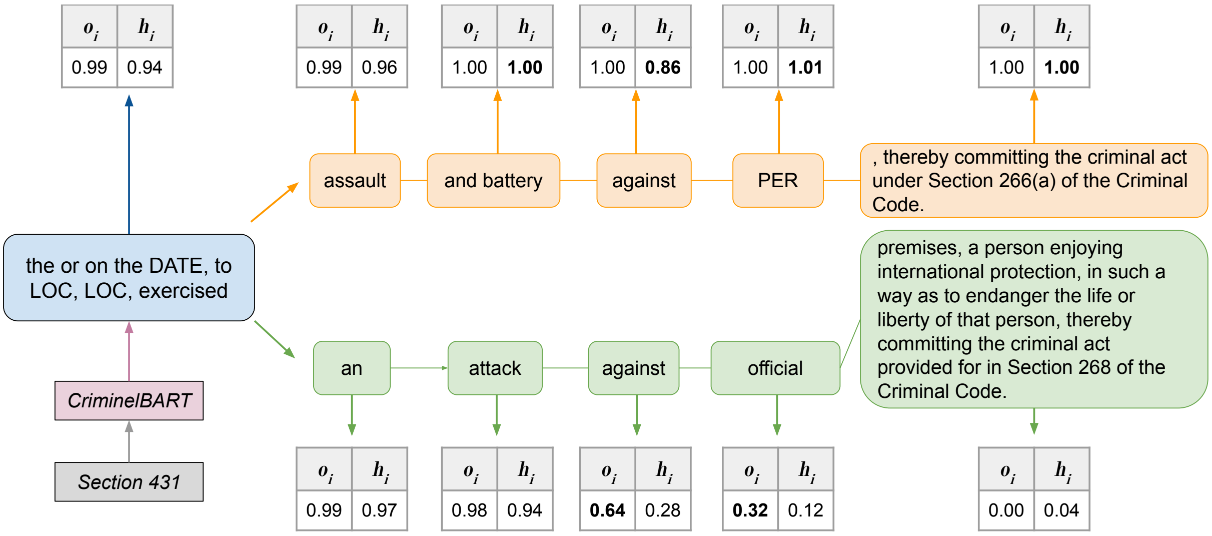

Section 431
CriminelBART
the or on the DATE, to LOC, LOC, exercised
o_i h_i 0.99 0.94
assault
and battery
against
PER
, thereby committing the criminal act under Section 266(a) of the Criminal Code.
o_i h_i 0.99 0.96
o_i h_i 1.00 1.00
o_i h_i 1.00 0.86
o_i h_i 1.00 1.01
o_i h_i 1.00 1.00
an
attack
against
official
premises, a person enjoying international protection, in such a way as to endanger the life or liberty of that person, thereby committing the criminal act provided for in Section 268 of the Criminal Code.
o_i h_i 0.99 0.97
o_i h_i 0.98 0.94
o_i h_i 0.64 0.28
o_i h_i 0.32 0.12
o_i h_i 0.00 0.04

Figure 3: Analysis of *CriminelBART* using the original (orange) and the guided (green) beam search algorithms on the generation of provision 431 (translated in English): "*Attack on-premises, residence or transport of internationally protected person.*". The predicted number of omissions (o_i) and hallucinations (h_i) are presented at each timestep.

The original beam search algorithm is illustrated in yellow, while the guided beam search is illustrated in green. Each time step in the figures may be an aggregation of several generation steps, for easier understanding. Each time step is associated with the predicted number of omissions and hallucinations, o_i and h_i respectively, with respect to their associated models.

In this particular example, the models start with omitted and hallucinated values of one regarding the initial generation "*the or on the DATE, at LOC, LOC, exercised*"[4]. The decoding algorithm branch out on the next token, generating "assault" for the original version and "an" for the guided one. It is only with two generation steps that the guided beam search obtains lower predicted values in terms of omissions and hallucinations (0.64 and 0.28 respectively). The final generation obtains scores of omissions and hallucinations of 0.00 and 0.04. The original version of the beam search on the other hand clearly omits to generate the proper provision, and hallucinates the provision of "assault", ending with both omission and hallucinations scores of 1.0. Finally, the original version of *CriminelBART* obtains a human evaluation score of 1.0, compared to the guided version having 8.33.

5 Generalization of the Approach

We illustrate the generalization of our proposed approach to improve the beam search algorithm to other data-to-text settings by using WebNLG, a well-known dataset in the NLP community. To this end, we used the same methodology described in Section 3.3:

- From the original WebNLG training dataset, we build two datasets, $\mathcal{O}$ and $\mathcal{H}$. Training instances of these datasets consist of a set of triplets each containing a table, a generation,

[4] DATE, PER and LOC are special tokens from the *Plum2Text¨ dataset where dates, persons, and locations have been anonymized.*

Provision	*CriminelBART*	Guided *CriminelBART*
46	1.00	8.00
57	3.00	8.00
58	2.33	7.00
83.04	2.67	8.00
83.08	3.00	8.00
83.21	5.33	8.00
83.181	1.00	8.00
123	1.00	8.00
148	7.67	8.67
150	3.67	8.33
170	2.33	5.00
173	2.33	8.33
202	1.00	4.67
218	1.00	5.67
243	4.33	6.67
245	2.00	7.33
253	6.00	8.00
267	6.33	8.00
270.1	3.33	8.67
318	7.00	8.33
342	8.67	9.00
342.1	2.33	9.67
344	4.00	8.67
345	7.67	1.00
347	1.00	6.00
351	7.00	9.00
354	3.00	8.00
355	5.00	7.67
356	1.00	7.67
364	1.00	8.67
368	7.33	9.00
374	4.67	5.00
382.1	8.33	4.00
398	8.00	6.00
402.2	8.00	8.33
406	3.33	8.00
431	1.00	8.33
432	5.00	4.33
437	1.00	4.33
438	5.67	8.33
439	2.33	8.33
445.1	3.00	9.00
446	2.33	8.67
467.111	8.33	8.67
810.2	2.33	5.67
Average	**3.9**	**7.4**

Table 4: Human evaluation of the original version of *CriminelBART* and the one using guided beam search on the 45 unseen provisions.

and the associated number of omissions or hallucinations.

- Using the previously created datasets, we train two models to predict the number of omissions and hallucinations given the input table and its corresponding generation.
- We use the trained models to predict, during the decoding process, the number of omissions and hallucinations and weigh the beams accordingly.
- We apply the finalization step to select the best hypothesis.

The omission dataset $\mathcal{O}$ of WebNLG consists of 20,448 examples resulting in train and test sets of 16,358 and 4,090 examples respectively using an 80%–20% split. The hallucination dataset $\mathcal{H}$ consists of 20,600 examples resulting in train and test sets of 16,480 and 4,120 examples respectively also using an 80%–20% split. Similar to the *Plum2Text* setting, we trained the English version of BERT (Devlin et al., 2019) available in the HuggingFace library using the same hyperparameters previously selected. For the actual data-to-text generation task we trained BART (Lewis et al., 2020) on the training set of WebNLG and evaluated it on the test set after performing a grid-search over the guided beam search hyperparameters. We used the same automatic evaluation metrics as with *Plum2Text*. We can see from Table 6 that the guided version of BART on WebNLG improves the performance on 6 metrics out of 9. Considerable improvements are made regarding NLI and RANK, two metrics that were proven to be correlated with manual evaluation of WebNLG test instances (Garneau et al., 2022). Similar to the *Plum2Text* case, using the post-processing step to re-rank candidates improved the generations for almost all metrics. While the hallucination and omission rates are already low, using the guided version of BART slightly improves them.

6 Conclusion

In this paper, we introduced a new guiding mechanism for the beam search algorithm in the data-to-text generation setting. We presented how to train two models, one to predict the number of omissions, and the other for the number of hallucinations. These predictors are used not only to mitigate the number of omissions and hallucinations

Input Data	**Provision:** 123: Influencing a municipal officer; is guilty of an indictable offence and liable to imprisonment for a term not exceeding five years or is guilty of an offence punishable on summary conviction if he influences or attempts to influence a municipal officer to do any of the things referred to in paragraphs (1)(a) to (d): by threats or deception.
***Criminel*BART** Score: 1.0	On the LABELD, PER is arrested for driving with more than 80 mg of alcohol per 100 ml of blood.
Guided ***Criminel*BART** Score: 8.0	On or about LABELD, at LOC, LOC, willfully attempted to influence a municipal official, thereby committing the criminal act under section 120(a) of the Criminal Code.

Table 5: Comparing the generation of the original *CriminelBART* and the model using guided beam search on provision 123: “Influencing a municipal officer”. The original version of *CriminelBART* generates one of the most common provisions, 320.14 (driving under the influence) resulting in a score of 1.0, whereas the model using guided beam search generates a description about the right provision resulting in a score of 8.0.

			BLEU									Rates	
ω	γ	β	1	2	3	4	ROUGE	METEOR	BScore	NLI	RANK	Hal.	Om.
0.0	0.0	5	0.81	0.71	0.64	0.58	0.55	0.54	0.94	0.63	0.64	0.11	0.00
0.2	0.2	15	**0.83**	**0.73**	**0.65**	**0.59**	0.53	0.54	0.94	**0.68**	**0.65**	0.10	0.00
Post processing			**0.84**	**0.74**	**0.66**	**0.60**	0.54	0.54	0.94	**0.68**	**0.66**	0.10	0.00

Table 6: Automatic evaluation results of the best performing BART model on WebNLG ($\omega = 0.0$, $\gamma = 0.0$, $\beta = 5$) and the best-performing BART model using the weighted beam search algorithm ($\omega = 0.2$, $\gamma = 0.5$, $\beta = 10$).

but also to favor the exploration of the possible generation space. This new mechanism improves the generation quality with respect to automatic evaluation metrics and shows significant generalization improvement regarding unseen data points during human evaluation. Moreover, our mechanism offers a new degree of a posteriori interpretability given a list of potential hypotheses, since the characterization models provide estimates of the number of omissions and hallucinations. Finally, we showed that our approach generalizes not only to *Plum2Text*, a challenging low-resource dataset but also to a well-known dataset such as WebNLG. In future works, it would be interesting to investigate the identification of omitted values and hallucinated tokens. The identification of omitted values is easier to perform since we already provide a way to build such a dataset and train a model accordingly. However, identifying the hallucinated tokens requires a sequence-to-sequence tagger and its respective training set, which most likely can only be obtained with manual annotations.

Ethics Statement

The scope of this work is to improve the faithfulness of neural data-to-text generators. Faithfulness is extremely important in the legal field since we do not want to generate false accusations about litigants. There is a potential risk to using neural data-to-text generators in production, and we provided not only improve their performance but also analyzed their behavior. In the end, the purpose of this work is largely motivated by the ethical use of neural text generators and a better understanding of their implications.

Acknowledgements

We thank the reviewers for their insightful comments on our manuscript. This work was partly funded by the Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG, https://www.nserc-crsng.gc.ca/index_fra.asp) and the Fonds de recherche du Québec – Nature et technologies (FRQNT, https://frq.gouv.qc.ca/nature-et-technologies/).

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 936–

945, Copenhagen, Denmark. Association for Computational Linguistics.

Anusha Balakrishnan, Jinfeng Rao, Kartikeya Upasani, Michael White, and Rajen Subba. 2019. Constrained decoding for neural NLG from compositional representations in task-oriented dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 831–844, Florence, Italy. Association for Computational Linguistics.

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.

David Beauchemin, Nicolas Garneau, Eve Gaumond, Pierre-Luc Déziel, Richard Khoury, and Luc Lamontagne. 2020. Generating intelligible plumitifs descriptions: Use case application with ethical considerations. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 15–21, Dublin, Ireland. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina, editors. 2020. *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*. Association for Computational Linguistics, Dublin, Ireland (Virtual).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the E2E NLG challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 322–328, Tilburg University, The Netherlands. Association for Computational Linguistics.

Nicolas Garneau, Eve Gaumond, Luc Lamontagne, and Pierre-Luc Déziel. 2021a. Criminelbart: A french canadian legal language model specialized in criminal law. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 256–257, New York, NY, USA. Association for Computing Machinery.

Nicolas Garneau, Eve Gaumond, Luc Lamontagne, and Pierre-Luc Déziel. 2021b. Plum2text: A french plumitifs–descriptions data-to-text dataset for natural language generation. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*, Sao Paulo, Brazil. International Association for Artificial Intelligence and Law.

Nicolas Garneau, Eve Gaumond, Luc Lamontagne, and Pierre-Luc Déziel. 2022. Evaluating legal accuracy of neural generators on the generation of criminal court dockets description. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 73–99, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Nicolas Garneau and Luc Lamontagne. 2021. Trainable ranking models to evaluate the semantic accuracy of data-to-text neural generator. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 51–61, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48, Vancouver, Canada. Association for Computational Linguistics.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Comput. Surv.* Just Accepted.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.

Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022. Rankgen: Improving text generation with large ranking models. *ArXiv*, abs/2205.09726.

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.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.

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.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Yufei Wang, Ian Wood, Stephen Wan, Mark Dras, and Mark Johnson. 2021. Mention flags (MF): Constraining transformer-based text generators. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 103–113, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A New Unseen Provisions

List of the 37 new unseen provisions and their associated texts.

- **46 (1) a)**: "*High treason. Every person commits high treason who, in Canada, wages war against Canada or does any act preparatory thereto;*"

- **57 (2)**: "*Misrepresentation in relation to a passport. Every person who, in Canada or elsewhere, for the purpose of obtaining a passport for himself or herself or for another person or for the purpose of obtaining a material alteration or addition to such a passport, makes a written or oral statement that he or she knows to be false or misleading is guilty.*"

- **58 (1) a)**: "*Fraudulent use of citizenship certificate. Every person who, while in Canada or outside Canada, as the case may be, uses a certificate of citizenship or a certificate of naturalization for a fraudulent purpose is guilty of an indictable offence and liable to imprisonment for a term not exceeding two years or is guilty of an offence punishable on summary conviction;*"

- **83.04 a)**: "*Using or possessing property for terrorist purposes. Any person who: directly or indirectly uses property, in whole or in part, for or to facilitate a terrorist activity is guilty of an indictable offense punishable by imprisonment for not more than ten years;*"

- **83.08 (1) a)**: "*Freezing of property. No person in Canada and no Canadian outside Canada shall: knowingly deal with property owned or controlled, directly or indirectly, by a terrorist group;*"

- **83.21 (1)**: "*Instructing a person to carry out an activity for a terrorist group. Every person who knowingly directs, directly or indirectly, any person to carry out any activity for the benefit of, at the direction of, or in association with a terrorist group for the purpose of enhancing the ability of any terrorist group to facilitate or carry out a terrorist activity is guilty of an indictable offence and liable to imprisonment for life.*"

- **123 (2)**: "*Influencing a municipal officer. Every person who influences or attempts to influence a municipal officer to do anything mentioned in paragraphs (1)(a) to (d) is guilty of an indictable offence and liable to imprisonment for a term not exceeding five years or is guilty of an offence punishable on summary conviction:*"

- **148 a)**: "*Assisting prisoner of war to escape. Every one who knowingly: aids a prisoner of war in Canada to escape from a place of confinement is guilty of an indictable offence and liable to imprisonment for a term not exceeding five years or is guilty of an offence punishable on summary conviction;*"

- **170**: "*Father, mother or guardian who procures. A parent or guardian of a person under the age of eighteen years who causes that person to engage in sexual acts prohibited by this Act with a third party is guilty of an indictable offence and liable to imprisonment for a term not exceeding fourteen years and to a minimum punishment of one year.*"

- **173 (2)**: "*Exhibitionism. Any person who, in any place whatsoever, for sexual purposes, exhibits his or her genitals in front of a person under the age of sixteen years is guilty of:*"

- **202 (1) a)**: "*Gambles, bookmaking, etc. Every person commits an offence who: uses or knowingly permits to be used any premises under his control for the purpose of registering or recording bets or selling a pool bet;*"

- **218**: "*Abandonment of child. Whoever unlawfully abandons or exposes a child under the age of ten years, so that the life of such child is actually endangered or exposed to be endangered, or the health of such child is actually permanently endangered or exposed to be endangered, is guilty of:*"

- **243**: "*Suppression of part. Whoever in any way causes the corpse of a child to disappear with the intention of concealing the fact that its mother gave birth to it, whether the child died before, during or after birth, is guilty:*"

- **245 (1)**: "*Administering deleterious substance. Whoever administers or causes to be administered to any person any poison or other destructive or deleterious substance, shall be guilty of:*"

- **270.1 (1)**: "*Disarming a peace officer. Every person commits an offence who takes or attempts to take a weapon from the possession of a peace officer acting in the performance of his or her duties, without the consent of the peace officer.*"

- **318 (1)**: "*Advocacy of genocide. Anyone who advocates or foments genocide is guilty of an indictable offence and liable to imprisonment for a term not exceeding five years.*"

- **342 (3)**: "*Unauthorized use of credit card data. Any person who fraudulently and without the appearance of right has in his possession or uses data, whether genuine or not, relating to a credit card, including a personal authenticator, which would enable the use of the same or the obtaining of services connected with its use, traffics in such data or allows another person to use the same, shall be guilty:*"

- **342.1 (1) a)**: "*Unauthorized use of computer. Every person who fraudulently and without colour of right, directly or indirectly, obtains computer services is guilty of an indictable offence and liable to imprisonment for a term not exceeding ten years or is guilty of an offence punishable on summary conviction;*"

- **345**: "*Stopping the mail with intent to rob. Anyone who stops a mail transport with the intention of stealing or searching it is guilty of a criminal act and liable to life imprisonment.*"

- **347 (1)**: "*Criminal rate of interest. Notwithstanding any other federal law, any person who enters into an agreement or arrangement to charge interest at a criminal rate or charges interest, even partially, at a criminal rate is guilty of:*"

- **351 (1)**: "*Possession of burglary tools. Whoever, without lawful excuse, has in his possession any instrument which may be used to break into any place, motor vehicle, vault or safe, knowing that the instrument has been used or is intended to be used for such purpose, is guilty of:*"

- **354 (2)**: "*Possession of motor vehicle with identification number obliterated. In proceedings under subsection (1), evidence that a person is in possession of a motor vehicle, or any part thereof, the identification number of which has been wholly or partly removed or obliterated is, in the absence of any evidence to the contrary, proof that it was obtained by the commission in Canada of an offence punishable on indictment;*"

- **356 (1) a)**: "*Theft of mail. Every person commits an offence who: steals anything sent by mail after it has been deposited in a post office and before it is delivered, or after it has been delivered but before it is in the possession of the addressee or any person who may reasonably be regarded as authorized by the addressee to receive the mail;*"

- **364 (1)**: "*Fraudulent obtaining of food and lodging. Any person who fraudulently obtains food, drink, or other commodities in any establishment dealing in them is guilty of a summary conviction offense.*"

- **368 (1) a)**: "*Using, possessing or trafficking in a forged document. Every person commits an offence who, knowing or believing that a document is counterfeit, as the case may be: uses, treats or acts with respect to it as if it were genuine;*"

- **374 (a)**: "*Unauthorized drafting of document. Any person who, with intent to defraud and without lawful authority, makes, subscribes, draws, signs, accepts or endorses a document in the name of or on behalf of another person, by proxy or otherwise, is guilty of an indictable offence and liable to imprisonment for a term not exceeding fourteen years;*"

- **382.1 (1) a)**: "*Insider trading. Every person who knowingly sells or buys securities, even indirectly, using confidential information that he or she holds as a shareholder of the issuer of the securities in question is guilty of an indictable offence and liable to imprisonment for a term not exceeding ten years or is guilty of an offence punishable on summary conviction;*"

- **398**: "*Falsifying record of employment. Every person who, with intent to mislead, falsifies a record of employment by any means, including* *the punching of a time clock, is guilty of a summary conviction offence.*"

- **402.2**: "*Identity theft. Every person commits an offense who obtains or has in his or her possession identifying information about another person with the intent to use that information to commit an indictable offence, one of the elements of which is fraud, deceit or falsehood.*"
- **406 a)**: "*Infringement of Trade-mark. For the purposes of this Part, a person who, without the consent of the owner of the trade-mark, makes or reproduces in any manner that trade-mark or a mark so nearly resembling it as to be likely to mislead;*"
- **431**: "*Attack on the official premises, private dwelling or means of transport of an internationally protected person. Any person who makes an attack accompanied by violence on the official premises, private dwelling or means of transportation of an internationally protected person in such a manner as to be likely to endanger the life or liberty of that person shall be guilty of an indictable offence punishable by imprisonment for a term not exceeding fourteen years.*"
- **432 (1)**: "*Unauthorized recording of a motion picture. Whoever, without the consent of the manager of the cinema, records a cinematographic work - as that term is defined in section 2 of the Copyright Act - that is shown in a cinema, or its soundtrack, is guilty of:*"
- **437**: "*False alarm. Any person who willfully, without reasonable cause, by shouting, ringing bells, using a fire alarm, telephone or telegraph, or in any other manner, sounds or spreads or causes to be sounded or spread a fire alarm, is guilty.*"
- **438 (2)**: "*Obstructing salvage of wreck. Every person who wilfully prevents or hinders, or wilfully seeks to prevent or hinder, the salvage of a wreck is guilty of an offence punishable on summary conviction.*"
- **439**: "*Disturbance of marine signals. Every person who moors a ship or boat to a signal, buoy or other landmark used for navigation is guilty of an offence punishable on summary conviction.*"
- **467.111**: "*Recruitment of members by criminal organization. Whoever recruits a person to be a member of a criminal organization-or invites, encourages, coerces, or solicits a person to be a member of a criminal organization-for the purpose of increasing the ability of the organization to facilitate or commit a criminal act under this or any other federal law is guilty of an indictable offense and liable:*"