Most NLG is Low-Resource: here's what we can do about it

David M. Howcroft and **Dimitra Gkatzia**
School of Computing
Edinburgh Napier University
Edinburgh, Scotland, United Kingdom
{d.howcroft,d.gkatzia}@napier.ac.uk

Abstract

Many domains and tasks in natural language generation (NLG) are inherently 'low-resource', where training data, tools and linguistic analyses are scarce. This poses a particular challenge to researchers and system developers in the era of machine-learning-driven NLG. In this position paper, we initially present the challenges researchers & developers often encounter when dealing with low-resource settings in NLG. We then argue that it is unsustainable to collect large aligned datasets or build large language models from scratch for every possible domain due to cost, labour, and time constraints, so researching and developing methods and resources for low-resource settings is vital. We then discuss current approaches to low-resource NLG, followed by proposed solutions and promising avenues for future work in NLG for low-resource settings.

1 Introduction

Natural Language Generation (NLG) is the process of generating text from structured or unstructured data and has recently received renewed attention due to the emergence of large pre-trained language models (e.g. Brown et al., 2020) that promise to generate output that is more natural, variable, and adaptable to new domains as compared to rule-based approaches. However, the development of robust, *controllable*, and usable NLG systems depends heavily on the availability of large, high-quality, labelled datasets that are appropriate for the task at hand (Fan and Gardent, 2020). Unfortunately, for most domains such data is unavailable, probably with the exception of weather, restaurant, and sports domains. Even in the aforementioned domains, data is fairly small compared to low-resource tasks & languages in other areas, such as in machine translation.

In this paper, we focus on *controllable* NLG tasks that can be framed as data-to-text generation tasks, rather than language prediction models. We

Language			Domain			Illustrative examples of
C	A	T	C	A	T	task or data availability
-	-	-	-	-	-	most languages, any domain
-	-	-	-	+	-	domains with linguistic analyses
-	+	-	-	-	-	well-studied minority languages
-	+	+	-	-	-	lg.s not very present online that are well-studied and have, e.g., parsers, etc, *but few* domain-specific resources
-	+	+	+	+	+	lg.s not very present online that are well-studied & have parsers, etc, *and* substantial domain-specific resources
+	+	+	-	-	-	novel NLG domains in English
+	+	+	-	+	-	domains with lots of analysis in a high-resource language but little specialised data or tooling (e.g. political rhetoric)
+	+	+	+	+	+	Restaurant reviews, weather forecasts, and sports reporting in English

Table 1: Dimensions of resource availability at the language & domain level, sketching variation with respect to C(orpora), A(nalysis), & T(ool) availability. Plus/minus indicate relatively more/less availability, respectively. Examples demonstrate the logical possibilities associated with different combinations of language and domain resource availability.

center our discussion around the *low-resource* nature of controllable NLG, that is, on the limitations to the creation of new NLG systems due to a lack of **corpora**, **analyses**, or **tools** for a target domain or language (illustrated in Table 1). Here, we define as *corpora* paired input representations and texts which serve as reference outputs. *Analyses* are linguistic analyses relevant to a domain of application, including but not limited to grammars of a target language or conversation analyses of a target domain. *Tools* can then be automated means of analysing linguistic data (e.g. parsers, part-of-speech taggers, etc.), secondary resources based on primary data (e.g. word embeddings), or software libraries for different NLG (sub)tasks.

This position paper contributes:

- a discussion of the availability of NLG resources in terms of Corpora, Analyses, & Tools for different languages and domains and the challenges this presents;

- a high-level overview of mitigation strategies for addressing low resource availability; and

- a call to focus future work on particularly promising directions which we highlight.

2 Corpora for NLG

Corpora are essential to the development of both data-driven and knowledge-driven NLG systems. We focus here on NLG corpora for generating full utterances of at least one complete sentence which include input *meaning representations* (MRs, ranging from raw sensor data to morphologically specified syntax trees)[1] and corresponding texts, in situations where existing resources are limited. Currently, the most studied language for NLG is English (cf. Fan and Gardent, 2020) with few data-to-text corpora available for other languages. The domains with the most data-to-text corpora available are restaurant descriptions, weather forecasts, and sports reporting[2]. Corpora for the widely studied restaurant domain range in size from 400 utterances (Mairesse et al., 2010) to 50k utterances (Dušek et al., 2020), some being based on hand-crafted systems while the majority are crowdsourced. We now highlight prominent strategies for building NLG corpora; however, even with a simple MR format (e.g. CUED slot value pairs (Young, 2007)), collecting high-quality parallel MR-text corpora is expensive so most such datasets remain small.

'Found' NLG Corpora Sometimes data-to-text corpora can be adapted from existing sources of semantic and textual data. For instance, Belz and Kow (2010) assembled a parallel corpus of sets of facts (from hobbyists) and corresponding texts (from Wikipedia) about British hills. Similarly, the WikiBio dataset (Lebret et al., 2016) pairs the first sentence of each article in the WikiProject Biography dataset with the facts reported in that article's 'infobox'. GenWiki extends these approaches even further, aiming to provide an automatically aligned corpus of texts from Wikipedia paired with graphs

from DBPedia identified based on overlapping entities for more than one million texts (Jin et al., 2020). Apart from Wikipedia, researchers have collected datasets from other online resources that contain both data (in metadata) and (somewhat) aligned text (e.g. Liang et al., 2009; Barzilay and Lapata, 2005). Note that these approaches rely on the fact that others, such as domain experts, have already chosen to create a semantic or tabular representation of important details; therefore, this method of building new datasets is not generalisable.

Creating meaning annotations Some research has annotated existing data-to-text corpora with discourse structures, such as Balakrishnan et al. (2019)—who semi-automatically added discourse structures to the E2E Challenge corpus—and Stevens-Guille et al. (2020)—who leveraged the rule-based Methodius system (Isard, 2016) to create a discourse-annotated corpus. Based on the automatically derived Methodius Corpus, Maskharashvili et al. (2021) observe that "discourse relations are enormously helpful when the dataset for the domain is limited", highlighting the importance of fine-grained MRs in low-resource domains.

Other work has sought to address the issue of content selection for corpus creation, independent of the actual text to be associated with each MR (see also Gkatzia, 2016). For example, Perez-Beltrachini et al. (2016) leverage DBPedia to construct trees of semantic triples based on their frequency and relationship to one another in a large ontology, with the goal of selecting content which forms a natural unit that can be later associated with a human-written text.

Eliciting texts for given meanings Early datasets typically relied on domain experts or NLG researchers directly, but most recent work uses crowdsourcing to quickly collect texts from a variety of speakers (e.g. Mairesse et al., 2010; Wen et al., 2015, 2016; Juraska et al., 2019). Where early work tended to use prompts similar to a set of slot-value pairs, later work observed that such prompts encouraged the use of particular words & phrases—thus reducing textual diversity—and found that using images (Novikova et al., 2016) or full sentences (Howcroft et al., 2017) as prompts resulted in better quality. Recent work has incorporated quality estimates and text suggestions to fill gaps in datasets (Chang et al., 2020). While these methods can quickly provide varied data, it remains

---

[1]While we would like to see more emphasis on representations based on semantic, linguistic, or logical representations of meaning and discourse structure, such as Montagovian semantics, HLDS (Kruijff, 2001), DRS (Kamp and Reyle, 1993), etc., common data-to-text MRs more commonly resemble tabular data, taking the form of slot-value pairs plus an optional dialogue act annotation or RDF triples. For simplicity, we refer to the input to a data-to-text NLG system as an MR, regardless of the degree to which the encoding is developed as a representation of *meaning* per se.

[2]https://aclweb.org/aclwiki/Data_sets_for_NLG

impractical to create large, targeted datasets for all domains, and most languages do not even have these small datasets available.

3 Analyses & Tools

While the primary data for NLG are parallel corpora, a variety of other resources can facilitate development. Language models (LMs) can approximate fluency measures and can be used to produce texts using sampling (Brown et al., 2020). Linguistic analyses provide insight into what makes a text well-formed and assist in the design of rule-based systems and architectures for ML-based approaches. Finally, tools & resources such as parsers, part-of-speech taggers, semantic role labellers, ontologies, morphological analysers, and word embeddings can help to decompose NLG subtasks and make the problem more approachable.

3.1 (Large) Language Models

Statistical language models have been used since the late 1990s to help rank potential NLG outputs (Langkilde and Knight, 1998; Knight and Hatzivassiloglou, 1995). With the rise of large pre-trained language models (e.g. BERT, GPT: Devlin, 2018; Brown et al., 2020), there has been renewed interest in sampling-based approaches to generation, where an LM trained exclusively on text (i.e. without parallel MRs) generates a continuation for an initial string. While the challenge of making sampling-based NLG more *controlled* is an active area of research[3], these tools continue to be helpful to rank texts with likelihood as a fluency approximation.

Unfortunately, good language models require large collections of text in the target language in order to perform well and only very limited digitized data is available for most languages. For example, Joshi et al. (2020) found that only seven languages (of the world's approximately 7000 languages) qualify as truly high-resource[4]. One major factor in corpus availability is the cost of technology relative to typical incomes in countries where a language is spoken (Ahia et al., 2021). Given the high cost of collecting & annotating data, this observation is not surprising. Recent approaches to text generation have emphasised the use of large-pretrained LMs (see Section 3.1) and task-specific fine-tuning in order to transfer general language

statistics to a particular task. Consider for instance GPT3 (Brown et al., 2020), trained on 45TB of text data in English. Training such a model for any language requires an amount of data and compute power unavailable in most regions, meaning that such models can typically be used only for low-resource domains in high-resource languages.

Candidate languages include the roughly 100 languages covered by mBERT (Devlin, 2018) and mT5 (Xue et al., 2021). More often, though, available language models for lower resource languages are often based on legal, journalistic, or governmental language rather than everyday language, resulting in a genre mismatch making them poor off-the-shelf models for many applications. For those cases where appropriate LMs do exist, Section 4.2 covers their usefulness in transfer learning.

3.2 Analysis

With enough data, we can hope that a powerful ML architecture might detect the patterns necessary to produce good texts. However, when corpora are not large enough for this, descriptive and automated linguistic analyses can help. Researchers can use linguistic documentation to develop their system, consulting formal grammars & lexica to understand the kinds of constructions possible in a target language. Researchers & developers can also partner with speakers of their target language to ensure that the system serves community needs while working together to understand the language they are generating (Hirmer et al., 2021).

Generally speaking, it is easiest to leverage these linguistic resources when developing a rule-based NLG system, where observations can be encoded explicitly. For example, three grammars of Rapanui have been published in English in the past 110 years (Churchill, 1912; Du Feu, 2012; Kieviet, 2017), giving insight into word order, morphology, and agreement phenomena.[5] Such features can then be input in grammar engineering tools like the Lingo matrix grammar construction toolkit[6] (Bender et al., 2002, 2010), to provide a starting point for building a rule-based NLG system.

Both rule-based and end-to-end ML approaches benefit from tools for automated linguistic analy-

---

[3]e.g. `https://ctrlgenworkshop.github.io/`

[4]English, Spanish, German, Japanese, French, Arabic, Mandarin

[5]We use Rapanui as an example of a very low-resource language which, nevertheless, has been the subject of multiple grammars published as monographs. For more information, see the cited grammars or `https://en.wikipedia.org/wiki/Rapa_Nui_language`.

[6]`https://matrix.ling.washington.edu`

sis, such as lemmatisers, part-of-speech taggers, parsers, and semantic role labellers. To learn to map input MRs to output texts, a system or developer must recognise useful generalisations and abstractions. For example, part-of-speech tags and constituency parses provide information about the order in which words of a given class should appear, while dependency parses and semantic role labels can relate individual words to each other. Similarly, normalization diminishes the impact of noise, lemmatization helps associate word 'stems' with meanings, morphological analysers/realisers help with word forms, parsers help get words in the right order, semantic role labellers & natural language understanding systems help associate chunks of form & meaning directly.

All of these individual analyses together help decompose the task. Although data-to-text systems have used modular architectures for decades (Reiter, 2007), there is now a shift towards end-to-end architectures. However, we argue here that modular architectures are important for low-resource settings, as they might require less training data than end-to-end. This is an important promising direction, therefore we discuss pipelines and problem structuring further in Section 4.5.

3.3 NLG Libraries

Similarly to general NLP tools, NLG-specific tooling shows variable availability. While SimpleNLG has been adapted to at least eight other languages[7] (English: Gatt and Reiter, 2009), this library still requires significant effort to develop supporting components. Other available tools for developing NLG systems include FUF/SURGE[8] (Elhadad and Robin, 1997, 1996) for surface realisation and OpenCCG[9] (White, 2006) for generation from hybrid logic dependency semantics, which can be used to represent meaning at deeper and shallower levels alike, using combinatory categorial grammar (Steedman and Baldridge, 2006). For a recent survey of surface realisation systems for low-resource languages, see (Mahlaza and Keet, 2022).

A number of neural NLG models are also publicly available, such as Wen et al.'s (2015) SC-LSTM[10] and Dusek & Jurcicek's (2016) TGen[11].

[7]https://github.com/simplenlg/simplenlg
[8]https://www.cs.bgu.ac.il/~elhadad/install-fuf.html
[9]https://github.com/OpenCCG/openccg
[10]https://github.com/shawnwun/RNNLG
[11]https://github.com/UFAL-DSG/tgen

While TGen has been widely used as a baseline for end-to-end NLG tasks, these systems generally represent the outcome of a particular research project rather than being intended to be used as a platform for developing future NLG systems.

4 Mitigation strategies

So far we have discussed the corpora, analyses, & tools that are often missing in low-resource settings. We now turn to mitigation strategies for dealing with this lack of resources, namely data augmentation, transfer learning, meta-learning, feedback-based learning, and rule-based methods. Table 2 gives an overview of the requirements for and outcomes of using these different mitigation strategies.

4.1 Data Augmentation

Data augmentation (DA) describes a family of approaches that aim to increase the number of training examples automatically, without manual data collection. Despite their popularity and demonstrated efficiency in other areas such as computer vision (Shorten and Khoshgoftaar, 2019) and NLP (Dhole et al., 2021), this area is still relatively under-explored in NLG, partly due to unique challenges. This section reviews existing approaches to DA for text generation; for DA approaches to NLP in general we refer the reader to the survey by Feng et al. (2021).

DA methods promise to enrich current datasets, potentially increasing the linguistic diversity of the dataset (e.g., by enhancing stylistic traits as proposed by Oraby et al. (2018)). In order to train NLG models, we require text that is not only grammatically correct but also semantically correct, coherent, and appropriate for the task at hand (cf. Dušek et al., 2019, on semantic noise). Therefore, straightforward approaches used in computer vision, such as cropping and rotation are not appropriate, though adaptations of these techniques can help with simpler NLP tasks like part of speech tagging for low-resource languages (Şahin and Steedman, 2018). More elaborate approaches such as back-translation & paraphrasing have been used in other areas of NLP, such as MT, and are also promising for NLG. Here, however, we only review methods specifically developed and applied in **NLG settings**.

There are two predominant DA methods used in NLG: (1) generation of new examples with pretrained LMs; and (2) generation of new examples

Mitigation Strategy	Requirements	Outcome
Data Augmentation	labelled data, rules	additional labelled data
Pre-training/Fine-tuning	existing LM or high-resource data, low-resource data	knowledge transfer to new domain
Zero-shot	existing LM	adapting to new scenarios with no labelled data
Few-shot	existing LM, few training examples	adapting to new scenarios with few labelled data
Prompt-based	existing LM, prompts	adapting to new scenarios with few or no labelled data
Meta-learning	high-resource data, low-resource dataset	weight initialisation for better training
Feedback-based	small domain dataset	better learning through interaction
Multi-task	multiple auxiliary tasks	learning more general representations, multiple tasks

Table 2: Summary of mitigation strategies for low-resource NLG, their requirements, and their outcomes.

with statistical or rule-based NLG systems.

An example of the former DA methods is Chang et al. (2021a)'s approach to generating new samples for NLG using a pretrained LM, by firstly creating an unannotated dataset with unlabeled MR instances by randomly selecting MRs from a pre-existing expert-annotated dataset and populating them with existing values. This dataset is then automatically annotated with noisy text labels generated by a pre-trained model, fine-tuned on joint MR and text conditioned on samples from the augmented MR set.

Examples of the latter include employing rule-based systems to generate new examples for developing or evaluating other NLG methods. For instance, Belz (2008) created an automatically generated version of SUMTIME meteo (Sripada et al., 2008), an expert annotated data-to-text dataset in the weather domain, which has been used to explore statistical NLG (Angeli et al., 2010, inter alia). Similarly, Oraby et al. (2018) utilised a statistical natural language generator to create a corpus of stylistic texts used to train seq2seq neural NLG models. Other apporaches have proposed utilising distant supervision can also be used to create new NLG corpora. For example, Agarwal et al. (2021) use distant supervision to verbalize knowledge graph subgraphs centered on entities, producing a large secondary dataset, a subset of which they were able to clean to a sufficiently high quality for fine-tuning their models.

4.2 Transfer Learning

Transfer learning uses the knowledge gained from a previous task (in a high-resource setting) to improve model performance for a related task in a lower-resource setting (Torrey and Shavlik, 2009). Typically, a model is trained with data from one or more high-resource domains/languages and then the model's weights are used to initialise the model to be trained in the low-resource setting through a

process called *fine-tuning*. Similarly, *few-shot* and *zero-shot* approaches aim to develop general language models which are applied to new tasks with limited intervention (Zhao and Eskenazi, 2018).

Pre-training & Fine-tuning   Transfer learning via *fine-tuning* typically requires the adaptation of a large pre-trained language model by updating and storing all of the parameters, resulting in a new language model for every task. One of the earliest works in this setting involved training a model from scratch in a related domain and then fine-tuning it in a new domain (Dethlefs, 2017). Kale and Roy (2020) propose transferring knowledge from a NMT model trained on English-Czech, which is then fine tuned for a data-to-text task in Czech. Pasricha et al. (2020) proposed a transfer learning approach which actually utilises one of the large language models (see Section 3.1) which is fine-tuned in the target task. In this setting, the data used for fine tuning is pre-pended with tags describing the part of speech as well as the type of the entity and are included in the vocabulary. Ribeiro et al. (2021) also show that pre-trained language models perform well in graph-based MR to text generation, even when the input is reduced to bags of nodes and edge labels. Most works in this area generate text in English. However, fine-tuning large models also works in other languages. For instance, Naous et al. (2021) propose fine-tuning AraBERT (Antoun et al., 2020) for empathetic NLG in Arabic.

Although fine-tuning requires significantly less computational power and time as compared to training models from scratch, it can still pose a considerable deployment challenge as the magnitude of pre-trained models continues to increase from millions to billions of parameters. As such, other data- and compute-efficient transfer learning approaches have been explored that try to minimize the number of parameters that are fine-tuned. Such methods include prompt-based, few- and zero-shot learning

approaches which are discussed next.

Prompt-based, Few- and Zero-shot learning
To alleviate the need to update all parameters of
a pre-trained model, researchers have explored
prompt-based methods that keep the parameters
of a model frozen and instead use *prompts* as part
of the input (Liu et al., 2021) to perform down-
stream tasks without further training. Although no
training is required, prompt-based methods require:
1) a prompting function that converts the input into
some specific form; 2) template prompts, which
can be created manually or automatically; 3) corre-
sponding filled & answer prompts; and 4) answers.
In many cases, prompt methods do not require any
further training and providing the aforementioned
prompting elements is enough for a model to per-
form *zero-shot* learning (as in Dou and Peng, 2022).
In other settings, though, prompting can be used for
further training/fine-tuning a model, when at least
a small amount of data is available. For instance,
Li and Liang (2021) proposed an approach to NLG
that keeps pre-trained LM parameters fixed, while
employing a task-specific prefix vector, which is
tuned for the task at hand. Clive et al. (2021) ex-
tend this approach for controlled text generation.
Prompt-learning is a very recent paradigm, so we
expect to see more work in this promising area.

Similarly to prompting, zero- and few-shot learn-
ing aim to achieve learning with minimal train-
ing/new data instances. Transfer learning aims
to 'learn' transferable features that can be used
in downstream tasks. In few-shot learning, there
might be only a few examples to learn from (or zero
in zero-shot). Ma et al. (2019) proposed decompos-
ing table-to-text generation into content selection
and surface realisation, so that each subtask can
be trained with a smaller dataset than it would be
needed for the end-to-end task. Chen et al. (2020)
also proposed pre-training a model from scratch,
although their paradigm employed distant supervi-
sion before fine-tuning the model to specific tasks.
Finally, Chang et al. (2021b) focused on improving
few-shot NLG by prioritising informative training
instances to fine-tune the model.

4.3 Meta-learning

Meta learning can be thought of as the machine
learning process of 'learning how to learn' (Mi
et al., 2019a). Meta-learning approaches are
split into *metric-learning* (Koch et al., 2015),
model-based (Andrychowicz et al., 2016), and

optimisation-based approaches (Finn et al., 2017).
Most of recent meta-learning approaches to NLG
follow an optimisation approach, the Model-
Agnostic Meta-Learning (MAML) algorithm, orig-
inally proposed by Finn et al. (2017). MAML
aims to make models achieve good generalisation
performance by adapting quickly to a new task
during training in low-resource settings, despite
a low quantity of training data, by learning a bet-
ter initialization of model parameters that facili-
tates fast adaptation to new low-resource scenarios.
Mi et al. (2019b) explored different adaptation set-
tings based on domain similarity and showed that
'nearer' domains can adapt better through a meta-
learning setting, outperforming other optimisation
methods such as multi-task learning. Meta-learning
has also shown promising results in MT. Gu et al.
(2018) compared MAML to transfer learning (Zoph
et al., 2016) and showed that meta-learning leads
to further improvements, despite training data for
the low-resource language being limited to a sig-
nificantly smaller dataset. As the corpus size of the
low-resource language decreases, transfer learn-
ing approaches suffer significantly more than meta-
learning approaches, demonstrating the effective-
ness of MAML for low-resource languages. How-
ever, as corpus size increases, the differences be-
tween the two approaches are much less signifi-
cant. Exploring the trade-off between data size and
learning paradigm (meta-learning versus transfer
learning) is a promising direction for this line of
work.

4.4 Feedback-based Learning

Low-resource settings have always been a bottle-
neck for NLG. Earlier work in this area employed
reinforcement learning (RL) in dialogue systems
for NLG (Rieser and Lemon, 2011; Dethlefs and
Cuayáhuitl, 2011) to overcome the issue of limited
data, whereas more recent approaches showed that
RL can help train better NLG systems (Panagiaris
et al., 2021). Low-resource settings can manifest
themselves also in domains where large collections
of unlabelled data are available without parallel
inputs. For instance, in MT one might have access
to large monolingual datasets in both source and
target languages but not an aligned one. In this
case, feedback-based methods can help, such as
the dual-learning setup from Kim et al. (2019), pre-
sented as a two-agent communication game. In this
setting, the first agent only understands language A,

and the second agent only understand language B. The two agents communicate through translation models and provide feedback on whether the translated message they received is a natural sentence in their own language. They then use this feedback to improve their individual models.

Similarly, Shen et al. (2019) propose treating language production as a game between speakers and listeners, where listeners must be able to reconstruct the intended meaning. Their models are trained to predict and avoid confusing outputs based on either reconstruction or distraction pragmatics (this is however not low-resource). Tran and Nguyen (2018) propose an adversarial training procedure for domain adaptation with two critics, which guide the generator to generate outputs similar to the sentences in the target domain, when limited amount of target domain data exist.

4.5 Task structure & rule-based approaches

While 'fully end-to-end' machine learning models are always enticing, careful thought about how to structure the task(s) can lead to significant improvements in outcomes. For example, in *multitask learning* a single model is trained on multiple tasks, allowing feedback from learning to perform well on one task to influence the others. Even without jointly learning to solve multiple tasks, decomposing generation into a sequence of stages in a pipeline can improve performance by simplifying what the model needs to learn.

Multi-task learning To our knowledge, there is no related work in multi-task learning for low-resource data-to-text generation. The closest work in this area jointly learns a semantically conditioned and unconditioned LMs for generation across multiple datasets (Zhu et al., 2019). For text-to-text generation, Magooda et al. (2021) showed that abstractive summarisation can benefit from being learnt in a multi-task framework, especially when combined with paraphrase learning. In addition to their target task, they train their model to perform extractive summarisation, concept & paraphrase detection, and language modelling. In a larger resource setting, Agarwal et al. (2020) adapt T5 (Raffel et al., 2020) to data-to-text generation for English and Russian while jointly learning text-to-data semantic parsing for both languages.

Pipelines & Problem Structuring Where in multitask learning the focus is on training a single neural network to perform multiple tasks, pipeline
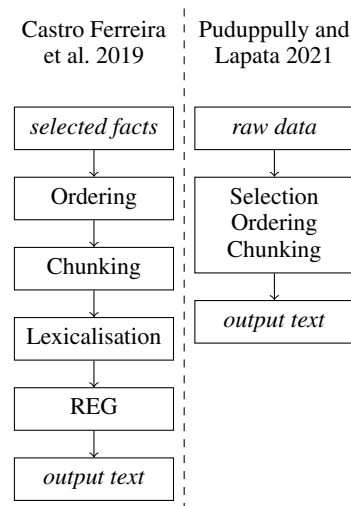

Figure 1: Two neural pipelines: Castro Ferreira et al. (2019) use four stages to produce a text based on chosen facts; Puduppully and Lapata (2021) use two stages to select & group facts and realise the text.

models use separate stages for each subproblem and do not require a single unified neural network (cf. Figure 1). This allows for specialisation, where solutions to subproblems can be refined independently. For example, a surface realisation module could be trained on multiple domains regardless of the handling required at the level of document or sentence planning. Of course, modularity comes with the risk of error propagation, as errors in an early component can break later ones.

Recent work has explored various decompositions of data-to-text generation based on the conventional NLG pipeline. For example, Howcroft et al. (2018) proposed a model to learn sentence planning rules for use with an off-the-shelf surface realiser while Moryossef et al. (2019) reversed this focus, using rule-based planning to handle content ordering & chunking and training a surface realiser.

Figure 1 sketches two fully neural NLG pipelines as examples. Castro Ferreira et al. (2019) decomposed generation into content ordering, content chunking, lexicalisation, referring expression generation (REG), & textual realisation and trained a neural model for each subtask, finding that pipelines improve text quality. Both Perez-Beltrachini and Lapata (2018) and Puduppully and Lapata (2021) created higher level pipelines which perform content selection (with the latter also performing content ordering and sentence chunking) before generating a text. Other work has used latent variable models to learn ordering and chunking

constraints instead of focusing on explicit pipelines (Shen et al., 2020; Xu et al., 2021).

These models generally preserve the 'fluency' expected of modern neural network LMs and improve semantic fidelity. This is especially encouraging for low resource NLG where the relative pay-off for adding intermediate annotations to train pipeline models is likely to be higher. Such approaches are further enhanced by 'delexicalisation', e.g. replacing certain values with placeholders (Shimorina and Gardent, 2019) to limit spurious variation and facilitate training.

For low-resource domains in better resourced languages, useful resources for AMR-to-text generation[12] and UD-to-text[13] generation (Mille et al., 2018, 2019, 2020) could be used for the final stage of generation, simplifying system development and allowing researchers to focus on developing components for document and sentence planning.

Complementary to these efforts are a number of new corpora (van der Lee et al., 2020) and extensions of existing corpora (Castro Ferreira et al., 2020, 2021) with annotations for intermediate representations in an NLG pipeline.

5 Discussion & Promising directions

So far, we have highlighted the ways in which NLG is influenced by the availability of data, analyses, and tools at both the language and the domain level and described a number of trainable approaches which aim to learn an NLG system given different resource limitations. It is clear that it is neither sustainable to collect large aligned datasets for every domain nor use general purpose language models off-the-shelf for *controlled* NLG, so researching and developing approaches for low-resource settings that distill the strengths of pre-trained language models while focusing on controllable approaches is vital. Here we outline some promising directions in this area.

Data Augmentation Although it is clear that controllable NLG approaches could benefit from as large corpora as possible, data augmentation is still under-explored for NLG. Approaches developed for other NLP tasks could be explored here whereas others will not be applicable for structured prediction problems. The potential future directions can be split into three main categories: (a)

Data augmentation through paraphrasing where a slot is replaced with another slot randomly: e.g. 'cheapest' is replaced with 'most expensive'; (b) using pre-trained models to distill knowledge; (c) back-translation to create resources in different languages.

Prompt-based Learning Prompt-based learning is a fairly novel direction for NLP (including NLG) in general and our understanding is quite limited to a few experimental works. Work on this area can highlight limitations of current pre-trained models and lead to potential improvements as well as can reveal situations where large models are safe (and might not need to be controllable) to use. Prompt-based approaches might also call for novel evaluation metrics in order to better understand the influence of prompts on the generated outputs.

Feedback-based approaches Although there has been a lot of discussion in AI in general about machines learning through feedback as humans can, research focused on increasing a NLG system's capabilities have been limited to game-based settings or simulation (as in the case of Reinforcement Learning). More research in this area, might be useful to endow models with new capabilities as in the frameworks proposed by Gkatzia and Belvedere (2021) and Wang et al. (2016).

Multitask learning & pipelines Very few of the combinatorial possibilities for neural NLG pipelines have been explored to date, so it remains unclear which tasks are best solved in sequence versus jointly. One especially promising direction is to approach such pipelines in a multitask setting: when annotated data exists for subtasks, the model can receive feedback for individual tasks during training while passing along the penultimate layers of the network to each subsequent task, thus allowing later tasks in the pipeline to influence the hidden representations learnt in earlier tasks.

When to use meta-learning and when to use transfer-learning? There is some evidence from MT that transferring models between related languages increases performance Zoph et al. (2016). On the other hand, data size matters (Kocmi and Bojar, 2018). Previous work has also shown that once you start increasing the amount of data in the target domain, transfer learning achieves better results. However, it is unclear where the 'sweet spot' lies, and more research in this area is required.

---

[12]https://nert-nlp.github.io/AMR-Bibliography/
[13]Universal Dependencies (de Marneffe et al., 2021)

Evaluation in low-resource NLG   In addition to the specific challenges and mitigation strategies for system development above, evaluation has its own challenges in the low-resource setting and is a promising direction for future work in itself. For instance, having less validation and test data reduces the applicability of automated, reference-based evaluations, necessitating alternative evaluation strategies such as an emphasis on error analysis (van Miltenburg et al., 2021) or standardised human evaluations (Howcroft et al., 2020). Methods for maximising the efficiency of input from domain and language experts will also be necessary for human evaluations when access to these persons is more limited than usual.

6 Conclusion

In this paper we have highlighted ways in which NLG is often low-resource in data, analyses, or tools with respect to either the target language or domain. The number of corpora for data-to-text generation remain limited, but there are a number of strategies for adding meaning annotations to existing texts, or vice versa, and crowdsourcing new data sets. Large language models are also not always available: they are typically only available for relatively high-resource languages. Linguistic analyses and automated analysis tools are more widely available and can help decompose the generation task to make it easier to develop systems in low-resource settings. Fortunately there are a variety of mitigation strategies available, including methods for data augmentation, transfer learning, meta- and feedback-based-learning, and different ways of structuring the task.

This makes for an interesting set of challenges and an exciting set of opportunities to improve the state-of-the-art for natural language generation in low resource settings. Our own current work is beginning to explore some of these promising directions, including multi-task learning, pipeline approaches to NLG, and transfer learning from related languages and general purpose LLMs. However, there is much work to be done, and we hope the community embraces these challenges.

Ethics Statement

One goal of our paper is explicitly to encourage further research in low-resource settings for natural language generation. While we have argued that many 'typical' NLG settings can be thought of

as relatively low-resource compared to other areas of NLP, it is worth noting that some topics (e.g., legal, medical) are especially sensitive and that working with low-resource languages in particular introduces challenges in privacy (for an extensive discussion on this topic see (Hirmer et al., 2021)). For work such as developing mental health support chatbots or developing systems for minoritised language communities, we especially encourage our fellow researchers to engage with their local ethics boards and adopt a participatory approach to data collection and system design to ensure that their efforts work in collaboration with the affected communities.

Limitations

While this position paper includes an overview of the current literature, a systematic survey (e.g. following PRISMA, Moher et al., 2009) of all work in low-resource NLG is beyond the scope of the paper: the relative novelty of the topic makes it difficult to determine appropriate selection criteria for a systematic survey on this topic. Usually a systematic survey would include papers based on keyword searches in academic databases, but almost no papers explicitly focus on natural language generation in low-resource settings, making it difficult to identify phrases which reliably indicate all and only the relevant works. This limits the coverage of the current paper, though we believe this limitation is a reasonable trade-off when highlighting an area requiring more attention in future work.

Acknowledgements

References

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565, Online. Association for Computational Linguistics.

Oshin Agarwal, Mihir Kale, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2020. Machine translation aided bilingual data-to-text generation and semantic parsing. In *Proceedings of the 3rd International*

Workshop on Natural Language Generation from the Semantic Web (WebNLG+), pages 125–130, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marcin Andrychowicz, Misha Denil, Sergio Gómez Colmenarejo, Matthew W. Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. 2016. Learning to learn by gradient descent by gradient descent. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NeurIPS'16, page 3988–3996, Red Hook, NY, USA. Curran Associates Inc.

Gabor Angeli, Percy Liang, and Dan Klein. 2010. A simple domain-independent probabilistic approach to generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 502–512, Cambridge, MA. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

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.

Regina Barzilay and Mirella Lapata. 2005. Collective content selection for concept-to-text generation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 331–338, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Anja Belz. 2008. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, 14(4):431–455.

Anja Belz and Eric Kow. 2010. Extracting parallel fragments from comparable corpora for data-to-text generation. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar Customization. *Research on Language and Computation*, 8(1):23–72.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starterkit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *COLING-02: Grammar Engineering and Evaluation*.

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. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. 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 552–562, Hong Kong, China. Association for Computational Linguistics.

Thiago Castro Ferreira, Helena Vaz, Brian Davis, and Adriana Pagano. 2021. Enriching the E2E dataset. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 177–183, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Ernie Chang, Jeriah Caplinger, Alex Marin, Xiaoyu Shen, and Vera Demberg. 2020. DART: A lightweight quality-suggestive data-to-text annotation tool. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 12–17, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Ernie Chang, Vera Demberg, and Alex Marin. 2021a. Jointly improving language understanding and generation with quality-weighted weak supervision of

automatic labeling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 818–829, Online. Association for Computational Linguistics.

Ernie Chang, Xiaoyu Shen, Hui-Syuan Yeh, and Vera Demberg. 2021b. On training instance selection for few-shot neural text 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 2: Short Papers)*, pages 8–13, Online. Association for Computational Linguistics.

Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2020. Few-shot NLG with pre-trained language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Online. Association for Computational Linguistics.

William Churchill. 1912. *Easter Island: The Rapanui Speech and the Peopling of Southeast Polynesia.* Carnegie Institution, Washington.

Jordan Clive, Kris Cao, and Marek Rei. 2021. Control prefixes for text generation. *CoRR*, abs/2110.08329.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Nina Dethlefs. 2017. Domain transfer for deep natural language generation from abstract meaning representations. *IEEE Computational Intelligence Magazine*, 12(3):18–28.

Nina Dethlefs and Heriberto Cuayáhuitl. 2011. Hierarchical reinforcement learning and hidden Markov models for task-oriented natural language generation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 654–659, Portland, Oregon, USA. Association for Computational Linguistics.

Jacob Devlin. 2018. Multilingual BERT README.

Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard H. Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Rishabh Gupta, Rishabh Gupta, Louanes

Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honore, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, and et al. 2021. Nl-augmenter: A framework for task-sensitive natural language augmentation. *CoRR*, abs/2112.02721.

Zi-Yi Dou and Nanyun Peng. 2022. Zero-shot commonsense question answering with cloze translation and consistency optimization. In *The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*.

Veronica. Du Feu. 2012. *Rapanui: A Descriptive Grammar.* Taylor and Francis, Hoboken.

Ondřej Dušek, David M. Howcroft, and Verena Rieser. 2019. Semantic noise matters for neural natural language generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 421–426, Tokyo, Japan. Association for Computational Linguistics.

Ondřej Dušek and Filip Jurčíček. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51, Berlin, Germany. 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.

Michael Elhadad and Jacques Robin. 1996. An overview of SURGE: a reusable comprehensive syntactic realization component. In *Eighth International Natural Language Generation Workshop (Posters and Demonstrations)*.

Michael Elhadad and Jacques Robin. 1997. Surge: A comprehensive plug-in syntactic realisation component for text generation.

Angela Fan and Claire Gardent. 2020. Multilingual AMR-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2889–2901, Online. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. *CoRR*, abs/1703.03400.

Albert Gatt and Ehud Reiter. 2009. SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93,

Athens, Greece. Association for Computational Linguistics.

Dimitra Gkatzia. 2016. Content selection in data-to-text systems: A survey. *CoRR*, abs/1610.08375.

Dimitra Gkatzia and Francesco Belvedere. 2021. *"What's This?" Comparing Active Learning Strategies for Concept Acquisition in HRI*, page 205–209. Association for Computing Machinery, New York, NY, USA.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Stephanie Hirmer, Alycia Leonard, Josephine Tumwesige, and Costanza Conforti. 2021. Building representative corpora from illiterate communities: A reviewof challenges and mitigation strategies for developing countries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2176–2189, Online. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

David M. Howcroft, Dietrich Klakow, and Vera Demberg. 2017. The Extended SPaRKy Restaurant Corpus: Designing a Corpus with Variable Information Density. In *Proc. of Interspeech 2017*, pages 3757–3761, Stockholm, Sweden. ISCA.

David M. Howcroft, Dietrich Klakow, and Vera Demberg. 2018. Toward Bayesian synchronous tree substitution grammars for sentence planning. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 391–396, Tilburg University, The Netherlands. Association for Computational Linguistics.

Amy Isard. 2016. The methodius corpus of rhetorical discourse structures and generated texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1732–1736, Portorož, Slovenia. European Language Resources Association (ELRA).

Zhijing Jin, Qipeng Guo, Xipeng Qiu, and Zheng Zhang. 2020. GenWiki: A dataset of 1.3 million content-sharing text and graphs for unsupervised graph-to-text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*,

pages 2398–2409, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Juraj Juraska, Kevin Bowden, and Marilyn Walker. 2019. ViGGO: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 164–172, Tokyo, Japan. Association for Computational Linguistics.

Mihir Kale and Scott Roy. 2020. Machine translation pre-training for data-to-text generation - a case study in Czech. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 91–96, Dublin, Ireland. Association for Computational Linguistics.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Studies in Linguistics and Philosophy. Springer Netherlands.

Paulus Kieviet. 2017. *A Grammar Of Rapa Nui*. Zenodo.

Young Jin Kim, Marcin Junczys-Dowmunt, Hany Hassan, Alham Fikri Aji, Kenneth Heafield, Roman Grundkiewicz, and Nikolay Bogoychev. 2019. From Research to Production and Back: Ludicrously Fast Neural Machine Translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 280–288, Hong Kong. Association for Computational Linguistics.

Kevin Knight and Vasileios Hatzivassiloglou. 1995. Two-level, many-paths generation. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 252–260, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Geert-Jan M. Kruijff. 2001. *A Categorial-Modal Logical Architecture of Informativity*. Ph.D. thesis, Charles University, Prague, Czech Republic.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 704–710, Montreal, Quebec, Canada. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural Text Generation from Structured Data with Application to the Biography Domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. 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.

Percy Liang, Michael Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 91–99, Suntec, Singapore. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Shuming Ma, Pengcheng Yang, Tianyu Liu, Peng Li, Jie Zhou, and Xu Sun. 2019. Key fact as pivot: A two-stage model for low resource table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2047–2057, Florence, Italy. Association for Computational Linguistics.

Ahmed Magooda, Diane Litman, and Mohamed Elaraby. 2021. Exploring multitask learning for low-resource abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1652–1661, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zola Mahlaza and C. Maria Keet. 2022. Surface realisation architecture for low-resourced african languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561, Uppsala, Sweden. Association for Computational Linguistics.

Aleksandre Maskharashvili, Symon Stevens-Guille, Xintong Li, and Michael White. 2021. Neural methodius revisited: Do discourse relations help with pre-trained models too? In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 12–23, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Fei Mi, Minlie Huang, Jiyong Zhang, and Boi Faltings. 2019a. Meta-learning for low-resource natural language generation in task-oriented dialogue systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3151–3157. International Joint Conferences on Artificial Intelligence Organization.

Fei Mi, Minlie Huang, Jiyong Zhang, and Boi Faltings. 2019b. Meta-learning for low-resource natural language generation in task-oriented dialogue systems. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 3151–3157. AAAI Press.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018. The First Multilingual Surface Realisation Shared Task (SR'18): Overview and Evaluation Results. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 1–12, Melbourne, Australia. Association for Computational Linguistics.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, and Leo Wanner. 2019. The Second Multilingual Surface Realisation Shared Task (SR'19): Overview and Evaluation Results. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, pages 1–17, Hong Kong, China. Association for Computational Linguistics.

Simon Mille, Anya Belz, Bernd Bohnet, Thiago Castro Ferreira, Yvette Graham, and Leo Wanner. 2020. The third multilingual surface realisation shared task (SR'20): Overview and evaluation results. In *Proceedings of the Third Workshop on Multilingual Surface Realisation*, pages 1–20, Barcelona, Spain (Online). Association for Computational Linguistics.

D. Moher, A. Liberati, J. Tetzlaff, D. G Altman, and for the PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*, 339(jul21 1):b2535–b2535.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. 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 2267–2277, Minneapolis, Minnesota. Association for Computational Linguistics.

Tarek Naous, Wissam Antoun, Reem Mahmoud, and Hazem Hajj. 2021. Empathetic BERT2BERT conversational model: Learning Arabic language generation with little data. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 164–172, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing NLG data: Pictures elicit better data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 265–273, Edinburgh, UK. Association for Computational Linguistics.

Shereen Oraby, Lena Reed, Shubhangi Tandon, Sharath T.S., Stephanie Lukin, and Marilyn Walker. 2018. Controlling personality-based stylistic variation with neural natural language generators. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 180–190, Melbourne, Australia. Association for Computational Linguistics.

Nikolaos Panagiaris, Emma Hart, and Dimitra Gkatzia. 2021. Generating unambiguous and diverse referring expressions. *Computer Speech Language*, 68:101184.

Nivranshu Pasricha, Mihael Arcan, and Paul Buitelaar. 2020. NUIG-DSI at the WebNLG+ challenge: Leveraging transfer learning for RDF-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 137–143, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Laura Perez-Beltrachini and Mirella Lapata. 2018. Bootstrapping generators from noisy data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1516–1527, New Orleans, Louisiana. Association for Computational Linguistics.

Laura Perez-Beltrachini, Rania Sayed, and Claire Gardent. 2016. Building RDF content for data-to-text generation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1493–1502, Osaka, Japan. The COLING 2016 Organizing Committee.

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(140):1–67.

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.

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021. Investigating pretrained language models for graph-to-text generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 211–227, Online. Association for Computational Linguistics.

V. Rieser and O. Lemon. 2011. *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*. Theory and Applications of Natural Language Processing. Springer.

Gözde Gül Şahin and Mark Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009, Brussels, Belgium. Association for Computational Linguistics.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. 2019. Pragmatically informative text generation. 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 4060–4067, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaoyu Shen, Ernie Chang, Hui Su, Cheng Niu, and Dietrich Klakow. 2020. Neural data-to-text generation via jointly learning the segmentation and correspondence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7155–7165, Online. Association for Computational Linguistics.

Anastasia Shimorina and Claire Gardent. 2019. Surface realisation using full delexicalisation. 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 3086–3096, Hong Kong, China. Association for Computational Linguistics.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60.

Somayajulu Gowri Sripada, Jin Yu, Ian P. Davy, and Wni Oceanroutes. 2008. Sumtime-meteo: Parallel corpus of naturally occurring forecast texts and weather data.

M. Steedman and J. Baldridge. 2006. Combinatory categorial grammar. In Keith Brown, editor, *Encyclopedia of Language Linguistics (Second Edition)*, second edition edition, pages 610–621. Elsevier, Oxford.

Symon Stevens-Guille, Aleksandre Maskharashvili, Amy Isard, Xintong Li, and Michael White. 2020. Neural NLG for methodius: From RST meaning representations to texts. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 306–315, Dublin, Ireland. Association for Computational Linguistics.

Lisa A. Torrey and Jude W. Shavlik. 2009. Transfer learning. In *Handbook of Research on Machine Learning Applications*.

Van-Khanh Tran and Le-Minh Nguyen. 2018. Adversarial domain adaptation for variational neural language generation in dialogue systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1205–1217, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chris van der Lee, Chris Emmery, Sander Wubben, and Emiel Krahmer. 2020. The CACAPO dataset: A multilingual, multi-domain dataset for neural pipeline and end-to-end data-to-text generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 68–79, Dublin, Ireland. Association for Computational Linguistics.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. Learning language games through interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2368–2378, Berlin, Germany. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. 2016. Multi-domain neural network language generation for spoken dialogue systems. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 120–129, San Diego, California. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Michael White. 2006. CCG chart realization from disjunctive inputs. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 12–19, Sydney, Australia. Association for Computational Linguistics.

Xinnuo Xu, Ondřej Dušek, Verena Rieser, and Ioannis Konstas. 2021. AggGen: Ordering and aggregating while generating. 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 1419–1434, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Steve Young. 2007. CUED standard dialogue acts. *Report, Cambridge University Engineering Department, 14th October*, 2007.

Tiancheng Zhao and Maxine Eskenazi. 2018. Zero-shot dialog generation with cross-domain latent actions. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 1–10, Melbourne, Australia. Association for Computational Linguistics.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2019. Multi-task learning for natural language generation in task-oriented dialogue. 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 1261–1266, Hong Kong, China. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.