Journal of Artificial Intelligence Research X (XXX) 1-15 Submitted ; published

Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning

Erkut Erdem ERKUT@CS.HACETTEPE.EDU.TR
Menekse Kuyu MENEKSE.KUYU@HACETTEPE.EDU.TR
Semih Yagcioglu SEMIH.YAGCIOGLU@HACETTEPE.EDU.TR
Hacettepe University, Ankara, Turkey

Anette Frank FRANK@CL.UNI-HEIDELBERG.DE
Letitia Parcalabescu PARCALABESCU@CL.UNI-HEIDELBERG.DE
Heidelberg University, Heidelberg, Germany

Barbara Plank BPLANK@ITU.DK
IT University of Copenhagen, Copenhagen, Denmark

Andrii Babii ANDRII.BABII@NURE.UA
Oleksii Turuta OLEKSII.TURUTA@NURE.UA
Kharkiv National University of Radio Electronics, Ukraine

Aykut Erdem AERDEM@KU.EDU.TR
Koç University, Istanbul, Turkey

Iacer Calixto IACER.CALIXTO@NYU.EDU
New York University, U.S.A.
University of Amsterdam, Netherlands

Elena Lloret ELLORET@DLSI.UA.ES
University of Alicante, Alicante, Spain

Elena-Simona Apostol ELENA.APOSTOL@UPB.RO
Ciprian-Octavian Truică CIPRIAN.TRUICA@UPB.RO
University Politehnica of Bucharest, Bucharest, Romania

Branislava Šandrih BRANISLAVA.SANDRIH@FIL.BG.AC.RS
University of Belgrade, Belgrade, Serbia

Albert Gatt ALBERT.GATT@UM.EDU.MT
University of Malta, Malta

Sanda Martinčić-Ipšić SMARTI@UNIRI.HR
University of Rijeka, Rijeka, Croatia

Gábor Berend BERENDG@INF.U-SZEGED.HU
University of Szeged, Szeged, Hungary

Gražina Korvel GRAZINA.KORVEL@MIF.VU.LT
Vilnius University, Vilnius, Lithuania

Abstract

Developing artificial learning systems that can understand and generate natural language has been one of the long-standing goals of artificial intelligence. Recent decades have witnessed an impressive progress on both of these problems, giving rise to a new family of approaches. Especially, the advances in deep learning over the past couple of years have

led to neural approaches to natural language generation (NLG). These methods combine generative language learning techniques with neural-networks based frameworks. With a wide range of applications in natural language processing, neural NLG (NNLG) is a new and fast growing field of research. In this state-of-the-art report, we investigate the recent developments and applications of NNLG in its full extent from a multidimensional view, covering critical perspectives such as multimodality, multilinguality, controllability and learning strategies. We summarize the fundamental building blocks of NNLG approaches from these aspects and provide detailed reviews of commonly used preprocessing steps and basic neural architectures. This report also focuses on the seminal applications of these NNLG models such as machine translation, description generation, automatic speech recognition, abstractive summarization, text simplification, question answering and generation, and dialogue generation. Finally, we conclude with a thorough discussion of the described frameworks by pointing out some open research directions.

1. Introduction

Humans communicate and express information through natural language. Research within Artificial Intelligence, in particular Natural Language Processing (NLP), is concerned with the automatic analysis, representation and generation of human language. Generation of language is the focus of the subfield of Natural Language Generation (NLG).

Generation is at the heart of human-machine interfaces. Examples of tasks that facilitate this are dialogue systems, question answering, machine translation, summarization and image captioning. Traditionally, NLG has been approached as a pipeline of several stages (Reiter & Dale, 2000), involving (i) macroplanning – deciding "what to say?", (ii) microplanning – choosing the appropriate structures and vocabulary, and (iii) surface realization – determining the final output, or "how to say it?", given the information provided in the previous stages. While early work on language generation relied on linguistic patterns which had been defined *a priori*, the field has witnessed a revolution in the past few years. The development and evaluation of statistical models based on neural architectures —Neural Natural Language Generation (NNLG) —has shifted the research focus away from knowledge-based approaches motivated by linguistic theories, which predominated in the 90s (Bateman & Zock, 2003). The increasing number of neural approaches for NLG has been also reported and make evident in the recent NLG surveys, e.g., (Gatt & Krahmer, 2018; Iqbal & Qureshi, 2020), further detailed in Section 2).

In parallel, another development has been the exponential increase in information, both in terms of volume, and in terms of type, format, language, etc. This has increased the importance of *multilinguality* and *multimodality* in recent NLG approaches. Two further developments which have an impact on NLG are *learning strategies* and *controllability*. These four dimensions—multilinguality, multimodality, learning strategies, controllability—are important in the context of NNLG to maximize its potential. Tasks such as Machine Translation, Question Generation, and Abstractive Summarization, among others, are examples of applications in which these dimensions can play a crucial role, both individually and in tandem. Therefore, the objective of this survey is to provide the reader with an overview of recent advances in NNLG from a *multidimensional perspective*, focusing on the most recent neural approaches. In this survey, both aspects—dimensions and tasks—are discussed in detail, demonstrating their implementation in NNLG tasks and applications.

To provide the reader with a general overview of the state of the art in NNLG, the rest of the paper is structured as follows: We first provide a brief list of the most recent surveys related to different aspects of NLG with an emphasis on neural approaches, together with the main scope of the present survey (Sections 2 and 3). In Section 4, we present the four main dimensions in which NNLG can be studied. In Section 5, we describe the fundamental building blocks, including common preprocessing steps, and basic neural architectures. After that, in Section 6, we investigate all major language- and speech-related NNLG applications widely studied in the community. In particular, we provide an in-depth review of seven popular tasks, namely machine translation, description generation,[1] automatic speech recognition, abstractive summarization, text simplification, question answering/generation,

1. Throughout this manuscript, the terms "image captioning" and "description generation" will be used interchangeably.

and dialogue generation. Finally, we give concluding remarks in Section 7 where we outline some open research directions.

2. Related Surveys

A number of recent survey papers related to different aspects of NLG can be found in the literature. These aspects include a focus on multimodality by integrating vision and language, for instance, and diverse language generation tasks, such as machine translation, description generation or abstractive summarization. In this section, we conduct a brief analysis of the most recent surveys (i.e., published in the period 2016-2020) that deal with any aspect that has some impact on language generation. Our goal in this Section is to clarify what has been reported so far, tackling NLG from a different angle from what we would like to report in our proposed NNLG survey.

General purpose NLG surveys. A general perspective together with the fundamentals and evolution of NLG is provided in (Gatt & Krahmer, 2018) and (Santhanam & Shaikh, 2019). Gatt and Krahmer (2018) mainly focus on data-driven approaches, and Santhanam and Shaikh (2019) on specific NLG tasks and applications, such as open domain dialogue systems. Recently, Iqbal and Qureshi (2020) propose another survey where the focus is on deep learning models that can be applied to NLG, extending the aforementioned work of Gatt and Krahmer (2018) and concluding that generative adversarial networks (GANs) seem to dominate NLG when working with images, whereas variational auto-encoders (VAEs) are predominantly used with texts. The evaluation of NLG systems is also an important aspect and was surveyed in (Celikyilmaz, Clark, & Gao, 2020) and (van der Lee, Gatt, van Miltenburg, & Krahmer, 2021), with the latter focusing primarily on human evaluation.

Undoubtedly, in line with the trend in other fields of Artificial Intelligence, neural models are currently the dominant technique used in NLG, which has witnessed a gradual shift from traditional, rule-based approaches, to statistical, data-driven ones (Gatt & Krahmer, 2018). Moreover, NLG is a very broad area and we note that despite in-depth treatments of several topics in previous surveys many important subareas were either left out or were not the main focus. Important dimensions such as *multilinguality*, *multimodality*, *controllability* and *learning strategies* are important in the context of NLG to maximize its potential. Tasks such as Machine Translation, Question Generation, Abstractive Summarization, among others, can be also considered a form of text production. Thus, a survey is warranted which covers the most recent trends and focusses on these dimensions and tasks.

Surveys on dimensions relevant to NLG. Ruder (2017) provides an overview of the state of multi-task learning with deep neural networks, finding that one of its weaknesses is its limited generalization capabilities. Belinkov and Glass (2019) review analysis methods in the context of neural networks. Among their main conclusions, we highlight two which are directly related to the aim of this survey: (i) the lack of methods and resources in languages other than English (and therefore the need to develop them); and (ii) the need for more challenge sets for evaluating tasks other than natural language inference and machine translation. Regarding multilinguality, Ruder, Vulić, and Søgaard (2019) surveys cross-lingual word embeddings due to their usefulness in building a common representation space that enables model transfer between languages. Baltrusaitis, Ahuja, and Morency (2019)

addresses multimodality in the context of machine learning approaches, whereas Wiriyathammabhum, Summers-Stay, Fermüller, and Aloimonos (2016), Mogadala, Kalimuthu, and Klakow (2021) focus more on how the different modalities in a system's input (e.g. images, videos, text) can be integrated. Regardless of whether the main focus is on the input side or the algorithmic one, the important aspect in these surveys is how to process and relate information from multiple modalities.

NLG Tasks Surveys. Among surveys that deal with specific NLP tasks that generates language as output and therefore can be considered NLG applications, Dabre, Chu, and Kunchukuttan (2020) review multilingual neural machine translation approaches. The survey concludes that more human evaluation of these models is necessary to have a better understanding of the impact of multilingualism. Sulubacak, Caglayan, Grönroos, Rouhe, Elliott, Specia, and Tiedemann (2020) analyze multimodality in machine translation, advocating the need for bigger and more challenging multimodal datasets in both the input and output space, as well as targeted evaluations to better compare model performance. Concerning abstractive summarization, there are also several surveys that provide the basic background about this type of NLG task (Aries, Zegour, & Hidouci, 2019; Baumel & Elhadad, 2019; Gupta & Gupta, 2019; Lin & Ng, 2019). Specifically, Lin and Ng (2019) and Baumel and Elhadad (2019) analyze neural network models for abstractive summarisation, especially in view of the fact that such models provide a viable framework for obtaining an abstract representation of the meaning of an input text and generating informative, fluent, and humanlike summaries. However, there is no survey that analyzes abstractive summarization with respect to the multilinguality or multimodality dimensions, as has been done for machine translation.

Research addressing tasks that include video or images as input and produce text as output can be considered multimodal by definition. These works mainly focus on video or image description/captioning and surveys such the ones in (Aafaq, Mian, Liu, Gilani, & Shah, 2020; Hrga & Ivašić-Kos, 2019; He & Deng, 2017; Bernardi, Cakici, Elliott, Erdem, Erdem, Ikizler-Cinbis, Keller, Muscat, & Plank, 2016) provide a good starting point to understand these tasks. According to existing surveys (Aafaq et al., 2020), research on video description is far from mature at present. This is partly due to the fact that the analysis of video description models is a very challenging task, and existing evaluation metrics fall short of measuring the agreement between machine-generated descriptions with those of humans. Regarding image captioning or description generation, current systems are limited in the diversity of output they generate, with a tendency to reproduce what appears in the training data. Furthermore, they still do not robustly infer the underlying semantics of images to generate novel descriptions. Another research line for future work in the context of description generation is to investigate multilingual approaches. Visual Question Generation is another task that has recently gained traction. Here, a model takes an image as input and generates meaningful questions based on that input image. Patil and Patwardhan (2020) review the state-of-the-art for this task, with an overview of the main techniques, datasets and evaluation metrics.

3. Scope of the present survey

There is no shortage of surveys on language generation, but to the best of our knowledge none of the existing surveys explore language generation from a multidimensional perspective. That is, they do not fully address issues related to the type of input, output language or architectures used, and do not cover broad range of applications that directly or indirectly involve language generation at their core, specifically when one is interested in the recent neural network based approaches. While the recent survey by Jin, Cao, Wang, Xing, and Wan (2020), Garbacea and Mei (2020) cover some aspects of neural NLG, mainly models and metrics but also some tasks, it does not pay attention to the multidimensional perspectives mentioned in Section 1.

With the present survey, we aim to provide a broader understanding and an overall perspective of the most recent advances in neural language generation from different angles. In this manner, this survey takes into account both the relevant dimensions for different language generation tasks, as well as a set of representative applications. We include five different dimensions in our analysis, all of which are further described in Section 4: multilinguality (Section 4.1), multimodality (Section 4.2), learning strategies (Section 4.3), controllability (Section 4.4). We then introduce fundamental preprocessing steps and the general neural architectures (Section 5), hence an advanced reader may skip this section. The specific tasks we survey are the following[2]: Machine Translation (Section 6.1), Description Generation (Section 6.2), Automatic Speech Recognition (Section 6.3), Abstractive Summarization (Section 6.4), Text Simplification (Section 6.5), Question Answering and Generation (Section 6.6), and Dialogue Generation (Section 6.7).

4. Neural Natural Language Generation

This survey studies methods to perform Natural Language Generation (NLG) using neural methods, with a special focus on *multiple dimensions*, as introduced in Section 1[3]. We consider *neural NLG methods* from the perspectives of *multiple languages*, *multiple modalities,* and in view of *multiple application tasks.*

NLG is an NLP task that comes with aims and in flavors that are as diverse as human life, people's interactions and the languages they speak. Put abstractly and from a bird's eye view, the essence of NLG is that *information* encoded in *inputs* stemming from a single or multiple *modalities* – may it be text, speech, images, videos, etc. – is *mapped* by a system to a single or multiple *outputs in natural language in textual form.* The mapping performed by the system, and the choice of methods that ensure high-quality NLG output crucially depend on the chosen input and the targeted output languages or language varieties, but most importantly will have to realize a *task-specific mapping* from the inputs to natural language outputs. Performing a specific task for a given input and a desired output may require specific *methods* to obtain optimal results. Hence, we consider the *specific kind of mapping* that has to be performed to solve a *specific NLG task* as an important dimension for designing suitable neural NLG methods. Clearly, such mappings can come in many

2. See `https://github.com/Multi3Generation/neural-natural-language-generation` for the list of official implementations of the papers (if any) reviewed in our survey.
3. In this survey, we are focusing on the main dimensions and tasks of the Multi3Generation COST Action (`https://multi3generation.eu`).

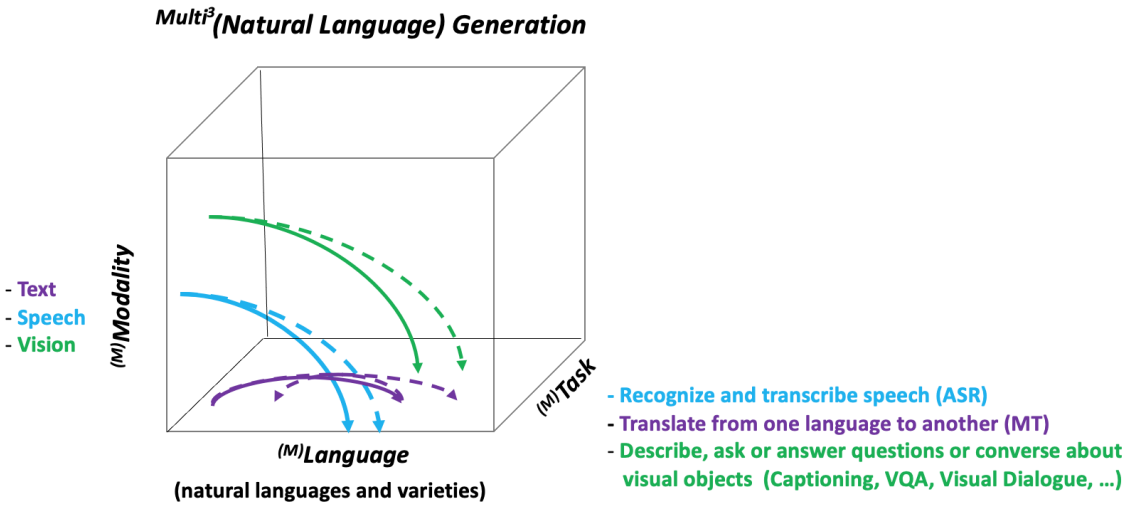

Figure 1: $^{Multi^3}$**NLG**: NLG from a multilingual, multimodal and multi-task perspective.

varieties: we may want to translate content delivered by textual or spoken input – ideally 1:1 – to the chosen output language; we may want to summarize important pieces of content from the input and convey it in a short, abstracted manner (as in text summarization) or by describing prominent information captured in an image in NL sentences, etc.

Hence, for the purpose of this survey, we define the NLG task from three perspectives (cf. Figure 1), in form of the targeted (input or output) *languages*, different types of *modalities* of the input, and the desired *application tasks*, as follows:

> NLG is the task of **mapping content** encoded in a given input – originating from a single or various modalities – to **NL output** in a chosen language, by translating, summarizing, simplifying, or otherwise modifying the input, or by describing or explaining the input, asking or answering questions about or reacting to it, in performing a specific **application task**. When solving such task-specific mappings, NLG methods apply a variety of *learning and control strategies* to achieve optimal results.

How does this definition apply to multiple languages, modalities and tasks addressed in this survey?

$^{(M)}$**Modality.** The input to a NLG system may originate from different modalities, such as text, speech, or images. While these constitute *unstructured* inputs, we may also consider *structured* inputs, as in data2text generation where a system may be tasked to answer a user's question in NL on the basis of a structured knowledge graph or database, or where a system must verbalize information contained in a table (as in financial or weather reports) or in a structured meaning representation (e.g., in AMR2Text generation).[4]

4. We do not consider data2text variants of NLG in this survey although it is covered by the above definition.

$^{(M)}$**Language.** Crucial for our definition of NLG is that a system's output is always *natural language in textual form*.[5] However, the generated output[6] may be chosen from manifold languages, including varieties such as dialects, or linguistic styles or genres.

$^{(M)}$**Task.** The core function of NLG is to map a given input to a textual output in NL. This function crucially depends on the chosen NLG task, as well as the chosen input and output variables. For any such task, it is crucial that the generated output conforms to the *task's specific mapping function*. The spectrum of such mapping functions is large, and may continue to grow in the future. It ranges from translating (1:1) between languages, over summarizing relevant pieces of content of a text, simplifying its language or varying its language style, or may consist in triggering situation-adapted conversational turns. All these different mapping functions may be applied in mono- or multilingual settings, with single or multiple modalities, e.g., with or without grounding language in vision and speech. Performing NLG from spoken language inputs has its own challenges, tackled in ASR[7], but may also be combined with translation or summarization; also, performing NLG in combination with visual inputs covers various functions based on the content encoded in them: visual inputs may be described or explained using NL, and they may serve as a basis for generating questions or answers about them, by combining vision and language inputs.

$^{Multi^3}$**NLG.** Hence, the *nature of the various kinds of mappings that NLG needs to perform* is crucially determined by the task at hand, in interaction with the chosen input and output dimensions. Given the manifold combinations, research needs to develop suitable methods of learning and controlling NLG, to ensure optimal results.

4.1 Multilinguality

Multilinguality in NLP refers to the problem of processing more than one language. In Natural Language Generation in particular, multilinguality refers to the problem of automatic generation of texts in multiple languages. The exact definition of multilinguality in NLP and NLG depends however on the way an architecture deals with multiple languages.[8]

Traditional NLG architectures that focused on this aspect distinguish between weak and strong multilinguality (Bateman & Licheng, 1999). A system is said to exhibit **"weak" multilinguality** if it consists of different generators for each language, each with its own algorithms and data structures. Thus, even though the final NLG system addresses several languages, the individual components are only aware of one of the languages at a time. In contrast, **"strong" multilinguality** is a term used for systems in which a single generator applies to all of the languages to be covered by the multilingual system.

5. Hence, our definition does not consider speech synthesis as a NLG task.
6. and equally the input, for that matter
7. Converting spoken language to text can be problematic. ASR produces raw word sequences without punctuation and capitalization. Furthermore, the information present in speech such as intonation, hesitation, etc. is reduced by converting speech into text.
8. In the early multilingual NLP literature, the term *multilingual* was used in the meaning of *independent evaluation on multiple languages*, i.e., a single system was trained independently on many languages (without sharing) and it was typically evaluated on each language individually. In recent years, *multilingual* typically refers to a single system trained jointly on many languages. To distinguish this terminological confusion, the term *polyglot* learning was proposed (Mulcaire, Kasai, & Smith, 2019) to refer to training a single model on data from multiple languages. Albeit the term *polyglot* is not widely used, most *multilingual* models today are essentially polyglot-trained systems.

Two of the advantages of strong multilinguality identified by Bateman and Licheng (1999) are modularity and generalizability, i.e., having a clear separation of process (generation) and data (linguistic description), where the generation process must from the outset be designed to be adequate for handling a variety of language specifications, thus making it more likely to support general solutions. In the last decades, the NLP research community has developed a wide range of algorithms, methods and applications for addressing multilinguality. The predominant approaches rely on statistical methods (Bikel & Zitouni, 2012) and the creation of machine learning models able to understand and process different languages.

A cornerstone of recent multilingual models is multilingual word representations, that is, embeddings of words in two or more languages in a single, high-dimensional space, as further discussed in Section 4.3.1.

However, some concerns have been raised regarding the limitations and deficiencies of multilingual models for NLG tasks and their effectiveness for minority languages (Rönnqvist, Kanerva, Salakoski, & Ginter, 2019; Pires, Schlinger, & Garrette, 2019).

4.2 Multimodality

Multimodal research is already an established sub-field of machine learning, but (to the best of our knowledge) still lacks a rigorous definition of *modality*.

Baltrusaitis et al. (2019) propose a taxonomy of approaches to multimodal learning, using as their starting point a **perceptual, human-centered view** of modality, as follows: "Our experience of the world is multimodal - we see objects, hear sounds, feel texture, smell odors, and taste flavors. Modality refers to the way in which something happens or is experienced". The downside of this definition arises because machines are sensing the world very differently compared to humans: While we can read and comprehend both text (*e.g.* in a text file) and an image of text, a machine has to be programmed very differently if it has to generate text as a response to ASCII-based input, rather than images capturing the same text.

An alternative definition is the **machine-centered view**. Guo, Wang, and Wang (2019) formulate it as follows: "In the representation learning area, the word "modality" refers to a particular way or mechanism of encoding information." Similarly, Bernsen (2008) state: "a modality is defined by its physical medium and its particular "way" of representation." The upside of this definition consists in capturing the case where modalities could be represented differently but bear the same information; a case in which machines would not automatically notice the content overlap. But this focus on representation might miss crucial aspects of multimodality in machine learning, since *e.g.* PNG or JPEG are different encodings of the same image, but not different modalities. Conversely, an infrared image is usually represented as a PNG like an usual image, but delivers different information entirely.

While both human-centered and machine-centered definitions have pros and cons, for the scope of this paper, we will rely on the **perceptual, human-centered view**. This view better captures the motivation for multimodal learning, in that it draws attention to the complementarity of information reaching us via multiple channels, and it is the combination of these channels that is the primary goal of multimodal learning.

The fact that information may be present in complementary modalities raises challenges for multimodal NLG. Firstly, models should focus on multiple multimodal inputs, otherwise they miss information present in one but missing in the other. Secondly, the relationship between the two modalities need not be symmetrical from the generation perspective. For example in image captioning, there are multiple textual descriptions for the same image; conversely, there are many images realizing the same textual input.

4.3 Learning Strategies

This section provides an overview of the main learning strategies. As it turns out, multi-task learning is a very popular learning strategy across the tasks surveyed here. We refer the reader to the details in the task-specific sections (Section 6) as they differ per application. In the following subsections (4.3.1 and 4.3.2), we outline important learning strategies in the context of multilingual and multimodal modeling, respectively.

4.3.1 Multilingual Learning Strategies

Learning strategies for neural multilingual models differ largely in the way they use multilingual information. Early work on learning **static cross-lingual word representations** takes different assumptions in data sources and algorithms, and is surveyed in (Ruder et al., 2019). This includes the type of *alignment* (alignment at the word level, the sentence or the document level), and the *comparability* of the sources (whether exact translations as parallel data is required, or comparable data is sufficient). Algorithms are broadly classified into *mapping-based* and *joint* approaches, which differ in whether first monolingual representations are learned and then mapped to a multilingual space, or a multilingual space is directly learned (jointly on all languages). The authors point out that "the data a method requires to learn to align a cross-lingual representation space—is more important for the final model performance than the actual underlying architecture". Some recent examples of such multilingual representations include LASER embeddings[9] (Schwenk & Douze, 2017) or Multilingual USE embeddings[10] (Yang, Cer, Ahmad, Guo, Law, Constant, Abrego, Yuan, Tar, hsuan Sung, Strope, & Kurzweil, 2020), covering 93 and 16 languages, respectively.

In recent years, the advent of **contextualized multilingual word representations** has significantly pushed the field with seminal architectures like ELMo (Peters, Neumann, Iyyer, Gardner, Clark, Lee, & Zettlemoyer, 2018) and BERT (Devlin, Chang, Lee, & Toutanova, 2019) first proposed for monolingual learning. The latter, BERT, uses multi-task learning by optimizing two objectives—masked language-modeling and next-sentence prediction—while learning on large quantities of raw text. Multilingual variants like multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-R (Conneau, Khandelwal, Goyal, Chaudhary, Wenzek, Guzmán, Grave, Ott, Zettlemoyer, & Stoyanov, 2020) were soon after proposed, obtained by training on data from multiple languages jointly. In contrast to early research on static word embeddings, the focus has shifted to algorithms, as many architectures emerged. What they have in common is that these models enabled scaling to much larger quantities of data, yielding strong cross-lingual representations even without explicit cross-lingual supervision signal (like alignment of data). The use of language tags became

9. https://github.com/facebookresearch/LASER
10. https://aihub.cloud.google.com/p/products%2F4e63f320-d774-4772-aaaa-ccbe8f3f09f2

ubiquitous when training (or applying) such models, to indicate the language trained on (or asked to generate). This shift to contextualized models resulted in surprisingly effective multilingual models. Recent multilingual models like mBERT[11] and XLM-R[12] (Conneau et al., 2020) support 104 and 100 languages, respectively. There exist models that can generate texts (instead of only predicting masked tokens or their variants). These generative multilingual language models include GPT2-ML[13], mT5 (Xue, Constant, Roberts, Kale, Al-Rfou, Siddhant, Barua, & Raffel, 2021) and mBART (Liu, Gu, Goyal, Li, Edunov, Ghazvininejad, Lewis, & Zettlemoyer, 2020), which were pre-trained on very large corpora from Common Crawl-based data—the CCNet Corpus (Wenzek, Lachaux, Conneau, Chaudhary, Guzmán, Joulin, & Grave, 2020). mT5 and mBART support 101 and 25 different languages, respectively. Multilingual models can generalize across languages without supervision (zero-shot evaluation).

Problems. While multilingual models have without doubt brought significant advances for enabling NLP for many languages, the main issue currently is that multilingual models are not universal. For instance, mBERT covers only 104 languages, which is a third of the languages found in Wikipedia (Hedderich, Lange, Adel, Strötgen, & Klakow, 2021). Moreover, languages are not represented equally well in multilingual models: multilingual BERT works well for high-resource languages, but it is much worse on low-resource languages (Wu & Dredze, 2020). State-of-the-art multilingual models rely on a vocabulary based on subword units inferred from all data, which favors frequent subwords and thereby is more favorable to higher-resource languages. Recent work includes work on improving the vocabulary of multlingual models, for example by using language-clustered vocabularies (Chung, Garrette, Tan, & Riesa, 2020). This targets the balance of general subword sharing versus modeling language-specific information, and is one step towards improving such models for low-resource languages. Training data scarcity and balancing of data are further factors that challenge current models (Lauscher, Ravishankar, Vulić, & Glavaš, 2020; Conneau et al., 2020).

4.3.2 Multimodal Learning Strategies

We distinguish between two multimodal scenarios:

Multimodal input. When working with inputs from at least two different modalities (*e.g.* text and images in Multimodal MT, speech and gestures in speech recognition), it is common practice for one input modality to be used to ground the other input modality. In vision and language learning, for example, we can use the image to visually ground the language modality (as in Phrase Grounding, Multimodal MT, or Visual Dialogue), or vice versa: we can use language to textually ground the image (as in sentence-image retrieval). After grounding, the modalities have to be fused or aligned implicitly before producing the output of the downstream task, so that downstream applications are able to recover and further process information from each modality.

Among the first statistical approaches for multimodal fusion were algorithms based on CCA (Hotelling, 1992), which are applied on previously extracted image and textual

11. https://github.com/google-research/bert/blob/master/multilingual.md
12. https://github.com/facebookresearch/XLM
13. https://github.com/imcaspar/gpt2-ml

features[14]. More elaborate, neural multimodal methods combine well-established unimodal architectures (CNN for images, LSTM for text) as branches of a unified multimodal model (see Section 5 for a thorough review of fundemantal architectures in NLG). The unimodal sub-branches are usually fused by concatenation (Kiela & Bottou, 2014; Regneri, Rohrbach, Wetzel, Thater, Schiele, & Pinkal, 2013; Shekhar, Takmaz, Fernández, & Bernardi, 2019). or element-wise vector multiplication (Fukui, Park, Yang, Rohrbach, Darrell, & Rohrbach, 2016; Wang, Li, Huang, & Lazebnik, 2018), outer product (Fukui et al., 2016) or attention (Yu, Yu, Cui, Tao, & Tian, 2019). Other approaches map the resulting representations of the unimodal branches into a common space, by enforcing a rank-distance loss (Wang et al., 2018), training the model to keep representations of objects that are shared in both modalities close to each other in the joint space.

The fusion is typically followed by a task-specific head that provides the loss for the multimodal model. The advantages of embedding purely/full-fledged unimodal model components within the multimodal system architecture is that in this way, weights that are learnt from unimodal tasks (like image recognition or language modelling) can be transferred and further adapted within the multimodal architecture on specific multimodal tasks.

For Vision and Language tasks, the current state-of-the-art relies on multimodal Transformer architectures that come in two flavours: *single-stream models* such as VideoBERT (Sun, Myers, Vondrick, Murphy, & Schmid, 2019), VL-BERT (Su, Zhu, Cao, Li, Lu, Wei, & Dai, 2019), Unicoder-VL (Li, Duan, Fang, Gong, Jiang, & Zhou, 2020), or UNITER (Chen, Li, Yu, Kholy, Ahmed, Gan, Cheng, & Liu, 2020) employ a single Transformer for the visual and textual modalities. The *two-stream architecture*, as used by e.g., ViLBERT (Lu, Batra, Parikh, & Lee, 2019) or LXMERT (Tan & Bansal, 2019) combines two unimodal Transformers with a third Transformer that performs cross-modal fusion. Some recent work suggests that single and two-stream models can be unified in a common framework (Bugliarello, Cotterell, Okazaki, & Elliott, 2021).

In both types of architectures, the learning of visual-linguistic representations happens through different stages that employ transfer learning and multi-task learning strategies: First, often the textual branch is initialised with the weights of BERT (Devlin et al., 2019). Furthermore, the visual feature vectors are extracted by a visual backbone[15] which is already pre-trained on image recognition tasks. Secondly, a self-supervised multimodal and multi-task pre-training stage learns generic multimodal representations on tasks including multimodal masked language modelling, masked visual feature classification, or image-sentence alignment. Thirdly, the pre-trained model is further fine-tuned on a downstream task, such as image retrieval, phrase grounding, or VQA – in some cases again in a multi-task fashion. (Lu, Goswami, Rohrbach, Parikh, & Lee, 2020) show that multi-task learning over 12 different vision and language tasks can improve the performance on individual downstream tasks.

Translating between Modalities. In cases where we translate or transform input in one modality to output in another modality, the task and model qualify as multimodal according to our definition (cf. 4.2). Typical tasks are for instance image or video captioning, speech

14. See for example (Sargin, Yemez, Erzin, & Tekalp, 2007; Plummer, Wang, Cervantes, Caicedo, Hockenmaier, & Lazebnik, 2015; Massiceti, Dokania, Siddharth, & Torr, 2018)
15. For example, Faster-RCNN (Ren, He, Girshick, & Sun, 2015b), or MaskRCNN (He, Gkioxari, Dollár, & Girshick, 2017)

recognition, or image generation from text. The learning strategies generally employ a unimodal encoder for the input modality, and a unimodal decoder for the output. For example, image captioning might be performed using a a CNN encoder and LSTM decoder (Bernardi et al., 2016; Tanti, Gatt, & Camilleri, 2018). Text-to-image generation can be peformed using a GAN conditioned on the CNN-RNN text encoding of the textual modality (Reed, Akata, Mohan, Tenka, Schiele, & Lee, 2016).

Problems. The main issues found with multimodal architectures are currently that they tend to ignore one modality and instead focus on the other, despite their training objectives (Shekhar et al., 2019; Caglayan, Madhyastha, Specia, & Barrault, 2019; Cao, Gan, Cheng, Yu, Chen, & Liu, 2020; Agarwal, Bui, Lee, Konstas, & Rieser, 2020). The reasons are twofold: Firstly, datasets often contain statistical biases, so that tasks that require information from both modalities with equal importance become solvable by models which exploit data biases in a single modality to make predictions (Goyal, Khot, Summers-Stay, Batra, & Parikh, 2017; Massiceti et al., 2018; Agarwal et al., 2020). Secondly, the heterogeneity of modalities becomes a problem (Collell & Moens, 2018): The finite amounts of training data employed are usually enough for a successful data-point mapping from one modality into the other, showing high accuracies on the training set and on test data of a very similar distribution. However, the finite samples are not enough for a successful translation of the whole modality space into another modality space. Neural networks are only stretching/contracting high-dimensional spaces. If the input and output multimodal spaces are poorly sampled (relative to the complexity of the problem of translating one modality into the other), the finite sampling leads to degenerate neighbourhoods of the fitted data points during training (Collell & Moens, 2018), causing poor generalisation during testing and deployment.

4.4 Controllability

Controllability in Natural Language Generation tasks is mechanisms to generate natural sentences whose features can be defined a priori. The main problems in controlled language generation models range from the difficulty of generating text according to the given attributes, to the lack of diversity of the generated texts. The attributes to control features range from *style* such sentiment, formality, etc.; *content* such as information, keywords, entities, etc.; *structure* like ordering of information, events, like plot summaries (Prabhumoye, Black, & Salakhutdinov, 2020; Wiseman, Shieber, & Rush, 2018).

The state-of-the-art Language Models are trained with unsupervised learning. On the one hand, unsupervised learning allows to process a huge amount of training data, on the other hand, these data and output text of current task can have different structure, other features. In order to use the pre-trained Language Model to solve a specific text generation problem, it is necessary to have effective mechanisms to affected on the output text. Controllability of various attributes provides solutions for manifold NLG tasks. For instance, such as Machine Translation Crego, Kim, Klein, Rebollo, Yang, Senellart, Akhanov, Brunelle, Coquard, Deng, et al. (2016), Hokamp and Liu (2017), Susanto, Chollampatt, and Tan (2020), Description Generation Alikhani, Sharma, Li, Soricut, and Stone (2020), Question Answering and Generation, Abstractive Summarization Fan, Grangier, and Auli (2018), Text Simplification Maddela, Alva-Manchego, and Xu (2021), Dialogue Genera-

tion, Story Generation Chandu, Prabhumoye, Salakhutdinov, and Black (2019). Improved controllability of language generation is achieved through conditional training and explicit control hyper-parameters and additional input. Prabhumoye et al. (2020) discusses three abstraction levels of language generation to make machines generated text human-like. It regards structure, content and style.

Different type of source data are considered for controlled text generation, such as tables (Parikh, Wang, Gehrmann, Faruqui, Dhingra, Yang, & Das, 2020), structured data records (Nan, Radev, Zhang, Rau, Sivaprasad, Hsieh, Tang, Vyas, Verma, Krishna, Liu, Irwanto, Pan, Rahman, Zaidi, Mutuma, Tarabar, Gupta, Yu, Tan, Lin, Xiong, Socher, & Rajani, 2021), trees or graphs (Zhao, Walker, & Chaturvedi, 2020; Lin & Wan, 2021).

Structure of generated text directly depends on NLG tasks. It affects e.g. the length (Kikuchi, Neubig, Sasano, Takamura, & Okumura, 2016), and the relationship between sentences. Bridging between the semantic representation of story structure and the NLG engine discussed in (Rishes, Lukin, Elson, & Walker, 2013). In example, Orbach and Goldberg (2020) consider expanding a sequence of facts into a longer narrative. Wiseman et al. (2018) provide template-like structures. In BART, Lewis, Liu, Goyal, Ghazvininejad, Mohamed, Levy, Stoyanov, and Zettlemoyer (2020) propose an autoregressive decoder, which fine tuned for sequence generation tasks such as abstractive question answering and summarization. Su, Vandyke, Wang, Fang, and Collier (2021) propose a Plan-then-Generate (PlanGen) framework which consists of two components, namely a content planner and a sequence generator.

Content refers to the information conveyed by the generated text. Text generation tasks focused on generation based on structured input (for example, semantic representation) or free-form text (instructed data). There are various approaches to generating content. First group of approaches are based on control codes, e.g. (Keskar, McCann, Varshney, Xiong, & Socher, 2019; Dathathri, Madotto, Lan, Hung, Frank, Molino, Yosinski, & Liu, 2020; Krause, Gotmare, McCann, Keskar, Joty, Socher, & Rajani, 2020) Second line of works are based on conditioned language models, in which the generated text is conditioned on a context vector (Xia, Zhang, Nguyen, Zhang, & Yu, 2020a). Next group of approaches explore combining variational auto-encoders and input attributes, different types of content attributes (Shu, Papangelis, Wang, Tur, Xu, Feizollahi, Liu, & Molino, 2020; Hu, Yang, Liang, Salakhutdinov, & Xing, 2017; Shu et al., 2020; Bi, Li, Wu, Yan, Wang, Huang, Huang, & Si, 2020). Finally, the last group of approaches utilize planners for structure generation, e.g. (Zhao et al., 2020; Su et al., 2021). Special mention should be made of (Lin & Wan, 2021), which provides ways to improve diversity of generated texts.

Style of generated text is used for wide range of language generated subtasks. Style is usually understood to refer to features of lexis, grammar and semantics. The controlling style can affect of domain-specific dictionaries, sentiment, representation of emotions, personalization. Dathathri et al. (2020) provide a means for controlling topic and sentiment by attributes. Wiseman et al. (2018) consider two question separately "what to say" and "how to say", in which the latter is related to the style of generation. Lastly, Shu et al. (2020) propose controlling style attributes (linguistic register, readability, and many others) by the codebooks.

Concerning the implementation of controllability, there are three main approaches: **Control via hyperparameter**. Language models trained through self-supervision gen-

eralize knowledge from huge amount of texts. These texts can contain unequal (shifted) distribution of training data. In recent research discussed demographic groups, gender and racial biases (Sheng, Chang, Natarajan, & Peng, 2019). Controllable NLG by hyperparameters avoids this issues. Bias metrics and correlated human judgments, and empirical evidence discussed in (Sheng, Chang, Natarajan, & Peng, 2020). The method BTmPG (Lin & Wan, 2021) provided study how the hyper-parameter λ influences semantic guidance to paraphrase model. The GeDi (Krause et al., 2020) is used combination of hyper-parameters for sentiment, detoxification, and topic control generation. **Control via additional input**. This methods based on fine tuning models with extra inputs, such as tables, content for question answering tasks, question, history of dialog. The control code e.g. {*Education, Cars, Climate*} is used in GeDi (Krause et al., 2020) and CTRL Keskar et al. (2019). Metadata from tables is used as additional input in ToTTo (Parikh et al., 2020) for table-to-text generation. Attributes are used in PPLM (Dathathri et al., 2020) for controlling style and sentiment. The hierarchical input is used in DART (Nan et al., 2021) for two controllable properties, namely size and shape. Additional input to regular VAE is used in (Shu et al., 2020) for controlling style of text. Tree or graph as input is used in Zhao et al. (2020). **Conditional training** refers to the group of training methods that employ a discrete control variable to enrich the models with specific capabilities. For instance, approaches use **non-autoregressive models** to reduce the complexity of enforcing constraints at decoding time and include either soft or hard constraints during training. More details can be found in Section 6.1.

5. Fundamentals

5.1 Preprocessing

As for any machine-learning based task, a fundamental first step is to preprocess the input to derive a machine-readable input representation. The type of preprocessing required depends on the form the input takes, which in turn depends on the LG task at hand. A number of LG tasks, such as Machine Translation and Summarisation take language as their primary input. Over the past few years, a number of preprocessing techniques for text have emerged as a result of the turn to neural approaches.

Preprocessing of linguistic inputs primarily involves segmentation into sequences of basic processing units. Determining these basic processing unit typically involves sentence segmentation, tokenization (determining what a (sub)word is), and vocabulary formation. Commonly used choices for units are words, characters, or subword units extracted by methods such as WordPiece (Schuster & Nakajima, 2012), Byte-Pair Encodings (Sennrich & Haddow, 2016) and SentencePiece (Kudo & Richardson, 2018). Subword-like units have become increasingly popular since 2016, where they provided an elegant solution to the problem of out-of-vocabulary words, while providing a more space efficient vocabulary and allowing for open-vocabulary generation. Tokenization has been shown to affect model performance across different evaluation metrics as in Byte-Pair Encodings (Sennrich & Haddow, 2016) and SentencePiece (Kudo & Richardson, 2018) where subword-level tokenization performs better then word-level tokenization for Machine Translation tasks. Other techniques commonly used in the literature for preprocessing include stop-word elimination, stemming and normalization.

Commonly used techniques for preprocessing linguistic units of words involves representing words as one-hot encoded vectors where tokens in the input sequence are transformed to binary valued vectors where each token in the input is mapped to a vector through binarization as 1 indicating existence of the token in the vocabulary, 0 indicating the non-existence of tokens, resulting with with a vector of vocabulary size. Additionally, another strategy that is used in the literature is utilizing a learnable embedding matrix, accomplished by using one-hot encoding followed by matrix multiplication operations. Another common approach used in the literature for preprocessing input sequences consists of using a pre-trained neural network for extracting embeddings from this network as a preprocessing step. In particular linguistic tokens are processed with a pre-trained network where networks such as word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), GloVe (Pennington, Socher, & Manning, 2014), fastText (Bojanowski, Grave, Joulin, & Mikolov, 2017), ELMo (Peters et al., 2018) are commonly used as a feature extractor for input tokens.

Using pre-trained networks for pre-processing other input modalities are common in the literature, *e.g.* pre-processing visual tokens using pretrained CNNs such as Faster-RCNN (Ren et al., 2015b), or ResNet (He, Zhang, Ren, & Sun, 2016) is quite common in the literature where visual tokens are then converted to fixed sized embeddings of these networks. In a similar fashion, raw audio input is commonly preprocessed using a set of approaches. In particular, audio modality typically consists of thousands of samples per second, hence a spectogram representation is preferred which removes the phase information, as it is not informative. The spectogram is used to represent signal activity at different frequencies as a function of time. The frequency content is computed over frames where the length of frame determines the granularity of the spectral details. Each frame is then processed using a Mel-scale filterbank and the logarithm of filter outputs are used to compute log Mel-scale filterbank features providing compression that allows normalization to volume variations by an additive term. Common techniques used to further process the log Mel-scale filterbank features involves processing audio tokens via CNNs to extract embeddings for the input tokens, such as using VGGSound (Chen, Xie, Vedaldi, & Zisserman, 2020) or other pre-trained networks to embed audio tokens to fixed sized feature vectors.

5.2 Models

Recurrent Architectures. The family of Recurrent Neural Networks (RNNs) (Elman, 1990) shown encouraging performance in various NLP tasks such as language modeling and machine translation, and commonly used for NLG tasks typically by generating text in an autoregressive manner. One reason for their widespread adoption is their ease of use in various usage patterns, such as acceptors, encoders, or transducers (Goldberg, Hirst, Liu, & Zhang, 2018). In particular, RNNs are widely used for conditional generation using the encoder-decoder architecture where an RNN is used to encode inputs on the encoder side into a fixed-size vector representation, and another RNN is used to generate outputs on the decoder side, hence converting the (fixed-size) latent space representation into output tokens. The wide adoption of RNNs in the encoder-decoder architecture is also closely related to the success of RNNs in sequence-to-sequence tasks, i.e. a special case of the encoder-decoder architecture where input and output are modelled sequentially. One particular example of a sequence-to-sequence architecture is machine translation, where RNNs

are commonly used to encode a sequence of words in a source language and to decode a translation as a sequence of words in a target language (Sutskever, Vinyals, & Le, 2014; Bahdanau, Cho, & Bengio, 2015; Sennrich, Haddow, & Birch, 2016b) (See Sec.6.1). Another selling point of RNNs is their ability to encode variable-length input sequences into a fixed-length vector embedding including the ability of RNNs to adequately handle long-range dependencies in textual data.

The simplest form of an RNN can then be formulated as follows. For an input sequence $w = (w_1, w_2...w_i)$, an RNN updates its hidden states $h = (h_1, h_2...h_i)$ for each time step and the last state h_i represents the entire input sentence. The tokens in the input sequence are first converted to one-hot vectors which are later transformed into continuous word representations x_i using a word embedding matrix, which is jointly trained with the network. Then, the RNN's hidden state is updated as follows;

$$h_i = f(w_i, h_{i-1}) \tag{1}$$

In Equation 1, f is the function that changes according to the RNN type. The hidden state of the current time-step is dependent on all of the previous hidden states. The initial hidden state h_0 is usually the zero vector $\vec{0}$. Alternatively, the initial hidden state can be learned based on the task, e.g in sequence generation tasks, typically the initial state is set as the final hidden state of the encoder.

Long Short Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) networks were proposed to address the constraints of RNNs caused by vanishing gradients (the exploding gradients problem is not directly addressed with LSTMs and typically alleviated by clipping the gradients at a certain threshold). LSTMs are able to model long-range dependencies using separate hidden and memory cells in the hidden layer, which are capable of storing information for longer time-steps. Each hidden layer contains 3 gates: an input gate i_t which transforms the current input to update the cell, an output gate o_t which controls how much of the cell state to expose in the next hidden state, and a forget gate f_t which controls how much of the information in the previous cell state to forget. In Equation 2, $\odot$ stands for element-wise multiplication; σ is the sigmoid activation function, b is the bias term, W and U are learnable parameters.

$$\begin{aligned}
i_t &= \sigma\left(W^{(i)}w_t + U^{(i)}h_{t-1} + b^{(i)}\right) \\
f_t &= \sigma\left(W^{(f)}w_t + U^{(f)}h_{t-1} + b^{(f)}\right) \\
o_t &= \sigma\left(W^{(o)}w_t + U^{(o)}h_{t-1} + b^{(o)}\right) \\
u_t &= \sigma\left(W^{(u)}w_t + U^{(t)}h_{t-1} + b^{(t)}\right) \\
c_t &= i_t \odot u_t + f_t \odot c_{t-1} \\
h_t &= o_t \odot \tanh\left(c_t\right)
\end{aligned} \tag{2}$$

Gated Recurrent Unit (GRU) (Cho, van Merriënboer, Bahdanau, & Bengio, 2014a) was proposed to transform each recurrent unit to adaptively capture dependencies of different time scales and can be regarded as a simplification of LSTM and its gating mechanism. The gating mechanism in GRU allows flow of information inside the unit without having separate

memory cells. At each time step of the GRU, the activation h_t is a linear interpolation between the previous activation h_{t-1} and the candidate activation $\tilde{h}_t$, modulated by an update gate z_t which controls how much the unit updates its activation, or content and a reset gate r_t which allows to forget the previously computed state when off. In Equation 3, $\odot$ stands for element-wise multiplication; σ is the sigmoid activation function, tanh is the hyperbolic tangent function, b is the bias term, W and U are parameter matrices.

$$
\begin{aligned}
r_t &= \sigma\left(W^{(r)}w_t + U^{(r)}h_{t-1} + b^{(r)}\right) \\
z_t &= \sigma\left(W^{(z)}w_t + U^{(z)}h_{t-1} + b^{(z)}\right) \\
\tilde{h}_t &= \tanh\left(W^{(\tilde{h})}w_t + r_t \odot \left(U^{(\tilde{h})}h_{t-1}\right) + b^{(\tilde{h})}\right) \\
h_t &= (1 - z_t)h_{t-1} + z_t\tilde{h}_t
\end{aligned}
\tag{3}
$$

Attention-based Architectures. RNNs show promising results in sequence-to-sequence tasks, but if the length of the input sequence is large, RNNs often fail to generate a good fixed-sized representation of the input, i.e., summary vector. Experiments using variable-length inputs have shown that the performance of the model dramatically decreases with the increase of the sentence length (Cho et al., 2014a). To overcome this problem and to increase the effectiveness of neural machine translation models, Bahdanau et al. (2015) proposed an attention mechanism to be used in encoder-decoder architectures. The idea behind the attention approach is to represent each token with a vector referred to as an annotation vector rather than representing whole sentence with a single vector. Annotation vectors are later combined into a context vector c which is calculated in each time step. This strategy assists the decoder in attending to different parts of the input sequence while generating different portions of the target sequence.

$$c_j = \sum_{i=1}^{N} a_{ij}h_i, \tag{4}$$

$$e_{ij} = align(h_i, z_{j-1}), \tag{5}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j=1}^{L}\exp(e_{ij})} \tag{6}$$

To generate the j-th target token, a context vector c_j is computed as shown in Equation 4 where N corresponds to the input sequence length, a_{ij} is the attention weight, and h_i, z_j are annotation vectors for encoder and decoder, respectively. Alignment scores e_{ij} can be obtained by the alignment function, which aims to capture the relation between annotation vector h_i in the encoder and the previous hidden state of the decoder z_{j-1}. Attention weights α_{ij} are calculated with the softmax function which converts alignment scores into probabilities (Equation 6), where L corresponding to the target sequence length.

Soft attention mechanisms scan the entire input sequence when generating the output sequence. There are other attention mechanisms proposed in the literature such as monotonic attention for online tasks where the decoder RNN scans chunks of encoder states to

generate output sequences (Chiu & Raffel, 2018; Arivazhagan, Cherry, Macherey, Chiu, Yavuz, Pang, Li, & Raffel, 2019; Merboldt, Zeyer, Schlüter, & Ney, 2019). Recent advances in using attention mechanisms have brought a fully attention-based architecture (i.e., non-recurrent) called the Transformer (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, & Polosukhin, 2017) which achieved the state-of-the-art performance in many NLG tasks (see Section 6 for a broad overview).Transformer models employ encoder-decoder architectures where both encoder and decoder includes multiple stacked layers. The encoder consists of two parts: a multi-headed attention mechanism and a fully connected feed-forward layer. The decoder also has the same sub-layers as the encoder and additionally, there is a separate multi-headed attention layer which attends to the encoder representations. Transformers are the first neural architecture to be completely based on self-attention. The self-attention mechanism captures dependencies between input tokens in the sequence by interacting each token directly with each other token in the sequence. In addition to the self attention mechanism, some Transformer based models also utilize a cross attention mechanism in which tasks involve attending multi modal data on the input side such as VQA or multi modal machine translation, and vision and language grounding problems e.g. (Liu et al., 2019a; Tsai et al., 2019). Because of these "skip-connections" across time, Transformers are often better at handling long-range dependencies (Vaswani et al., 2017). The self-attention mechanism uses a dot-product as an alignment function, which uses transformations of an input $X \in \mathbb{R}^{T \times d}$ into query Q, key K, and value V, where $T_{(.)}$ denotes sequence length and $d_{(.)}$ denotes feature dimension. This attention mechanism executes in a multi-headed manner. First, inputs are projected into keys, queries, values, and then the attention function is applied (Equation 7).

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V. \tag{7}$$

In practice, the Transformer architecture uses multiple attention mechanisms as described in Equation 7 at each time-step, i.e., multi-headed attention. At the end, the output of each attention head is summed to calculate the final context vector C.

$$C = \sum_{i=1}^{h} Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)W_i^O, \tag{8}$$

where W matrices are trainable parameters, and h is number of self-attention heads and $W^Q \in \mathbb{R}^{d_o \times d_k}, W^K \in \mathbb{R}^{d_o \times d_k}$ and $W^V \in \mathbb{R}^{d_o \times d_v}$.

The Transformer architecture does not employ any recurrent or convolutional layers. Therefore, in order to encode the location of tokens in a sequence, the Transformer uses a "positional encoding" which is applied to both input and output sequences. For the positional encoding function, the authors used sine and cosine functions inspired by Fourier series as shown in Equation 9 where p is the position, d is the embedding dimension and i is the model's dimension. The authors also experimented with learned and fixed positional encoding which yields similar results.

$$\begin{aligned} PE_{(p,2i)} &= \sin\left(p/10000^{2i/d}\right), \\ PE_{(p,2i+1)} &= \cos\left(p/10000^{2i/d}\right). \end{aligned} \tag{9}$$

In recent years Transformer-based neural architectures are commonly used as a pre-training scheme for language modeling tasks. In particular, popular architectures such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) use pre-training for downstream language generation tasks. BERT uses pretraining on unlabeled text for extracting deep bidirectional representations by jointly conditioning on both left and right context and is trained using masked language modeling and next sentence prediction. It has become the state-of-the-art in several NLU tasks. Fine-tuning pretrained models like BERT for downstream tasks has become a popular approach in NLP (Xia et al., 2020b). Another transformer based model such as GPT (Generative Pre-Training) (Radford & Narasimhan, 2018) and its variants also utilizes pretraining for both NLG and NLU tasks with generative pre-training of a language model on a diverse corpus of unlabeled text.

The Unified pre-trained Language Model (UniLM) (Dong et al., 2019) architecture integrates multiple language modeling strategies into a joint framework, i.e. it involves standard unidirectional language modelling loss coupled with both a bidirectional and a sequence-to-sequence language modeling task. The combination of different language model objectives makes its learning strategy special, allowing UniLM to achieve high performances for both natural language understanding and natural language generation tasks – including abstractive summarization – as well. BART (Lewis et al., 2020) model formulation also offers a modification to BERT by giving a special emphasis towards natural text generation as well. BART combines a bidirectional encoder over corrupted texts and an autoregressive decoder. The model is trained by first perturbing the input sequences with some noising function (e.g. token deletion and masking), then trying to reconstruct the original text in a sequence-to-sequence manner. Multilingual extensions to BERT based models have been proposed in the literature such as mBART (Liu et al., 2020) and mT5 (Xue et al., 2021).

6. Neural Natural Language Generation Tasks and Applications

6.1 Machine Translation

In machine translation, a model is tasked with generating translations in a target language for sentences given in a source language. Let $X = (x_1, \cdots, x_M)$ be a sentence of length M in a source language L_S and $Y = (y_1, \cdots, y_N)$ be its translation of length N into a target language L_T. In multimodal machine translation (MMT), a model generates a translation Y given not just X but more additional context: an image I that illustrates both X and Y.

Different neural network models for machine translation (NMT) were proposed by Kalchbrenner and Blunsom (2013), Cho, van Merriënboer, Gulcehre, Bahdanau, Bougares, Schwenk, and Bengio (2014b), Cho et al. (2014a), Sutskever et al. (2014) more or less concomitantly. These models follow an encoder–decoder architecture, i.e. an *encoder* converts a source sentence X into a fixed-size dense hidden vector, and a *decoder* generates a translation Y by conditioning on this hidden vector. The use of this fixed-size bottleneck has fallen out of use and was replaced by *attention mechanisms*, which improved NMT models especially when translating longer sequences (Bahdanau et al., 2015; Luong, Pham, & Manning, 2015). In very simple terms, an attention mechanism directly connects each source word hidden state to the decoder at each time step (see Equations 4–6). Moreover, in the industry Google's adoption of NMT (Wu, Schuster, Chen, Le, Norouzi, Macherey, Krikun,

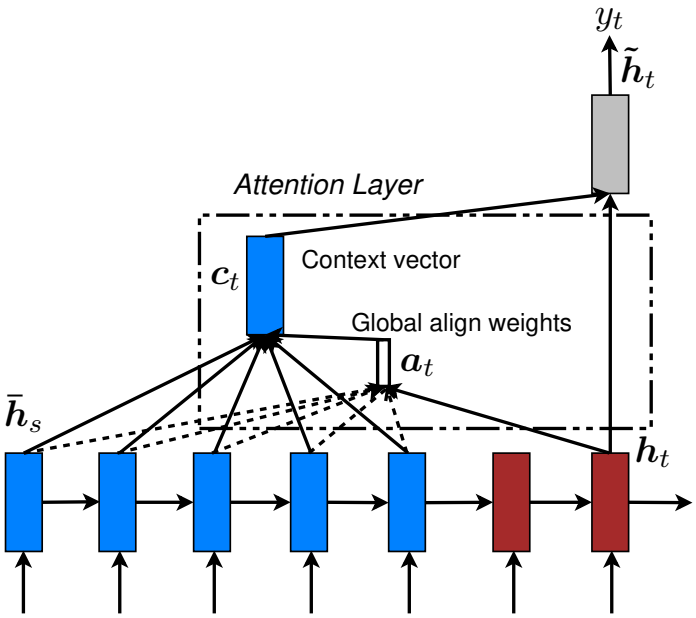

Figure 2: Illustration of an encoder-decoder architecture with attention (Luong et al., 2015).

Cao, Gao, Macherey, et al., 2016) was an indicator that neural networks were overtaking the MT field not only in research labs but also in commercial applications. Uptake of NMT is also evident at the institutional and government level. For example, the European Union's Directorate General for Translation (DGT) began to integrate neural MT technology in its computer-aided translation systems in 2018 (DGT, 2018). In Figure 2 we show an encoder-decoder architecture with attention as proposed in (Luong et al., 2015).

Multimodality. In MMT, the image is expected to be important to *visually ground* the model, and an important catalyst to research in this subarea of MT were the three WMT multimodal translation shared tasks (Specia, Frank, Sima'an, & Elliott, 2016; Elliott, Frank, Barrault, Bougares, & Specia, 2017; Barrault, Bougares, Specia, Lala, Elliott, & Frank, 2018).

The provision of annotated data is an important challenge in MMT. Such data can be expensive to produce. In the most common formulation of the task, training data needs to take the form of triples (X, I, Y). This is necessary since most models for MMT are fully supervised and trained using maximum likelihood estimation. Furthermore, models usually assume that both X and I are available at inference time.

In more sophisticated models, some or all of these assumptions can be relinquished. In **zero-shot and/or unsupervised MMT** the core idea is to alleviate the need for annotated triplets and rely instead on disjoint sets of images and captions in two languages, i.e. (X, I) and (Y, I) (Nakayama & Nishida, 2017; Chen, Liu, & Li, 2018; Su, Fan, Bach, Kuo, & Huang, 2019). However, these models usually perform worse than their fully-supervised counterparts. Zhang, Chen, Wang, Utiyama, Sumita, Li, and Zhao (2020b) propose to train a sentence–image retrieval model using (X, I) pairs, and use it to augment standard

MT training data consisting of pairs (X, Y), with images, yielding (X, Y, I). This allows for the inclusion of images into standard MT pipelines. Zhang et al. (2020b) showed small but consistent improvements compared to fully-supervised Transformer models across language pairs and on large data regimes. Another research line investigates **latent-variable models for MMT**, where the important idea is to better encode the semantics of source sentences using the image (Toyama, Misono, Suzuki, Nakayama, & Matsuo, 2016), and to learn visually grounded models that can translate without images at inference time (Calixto, Rios, & Aziz, 2019). In **domain-targeted MMT**, models are tailored to address common issues in e-commerce and e-fashion areas, such as noisy user-generated content (Calixto, Stein, Matusov, Castilho, & Way, 2017a; Calixto, Stein, Matusov, Lohar, Castilho, & Way, 2017b; Zhou, Cheng, Lee, & Yu, 2018; Laenen & Moens, 2019). An important research line investigates the **image contribution to MMT**, i.e. in which settings images actually help improve translations in MMT models, and also the impact adversarial examples have on multimodal MT (Elliott, 2018; Caglayan et al., 2019; Calixto & Liu, 2019; Dutta Chowdhury & Elliott, 2019). Initial investigations on using **images as context in simultaneous machine translation** have shown promising results, though experiments are still very preliminary (Imankulova, Kaneko, Hirasawa, & Komachi, 2020; Caglayan, Ive, Haralampieva, Madhyastha, Barrault, & Specia, 2020). Lala and Specia (2018) proposed a **translation quality metric specific for MMT** named multimodal lexical translation to try to better estimate the impact of images in visually grounding translations. Finally, in **video-based MMT** models are trained to translate subtitles in the context of videos. Raunak, Choe, Lu, Xu, and Metze (2019) and Wu, Ive, Wang, Madhyastha, and Specia (2019) use the How2 dataset (Sanabria, Caglayan, Palaskar, Elliott, Barrault, Specia, & Metze, 2018) to train video-based MMT models and report favourable results. Sigurdsson, Alayrac, Nematzadeh, Smaira, Malinowski, Carreira, Blunsom, and Zisserman (2020) recently approached unsupervised word translation using video, a different task but somewhat related to MMT.

Multilinguality. Multilingual machine translation models are proposed in cases where the same model is used to translate either from more than one source language (i.e., *many-to-one*), into more than one target language (i.e., *one-to-many*), or both (i.e., *many-to-many*). **Many-to-one** models can directly include multiple source languages by training one separate encoder for each language (Zoph & Knight, 2016), and can also ensemble existing standard NMT models trained for each different source language together (Garmash & Monz, 2016). One main downside of such models is the fact they *require* multiple source languages to be available at both training and test time, though alternatives without these requirements exist (Firat, Sankaran, Al-onaizan, Yarman Vural, & Cho, 2016b; Nishimura, Sudoh, Neubig, & Nakamura, 2020). A related topic involves using additional features available for the source language and/or tasks to improve translation, e.g. dependency parses for the source language (Luong, Le, Sutskever, Vinyals, & Kaiser, 2016; Sennrich & Haddow, 2016; Currey & Heafield, 2019). When translating from one source into many target languages, **one-to-many** models improved translations by simply adding a decoder for each target language (Dong, Wu, He, Yu, & Wang, 2015) and selectively sharing decoder weights between languages (Wang, Zhang, Zhai, Xu, & Zong, 2018). **Many-to-many** models translate from multiple source languages into multiple target languages. Firat, Cho, and Bengio (2016a) and Lu, Keung, Ladhak, Bhardwaj, Zhang, and Sun (2018) use a

shared attention between per-language encoder-decoder networks to achieve many-to-many translation, however the number of parameters grows linearly with the number of languages. This can be alleviated by using a single shared encoder-decoder network to translate between all languages (and a shared vocabulary), which is possible with the use of special language tokens to signal to the model which language to translate from/into (Johnson, Schuster, Le, Krikun, Wu, Chen, Thorat, Viégas, Wattenberg, Corrado, Hughes, & Dean, 2017; Aharoni, Johnson, & Firat, 2019; Freitag & Firat, 2020). The use of this special language token is the current state-of-the-art approach when using multilingual translation models.

Learning strategies. A few promising research directions in neural machine translation worth mentioning include *non-autoregressive*, *simultaneous*, *unsupervised*, and *latent variable* MT. Non-autoregressive models can make MT models much faster, simultaneous MT investigates using MT models in real-time, unsupervised models can increase the applicability of MT into hundreds of languages with at least some monolingual corpora, and latent variable models can acquire knowledge and structural relationships in complex datasets.

Non-autoregressive models do away with the autoregressive factorization of the translation probability $\log p(y_t|x_{1:M}, y_{<t})$ and instead model the generation of each word or token in parallel $\log p(y_k|x_{1:M})$. Important points in many of the proposed non-autoregressive translation (NAT) models include (i) modelling fertility (Della Pietra, Epstein, Roukos, & Ward, 1997) (i.e., how many target words a source word "generates"); (ii) a *knowledge distillation* training phase where an autoregressive model is used to teach the NAT model, which can still be used in parallel at inference time; (iii) the notion of iterative refinements of the translations, so that a first model iteration computes $\log p(y_k|x_{1:M})$, and further iterations refine the previous predictions $\log p(y_k|x_{1:M}, \hat{y}_{1:N})$ (Gu, Bradbury, Xiong, Li, & Socher, 2018; Libovický & Helcl, 2018; Kaiser, Bengio, Roy, Vaswani, Parmar, Uszkoreit, & Shazeer, 2018; Lee, Mansimov, & Cho, 2018; Zhou, Gu, & Neubig, 2020). Recently, (Gu & Kong, 2021) further reduce the gap between fully NAT models, models that do iterative refinement, and autoregressive translation models. **Simultaneous MT models** are design to be used in real-time translation scenarios, e.g. continuously translating utterances as the speaker is talking. Simultaneous MT models must balance latency and translation quality, and many methods have been proposed lately to improve models on both latency and quality dimensions (Ma, Huang, Xiong, Zheng, Liu, Zheng, Zhang, He, Liu, Li, Wu, & Wang, 2019; Arivazhagan et al., 2019; Ma, Pino, Cross, Puzon, & Gu, 2020; Zhang, Zhang, He, Wu, & Wang, 2020a). **Unsupervised MT models** do not assume that parallel training sentences (X,Y) are available. These methods build on unsupervised techniques to induce cross-lingual word embeddings, and leverage denoising auto-encoders, generative adversarial networks (Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, & Bengio, 2014), dual learning (He, Xia, Qin, Wang, Yu, Liu, & Ma, 2016), and back-translation (Sennrich, Haddow, & Birch, 2016a; Edunov, Ott, Auli, & Grangier, 2018) to achieve fully or almost fully unsupervised MT (Artetxe, Labaka, Agirre, & Cho, 2018; Lample, Conneau, Denoyer, & Ranzato, 2018; Marie, Wang, Fujita, Utiyama, & Sumita, 2018; Artetxe, Labaka, & Agirre, 2019; Ren, Zhang, Liu, Zhou, & Ma, 2019; Yang, Chen, Wang, & Xu, 2018). (Garcia, Foret, Sellam, & Parikh, 2020) proposes a method to train multilingual unsupervised MT models, and (Kim, Graca, & Ney, 2020) investigate the limitations of unsupervised MT and find that models still fail to translate properly, especially in cases

where languages L_S and L_T are very dissimilar and there is a domain mismatch between the different monolingual corpora used in the model. Finally, **latent variable MT models** introduce one or more continuous latent variables to capture notions such as model uncertainty and semantic consistency and interaction between source and target texts (Zhang, Xiong, Su, Duan, & Zhang, 2016; Shah & Barber, 2018; Eikema & Aziz, 2019). Related models include Yang, Liu, Xie, Wang, and Balasubramanian (2019), who include a latent variable to model part-of-speech in MT; and Bastings, Aziz, Titov, and Sima'an (2019), who introduce latent dependency parses in the decoder.

Controllability. Controllability in machine translation provides mechanisms to provide humanlike translated text. Explicit control mechanisms include measures to translate special entities, domain-specific dictionaries and other weighted decoding approach. Some approaches use **placeholders** for special entities and replace them through post-processing (Crego et al., 2016; Calixto et al., 2017b). Other enforce tokens in **beam search** at decoding time (Hokamp & Liu, 2017; Hasler, de Gispert, Iglesias, & Byrne, 2018; Post & Vilar, 2018), the same approach proposed for image captioning in (Dinu, Mathur, Federico, & Al-Onaizan, 2019; Yang, Gao, Wang, & Ney, 2020). Similar techniques have also been explored for controlling output length (Murray & Chiang, 2018; Melaku, Di Gangi, & Federico, 2019). Conditional training approaches use **non-autoregressive models** to reduce the complexity of enforcing constraints at decoding time (Susanto et al., 2020), and include either soft or hard constraints during training (Dinu et al., 2019).

In conclusion, we have seen remarkable performance boost in Machine Translation with the introduction of neural MT models. RNN-based models with attention mechanisms were the major driving force behind this progress, though nowadays they are replaced with Transformers-based neural architectures. Multimodal and/or multilingual extensions to these models have also been extensively studied in the literature, but they require a notable amount of paired training data. In this regard, a recent trend in MT community is to utilize less or even no labeled data to relax the supervised formulation of the task, which opens up the possibility of applying novel learning strategies.

6.2 Description Generation

The goal of description generation is to create natural text describing the contents of non-textual input data. Description generation can be categorized into two main types depending on the form of input, i.e., image-to-text approaches that handle visual inputs and data-to-text approaches that operate on non-visual inputs. Here we focus on the first form of description generation, i.e., when the goal is to generate descriptions for either images or videos.

The general goal of description generation from a formal standpoint is to find a mapping $\mathcal{X} \rightarrow Y$, where $\mathcal{X}$ denotes some (typically) non-textual input and Y refers to a description that is supposed to cover the most salient information about $\mathcal{X}$ in the form of a sequence of natural language tokens. $\mathcal{X}$ can take up many different forms, including images and videos (Kiros, Salakhutdinov, & Zemel, 2014; Sun et al., 2019), tabular data (Shahidi, Li, & Lin, 2020; Wang, Wang, An, Yu, & Chen, 2020; Chen, Chen, Su, Chen, & Wang, 2020; Parikh et al., 2020) as well as Resource Description Framework (RDF) triplets (Gardent, Shimorina, Narayan, & Perez-Beltrachini, 2017) or abstract meaning representations (AMRs) (Fan &

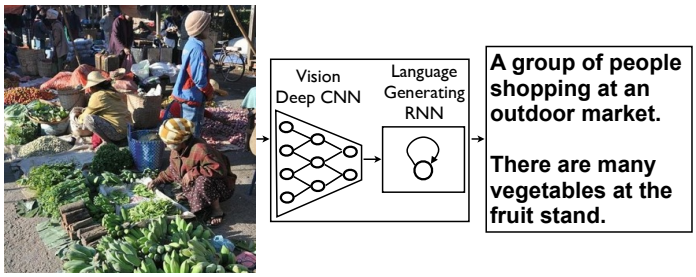

Figure 3: Illustration of the Neural Image Captioning (NIC) image-to-text description generation from (Vinyals et al., 2015).

Gardent, 2020). Here, we focus on image-to-text description generation, that is, we overview approaches that take some visual data as input and turn them into textual output.

Seminal work in image captioning (Kiros et al., 2014; Vinyals et al., 2015; Xu et al., 2015) aims at providing a textual description of input images and it can be considered as the most common form of image-to-text description generation.

Figure 3 illustrates a schematic overview of neural image captioning (Vinyals et al., 2015), a prototypical image-to-text description generation task. The general end-to-end approach, also employed in (Vinyals et al., 2015), is to extract a set of features from a CNN-encoder network and use these features to generate the descriptions with the help of a recurrent neural network. (Xu et al., 2015) demonstrated how visual attention can improve the quality of image captioning. A thorough survey covering additional seminal work related to automatic description generation from images can be found in (Bernardi et al., 2016).

Multimodality. As the core goal of description generation is to generate textual output from non-textual input, multimodality is trivially met for all the approaches discussed in this section. To this end, we decide not to discuss any of the papers in details from this respect. Additionally, as mentioned earlier in Section 4.3.2, multimodality can also refer to the case when a model is trained using inputs of different modalities simultaneously, e.g., (Lu et al., 2019; Tan & Bansal, 2019).

Multilinguality. Miyazaki and Shimizu (2016) created a small version of the MS COCO dataset (Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollár, & Zitnick, 2014) in Japanese. Their best model exploited the English MS COCO dataset in a transfer learning setup, i.e., they first trained an encoder-decoder model that used VGGNet (Simonyan & Zisserman, 2015) for extracting features from images and providing those as initial inputs for the LSTM module generating the captions. After an English model was trained until convergence, the fully connected layer responsible for transforming the image features from the pre-trained VGGNet to the LSTM component was re-utilized for the generation of Japanese captions. Wang, Wang, Zhang, Su, Wang, and Xu (2020) also relied on this kind of cross-lingual transfer learning strategy for caption generation in Chinese.

Lan, Li, and Dong (2017) proposed a model for generating captions in Chinese without relying on image-caption pairs in their target language. The lack of human-provided

captions in the target language was overcame via the application of machine translation. However, since the quality of machine translations are not necessarily satisfying, the authors trained an LSTM model for assigning fluency scores to machine translated captions. The target language image captioning model is then trained following (Vinyals et al., 2015) by jointly relying on the machine translated captions and their predicted fluency scores.

Gu, Joty, Cai, and Wang (2018) dealt with generating image descriptions in English as the target language, while using Chinese as a pivot language. Their model assumed access to gold standard image captions in the pivot language as well as an independent parallel corpus containing translations from the pivot language to the target language, i.e., the model consists of an image captioning and a neural machine translation submodule. The authors experimented with soft parameter sharing between the two tasks and joint training.

The Second Shared Task on Multimodal Machine Translation and Multilingual Image Description (Elliott et al., 2017) featured the task of multilingual image description for English and German. The baseline provided by the organizers was an image captioning system that was trained exclusively on German data in a monolingual setting. The only submission that managed to outperform the performance of the baseline system (Jaffe, 2017) leveraged the multilingual training data in a way that the final hidden layer of the LSTM network producing target language captions was fed into another LSTM as input for generating source language captions (without relying on the actual image as input).

Wang, Wu, Chen, Li, Wang, and Wang (2019) recently released their multilingual and multimodal dataset of video captionings, VaTeX. The dataset contains more than 40K videos and 800K captions in English and Chinese for which data the authors introduced the tasks of multilingual video captioning and video-guided machine translation.

Learning strategies. Zhao, Wang, Ye, Yang, Zhao, Luo, and Qiao (2018) introduced the Multi-task Learning Approach for Image Captioning (MLAIC). MLAIC encompasses a multi-class image classification CNN component for encoding images and a stacked LSTM decoder which is responsible for generating the caption as well as the syntactic structure of the caption. The MLAIC framework is further extended with a reinforcement learning component which uses the CIDEr (Vedantam et al., 2015) scores as its reward function. The framework proposed in (Li & Gong, 2019) builds on top of the encoder-decoder architecture S2VT (Venugopalan, Rohrbach, Donahue, Mooney, Darrell, & Saenko, 2015) for video captioning, and it implements a multi-task reinforcement learning model, where the auxiliary tasks are related to the prediction of attributes extracted from videos.

Bor-Chun Chen and Chen (2017) incorporated video summarization and video captioning into the Video to Text Summary (V2TS) architecture. V2TS is jointly trained to perform summarization, i.e., to extract the most salient frames from a video, and to generate a caption for the summarized frames using a multi-task objective. Video summarization is treated as a regression problem in V2TS, in which the task is to predict the correct importance score of the video frames. The selection of the most salient sequences of frames (shots) from a video is then treated as an instance of the knapsack problem, the objective of which is the maximization of the summed utility of the selected subset of items (shots in this case), such that the constraint originating from the total capacity of the knapsack (the total length of the video summary) is not violated.

A few methods that provide descriptions to videos have already been briefly discussed in Section 4.3.2 on the fundamentals related to multimodality. Here, we provide additional practical details on some of the earlier mentioned architectures. VideoBERT (Sun et al., 2019) extends self-supervised learning via the application of bidirectional Transformers to visual inputs by deriving 'visual words' from video inputs via the application of hierarchical vector quantization on the features derived from videos. Vision-and-Language BERT (ViLBERT) (Lu et al., 2019) follows a similar approach, with the notable difference that instead of incorporating the textual and quantized visual input into a single encoder, it feeds the different input modalities to separate Transformers that interact each other via a co-attention mechanism. ViLBERT is jointly trained for masked multimodal learning and multimodal alignment prediction, i.e., it aims at reconstructing masked words from captions and masked region categories from images, as well as deciding whether a caption is well aligned with the content of the image it belongs to.

Similar to other state-of-the-art image captioning systems, Oscar (Li, Yin, Li, Zhang, Hu, Zhang, Wang, Hu, Dong, Wei, Choi, & Gao, 2020) employs a vision-and-language Transformer architecture. Oscar is based on the observation that salient objects (object tags) are also often included in the image descriptions, hence the training samples are considered as triples, pertaining of a word sequence (the caption itself), a set of object tags originating from the image, and a set of image region features. The model is then trained via the aggregation of a masked token loss and a contrastive loss, which needs to decide if the input caption of an image got perturbed.

Zhou, Palangi, Zhang, Hu, Corso, and Gao (2020) has also demonstrated the utility of large vision-language pre-training (VLP) in both image captioning and visual question answering. The VLP in (Zhou et al., 2020) was trained jointly on two tasks: bidirectional and sequence-to-sequence (seq2seq) masked vision-language prediction.

Controllability. Deshpande, Aneja, Wang, Schwing, and Forsyth (2019) trained a neural image caption generation framework by imposing restrictions on the set of allowed part-of-speech (POS) sequences on the generated descriptions. The (quantized) POS sequence conditioned generation of image descriptions makes the fast generation of diverse captions possible compared to the application of beam search for generation. The architecture proposed in (Cornia, Baraldi, & Cucchiara, 2019) allows for the generation of controllable image descriptions, i.e., the generation process is designed such that it can be grounded on specific regions of an image. The model treats the generated captions as a sequence of chunks generated for the image regions, where the transitions between image regions are controlled by a gating mechanism. This approach offers a radically different way of controllability compared to (Deshpande et al., 2019), as it suggests a way of controllability from the *input* side, by encouraging the generated content to focus on various regions of the input image. On the contrary, (Deshpande et al., 2019) achieved controllability related to the grammatical structure of the generated *output* description. The main difference is that while the former approach affects the generated output implicitly, the latter one does so in a more explicit form.

Another goal of controllability could also be to improve the consistency of the generated descriptions with certain information needs. In that vein, Alikhani et al. (2020) annotated 10,000 image-caption pairs towards their coherence relations. The five coherence categories

(*Visible*, *Subjective*, *Action*, *Story* and *Meta*) investigated were inspired by computational models of discourse. Based on the coherence-annotated dataset, the authors also designed a coherence-aware image captioning framework which is based on an encoder-decoder architecture, using an image feature extractor and an object classifier as input.

Image captioning models typically assume that captions are sampled from a single distribution. In contrast, Fisch, Lee, Chang, Clark, and Barzilay (2020) argue that appropriate captions differ based on the information need of users. They recast the task as one of outputting a caption for an image, such that the caption entails the answer to an (implicit) question-answer pair, using a reinforcement learning paradigm, where the reward is contingent on a caption being contextually appropriate, given some (q, a) pair.

Another frequently studied aspect of controllability relates to the generation of descriptions according to some desired mood, style or personality. The rest of this section reviews such approaches. Context Sequence Memory Networks (Park, Kim, & Kim, 2017) are capable of creating personalized image captions via the incorporation of a memory network component that encourages the model to include words from the active vocabulary of users during description generation. Besides the user context memory, the architecture also includes an image memory and a word memory component that offers an improved treatment of long-range dependencies.

Chen, Pan, Liu, and Sun (2019b) proposed an unsupervised stylish image generation approach. Their network is trained towards the minimization of a joint objective comprising of a standard image description generation loss, i.e., the sum of the cross-entropy of correct description words, and an unsupervised image description reconstruction loss that is intended to learn the idiosyncrasies of generating descriptions according to a particular style. Adaptation to styles unseen during the training phase requires training additional parameters via soft parameter sharing.

Shuster, Humeau, Hu, Bordes, and Weston (2019) introduced the PERSONALITY-CAPTIONS dataset consisting of more than 240,000 images labeled with one of the 215 personality traits, such as *anxious* and *optimistic*. The authors extended three image captioning models (Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018) by learning a personality embedding for each of the personality traits and concatenating those to LSTM decoders during image captioning.

A predominant approach in Description Generation is the application of encoder-decoder architectures, e.g. a CNN for processing images/videos and an RNN generating descriptions. Most recently, the application of vision-language pre-training has gained an increasing popularity and the best performing models are dominantly using joint vision-language Transformer models. Due to the demonstrated effectiveness of large-scale self-supervised pre-training, this trend is likely to continue in description generation as well.

6.3 Automatic Speech Recognition

The primary goal of scientists dealing with Automatic Speech Recognition (ASR) is to transcribe speech into a sequence of words accurately. ASR traditionally consists of two components: an acoustic model $P(S|W)$ that gives probability, that the list of words W sounds like utterance S, and a language model $P(W)$ that is a probability distribution over sequences of words. Recently, much works has been centered around end-to-end models

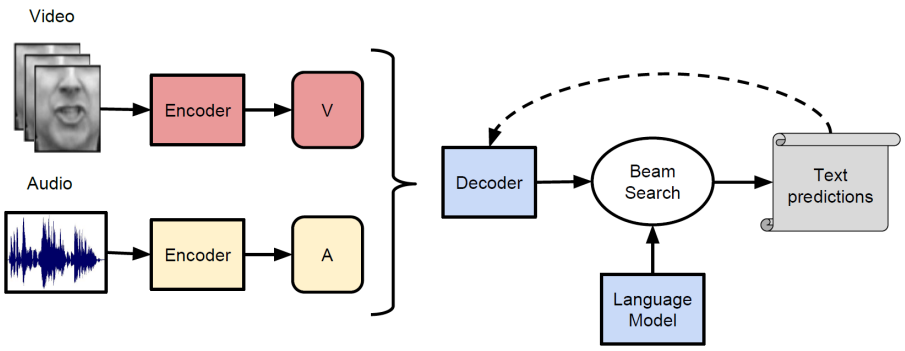

Figure 4: Outline of the audio-visual speech recognition pipeline (Afouras et al., 2018).

models. Such models replace the traditional components of an ASR system with a single end-to-end training.

The first work that showed that DNNs work for large-vocabulary ASR in realistic settings is by Seide, Li, and Yu (2011). The end-to-end systems operate using the connectionist temporal classification (CTC) (Graves, Fernández, Gomez, & Schmidhuber, 2006), the recurrent neural network transducer (RNN-T) (Graves, 2012) or attention-based sequence-to-sequence model (Bahdanau, Chorowski, Serdyuk, Brakel, & Bengio, 2016; Chan, Jaitly, Le, & Vinyals, 2016). Recently, speech recognition with Transformers has been explored (Vaswani et al., 2017).

Multimodality. In multimodal ASR most of the research papers explore the integration of the visual modality. Graphical representation of the audio-visual speech recognition pipeline is given in Figure 4. The use of this modality is focused on improving the performance of ASR under both clean and noisy conditions (Wei, Zhang, Hou, & Dai, 2020; Srinivasan, Sanabria, & Metze, 2019). Commonly, such recognition systems adopt visual features extracted from the speaker's mouth region. These features are distances between the facial markers (Cygert, Szwoch, Zaporowski, & Czyzewski, 2018; Tao & Busso, 2018b, 2018a) or appearance-based features (Fernandez-Lopez & Sukno, 2018; Tao & Busso, 2018b, 2018a). To build knowledge from mouth region features special multimodal corpora are needed (Czyzewski, Kostek, Bratoszewski, Kotus, & Szykulski, 2017; Kawaler & Czyzewski, 2019).

In recent studies, instead of speaker's mouth localization, visual semantic features are considered as an alternative. Examples of these features are objects and scenes that can be automatically detected in the video (Miao & Metze, 2016; Gupta, Miao, Neves, & Metze, 2017). Initial results of end-to-end speech recognizer adaptation to the visual semantic features were presented by Palaskar, Sanabria, and Metze (2018). To adapt a CTC bidirectional LSTM acoustic model and a sequence-to-sequence model, the authors evaluated visual adaptive training and early feature concatenation, respectively. The work of Caglayan, Sanabria, Palaskar, Barraul, and Metze (2019) can be considered as an extension of this research. The authors analyzed the behavior of adaptive training in sequence-to-sequence models. A promising approach is the system based on a RNN-T – regardless of the training

difficulties specific to this architecture (Makino, Liao, Assael, Shillingford, Garcia, Braga, & Siohan, 2019; Braga, Makino, Siohan, & Liao, 2020). Srinivasan et al. (2019) analyzed to what extent auxiliary modalities improve performance over unimodal models, and under what circumstances the auxiliary modalities are useful. Experimental results show that all of the considered multimodal models i.e., hierarchical feature attention, encoder initialization, early decoder fusion, and encoder-decoder initialization considerably outperform the unimodal baseline model (sequence-to-sequence model with attention) on the full unmasked test set. However, these models do not incorporate visual information when the audio signal has been corrupted. This issue remains unresolved. In Sterpu, Saam, and Harte (2020) it was found that Transformers also learn cross-modal monotonic alignments, but suffer from the same visual convergence problems as the LSTM model, calling for a deeper investigation into the dominant modality problem (as outlined in Section 4.3.2), where patterns need to be discovered in the weaker visual signal.

Multilinguality. A multilingual ASR system is an ASR system trained on data from more than one language. Multilingual training is focused on improving the performance of speech recognition, especially for low-resource languages (Yi, Tao, Wen, & Bai, 2018; Alumäe, Tsakalidis, & Schwartz, 2016; Dalmia, Sanabria, Metze, & Black, 2018).

In state-of-the-art multilingual ASR systems, only the acoustic model is multilingual (Thomas, Audhkhasi, Cui, Kingsbury, & Ramabhadran, 2016; Sercu, Saon, Cui, Cui, Ramabhadran, Kingsbury, & Sethy, 2017). Language-specific pronunciation and language models are still required. The hidden layers of the network are trained jointly using data from multiple languages (Sercu, Puhrsch, Kingsbury, & LeCun, 2016; Zhou, Zhao, Xu, & Xu, 2017). To improve the recognition accuracy, acoustic adaptation methods have been proposed (Liu, Wan, Xu, & Zhang, 2018; Tong, Garner, & Bourlard, 2017b).

Nowadays, end-to-end multilingual ASR systems have proved promising in extending multilingual speech recognition because they have simplified training by eliminating the need for linguistic information (Kim & Seltzer, 2018; Cho, Baskar, Li, Wiesner, Mallidi, Yalta, Karafiát, Watanabe, & Hori, 2018; Toshniwal, Sainath, Weiss, Li, Moreno, Weinstein, & Rao, 2018). The models are trained jointly on data from all languages using the union of language-specific grapheme sets. For example, Kannan, Datta, Sainath, Weinstein, Ramabhadran, Wu, Bapna, Chen, and Lee (2019) experiments on multilingual RNN-T, a streaming end-to-end model, which handles a key challenge of real world data, namely, the imbalance in the training data across languages. Results showed that the model significantly outperforms both the monolingual RNN-T models, and the state-of-the-art monolingual conventional recognizers, when language-specific adapter modules were added. To improve language discriminability in end-to-end multilingual ASR systems, language-adaptive training methods have been applied (Liu, Xu, & Zhang, 2020; Toshniwal et al., 2018; Miiller, Stiiker, & Waibel, 2018).

The above-mentioned models have been successfully used in various multilingual ASR tasks. However, the challenge faced by the models is related to building an ASR system that can successfully deal with code-switching scenarios. To solve the system problem of dealing with two or more languages at the same time, a language identification based approach is incorporated (Li, Li, Ye, Zhao, & Gong, 2019; Zeng, Khassanov, Pham, Xu, Chng, & Li, 2018) or a code-switch language model is integrated (Yue, Lee, Yilmaz, Deng, & Li, 2019).

One main challenge for code-switching task is a scarce resource of data, therefore different augmentation techniques are proposed (Long, Li, Zhang, Wei, Ye, & Yang, 2020; Du, Li, Lu, Wang, & Qian, 2021).

Learning strategies. Training a single system to solve multiple tasks in parallel is considered an effective method used to improve speech recognition performance (Tang, Li, & Wang, 2016). Multi-task learning can be divided into two categories: monolingual and multilingual. Within monolingual multi-task learning, network training is performed by paying attention to the auxiliary tasks or via ignoring them by using the adversarial learning (Meng, Li, Chen, Zhao, Mazalov, Gang, & Juang, 2018; Sun, Yeh, Ostendorf, Hwang, & Xie, 2018). The multi-task learning is incorporated by using additional linguistic features (Toshniwal, Tang, Lu, & Livescu, 2017; Pironkov, Dupont, & Dutoit, 2016) as well as by using speaker (Saon, Kurata, Sercu, Audhkhasi, Thomas, Dimitriadis, Cui, Ramabhadran, Picheny, Lim, Roomi, & Hall, 2017) or recording (Serdyuk, Audhkhasi, Brakel, Ramabhadran, Thomas, & Bengio, 2016; Shinohara, 2016) information. The CTC objective function as an auxiliary task is used in Kim, Hori, and Watanabe (2017). In multilingual multi-task approach, each language is considered as a different task (Bukhari, Wang, & Wang, 2017).

More recently, transfer learning has gained popularity in ASR. Transfer learning is widely applied in multilingual ASR to improve the performance on target languages by learning to share model parameters across languages (Kim & Seltzer, 2018; Cho et al., 2018; Popović, Pakoci, & Pekar, 2019; Joshi, Zhao, Mehta, Kumar, & Li, 2020). An adaptation of transfer learning for child speech recognition is given in (Tong, Wang, & Ma, 2017a; Matassoni, Gretter, Falavigna, & Giuliani, 2018; Shivakumar & Georgiou, 2020). In multimodal speech recognition transfer learning implies pre-training the visual front-end on word excerpts (Petridis, Stafylakis, Ma, Tzimiropoulos, & Pantic, 2018; Afouras et al., 2018).

With regards to learning strategies, there have been studies examining knowledge distillation to improve the performance of ASR system. In Xu, Hou, Song, Guo, and Dai (2019) several same sized encoder-decoder models are selected as multiple teacher models including multilingual teacher and monolingual teachers for each language. An investigation of the effectiveness of knowledge distillation in the context of multilingual models is given by Cui, Kingsbury, Ramabhadran, Saon, Sercu, Audhkhasi, Sethy, Nussbaum-Thom, and Rosenberg (2017).

A review of the literature revealed that a lot of effort is currently being made to provide multilingual and multimodal research on ASR. Most of the papers on multimodal ASR deal with the integration of the visual modality, which is realised as features of the mouth region or more recently as visual semantic features. Multilingual training is dealing with data from more than one language and usually involves tasks in low-resource languages. As regards the issue of achieving high accuracies in speech recognition, learning strategies such as transfer-learning and knowledge distillation are used.

6.4 Abstractive Summarization

The aim of summarization is to condense one or more information sources while preserving their relevant content and meaning. In this manner, depending on the input, sources of different types can be summarized. Of these, text, video, and images are among the most

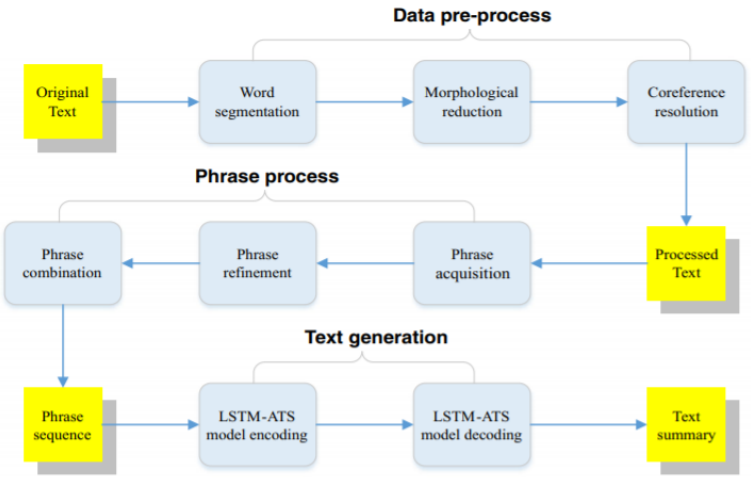

Figure 5: Outline of a standard abstractive summarization pipeline (Song et al., 2019).

common. Abstractive summarization is a technique in which the summary is generated by creating novel sentences, instead of simply extracting the important sentences verbatim (Gupta & Gupta, 2019). This is done by either rephrasing or using new words, and it helps to increase the quality of summaries produced from textual data, as well as the ones where non-textual or cross-language data are given as input. Multimodal abstractive summarization is closely related to description generation (see Section 6.2), with the main distinction being that for abstractive *text* summarization, the primarily input – as well as the expected output – is textual, whereas description generation deals with non-textual input.

Works that exclusively focus on abstractive summarization, paying attention to the recent developments concerning to the use of neural networks can be found in (Gupta & Gupta, 2019), (Lin & Ng, 2019), (Baumel & Elhadad, 2019) and (Shi, Keneshloo, Ramakrishnan, & Reddy, 2021). Figure 5 and Figure 6, both extracted from (Song, Huang, & Ruan, 2019), show the standard process description and an encoder-decoder architecture for generating abstractive summaries, where the input consists of phrases instead of words.

To define abstractive summarization more formally, let $S = (s_1, \ldots, s_N)$ be a collection of information sources (e.g., a single document, a collection of documents, a collection of facts) of variable length L, and let R be the set of relevant information units in S. An abstractive summary Abs is a generated text of length $|Abs| < L$, such that Abs maximises the coverage of R, while expressing it in a different way from S.

Multimodality. Abstractive summarization can be said to be multimodal whenever the summarization algorithm is applied to information in more than one modality, either in the input or the output (e.g., video, text, images, among others). To some extent, multimodal summarization can be related to description generation (explained in Section 6.2), i.e. the most salient information from the input has to be covered in the generated output. An important difference between the two tasks is that for multimodal abstractive summarization at least part of the input is textual, whereas for description generation, this is not the case. Focusing on multimodal abstractive summarization, we can distinguish between

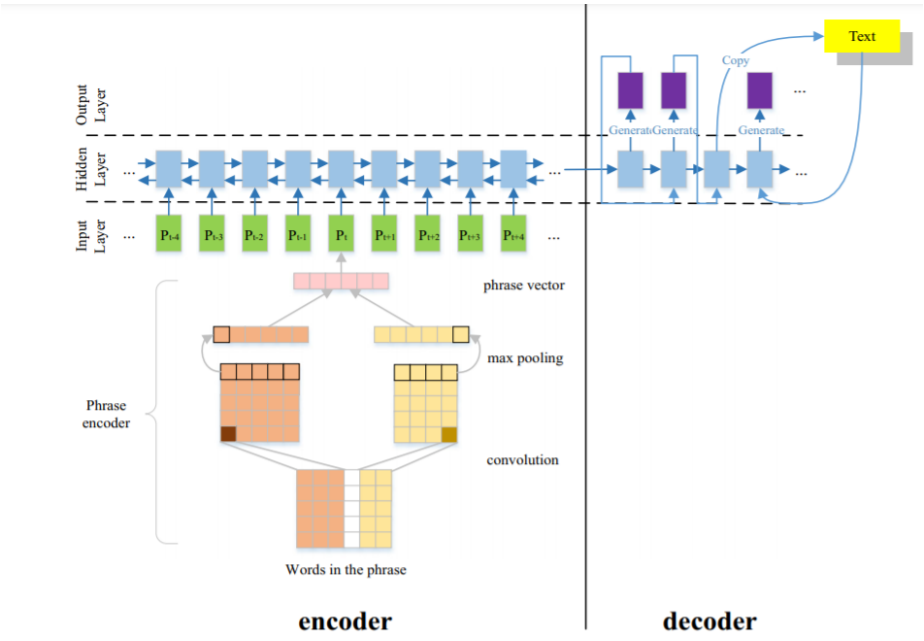

Figure 6: Example of an encoder-decorder architecture for abstractive summarization (Song et al., 2019).

multi-modal to text, and multi-modal to multi-modal scenarios. Concerning the former (i.e., multi-modal to text), Li, Zhu, Ma, Zhang, and Zong (2017) proposed a multimodal summarization method that can automatically generate a textual summary given a multimodal input containing documents, images, audio and videos. The authors further introduced the MMS corpus.[16] Similarly, in order to produce a short textual summary from a pair consisting of a sentence and an image, Li et al. (2018) improved the MMS corpus, and constructed the MMSS dataset[17] for the task of multimodal sentence summarization. The How2 dataset (Sanabria et al., 2018) collects instructional videos about different topics. This dataset has been used for video summarization in several reseach works (Libovický et al., 2018; Palaskar et al., 2019). In these previous works, the multimodality aspect in the input is limited to images (including video frames) and text, however, in Li et al. (2017), video transcripts are also used for dealing with the audio/speech.

As for multi-modal to multi-modal summarization, for the problem of text-image-video summary generation (TIVS) from multimodal input, Jangra, Jatowt, Hasanuzzaman, and Saha (2020) created a text-image-video dataset by extending and manually annotating the previously developed MMS dataset (Li et al., 2017). In (Chen & Zhuge, 2018), the DailyMail dataset[18] was extended by collecting images and captions from the Web, and then text and images are used simultaneously as input and output for the summary generation process. The proposed architecture relies on an attentional hierarchical Encoder-Decoder, where in

16. MMS corpus, `http://www.nlpr.ia.ac.cn/cip/ZhangPublications/emnlp_en.htm`
17. MMSS dataset, `http://nlpr-web.ia.ac.cn/cip/ZhangPublications/ijcai2018-en.htm`
18. https://github.com/abisee/cnn-dailymail

the encoding stage, a bi-directional RNN and a CNN network are used for the text and the images, respectively. Also using DailyMail dataset as a basis, in (Zhu et al., 2018), some pictorial summaries were first annotated. Then, a model based on a Pointer-Generator Network (See, Liu, & Manning, 2017) is extended consisting of four modules: (i) text encoder (a BiLSTM), (ii) image encoder (the VGG19 model pretrained on ImageNet) to extract global or local features from images, (iii) multimodal attention layer, which aims to fuse textual and visual information during decoding stage; and finally a (iv) summary decoder, through a unidirectional LSTM. The authors in (Zhu et al., 2018) propose a novel multimodal automatic evaluation (MMAE) method achieving a better correlation with human judgements of multimodal summaries.

Multilinguality. When multilinguality is the main aim of the summarization process, many researchers have used off-the-shelf methods and pretrained language models in order to examine their influence on the task of abstractive summarization. Aksenov et al. (2020) conditioned the encoder and decoder of a Transformer-based neural model on the BERT language models for English and German.

When the input and output language is different, the task is seen as cross-lingual summarization, where traditionally machine translation and summarization have been integrated and combined into the same approach, by either first translating the original document into the target language and then summarizing it, or vice versa Shreve (2006), Wan, Li, and Xiao (2010), Grega, Smaïli, Leszczuk, González-Gallardo, Torres-Moreno, Pontes, Fohr, Mella, Menacer, and Jouvet (2018). Zhu et al. (2019) propose a different approach, where the task of cross-lingual summarization is addressed end-to-end with the support of automatically constructed corpora. In this case, the languages involved are English and Chinese.

In contrast, other researchers have been proposing language-independent methods that were later evaluated on datasets in various languages for the purpose of experimenting with different languages. This is the case of the research proposed by Li et al. (2017), where the multilingual MMSS corpus, available for English and Chinese, was used for evaluation. As previously mentioned, this corpus is also multimodal. In order to reconstruct the title of Wikipedia articles in which the title, subtitle and the summary were masked, Liu et al. (2019c) evaluated their proposed approach in 42 languages.

Learning strategies. In some cases, the summarization task is addressed by relying on other tasks (e.g. textual entailment), or by joining training extractive and abstractive summarization strategies.

Specifically, for the task of sentence summarization that creates a condensed version of a long source sentence, Li et al. (2018) proposed an entailment-aware encoder under a multi-task framework comprising summarization generation and entailment recognition. Chen et al. (2019) proposed a general unified framework for abstractive summarization which incorporated extractive summarization as an auxiliary task, by constraining the attention learned in the abstractive task with the labels of the extractive task to strengthen the consistency between the two tasks.

Pasunuru et al. (2017) employed a multi-task learning setting for increasing the consistency of the summaries generated by the decoder of their neural sequence-to-sequence architecture. This was achieved by sharing the parameters of their decoder with an encoder and applying them for the generation of such output sentences for which the entailment

relation holds relative to some premise sentence from the SNLI (Bowman et al., 2015) natural language inference dataset. The model proposed by Guo et al. (2018) also incorporates an encoder for the entailment generation component, however, it was additionally trained towards question generation as well, while advocating soft parameter sharing between the primary task (abstractive summarization) and the auxiliary learning objectives (question generation and entailment generation). Xu et al. (2020) introduced an encoder-decoder architecture by defining such a multi-task learning framework which also involves loss terms for extracting key sentences and keywords. Their pointer-generator architecture benefits from the auxiliary losses, which encourage the summarizer to cover the most salient parts of the input documents in the generated summary.

Controllability. Controllability in abstractive summarization includes mechanisms that enable control of important aspects of a generated summary contributing to the generation of content tailored to user needs. The aspects of the generated output are the length of summary (Fan et al., 2018; Liu et al., 2018; Amplayo & Lapata, 2021; Wang et al., 2020), entities in the focus of interest (Fan et al., 2018), the style of generated text (Fan et al., 2018), sentiment or politeness of summarized text (Amplayo & Lapata, 2021), the desired content (Amplayo & Lapata, 2021) or summarization of desired parts of inputs (i.e. the remaining part of text) (Fan et al., 2018).

The method by Kikuchi et al. (2016) provided early motivation for work on controllable abstractive summarization. They achieve explicit control of desired output length either in the encoder during training, or in the decoder during generation. Important principles on the length controllability in abstractive summarization were subsequently laid down by (Fan et al., 2018). The control of summary length is enabled with a quantized length marked during learning. The output length, in the form of a five-word headline, single sentence or paragraph, is controlled by a marker token and improves the quality of generated output.

Entity-centric summarization sets the focus of a generated summary on a person, location or organization. During generation the desired entity token is prepended to the input contributing to the quality of generated content even when several entities are included in generation. Similarly, an extra marker is used to represent the source style and used as the marker for generated content.

In the work on the summarization from only selected parts of text authors in (Liu et al., 2018) extended the standard model with the explicit coding of the summary length provided from gold standards during training and controlled as input parameter during generation. Specifically, the model learns the probability of generating an EOS (end of summary) token, hence has the ability to use its internal state for the definition of the output length without sacrificing semantic information or lowering ROUGE score[19]. In (Amplayo & Lapata, 2021) a two-stage *condense-abstract* method is proposed. In the first stage all documents (opinions) are encoded, condensing their meaning and distilling information like sentiment or aspects. In the second stage condensed representations are aggregated into a multi-source fusion model. The generated summary can be controlled for the polarity of the reviews or the aspects covered.

19. ROUGE is a well-known evaluation metric for automatic summarization that measures the content overlap between an automatic summary and a reference one.

Although a lot of progress has been done in the context of Abstractive Summarization, there is still a room for improvement of the generated content quality and coherence. Regardless whether the studied aspect is multimodal, multilingual or controlled the final summary should take advantage of pre-trained models that work well for other text generation tasks. Specifically, multimodality is mainly studied as the multi modal input (i.e. text, video, images) and either textual or multimodal output. The reported progress is indicating that multimodality is beneficial for abstractive summarization performance. It seems that mulitilinguality remains the main challenge in abstractive summarization task, while reported research is mainly relaying upon the machine translation to provide multilinguality of the summaries and crosslingual summarization is scarcely addressed. The aspect of controllability is propelling a research in abstractive summarization with either controlled length, content or polarity, achieving better semantic content without deterioration of the performance and enabling better coherence of generated content. In all cases, coherent, truthful and useful summaries must be guarantee to avoid, for instance, misinformation problems.

6.5 Text Simplification

Text Simplification is the process of reducing the linguistic complexity of a text, while retaining the original information content and meaning (Saggion, 2017). Text Paraphrasing is defined as an approximate equivalence of meaning across sentences or phrases (Bhagat & Hovy, 2013). The main differences between Text Simplification and Text Paraphrasing lies in their handling of text complexity and in the aim of the text reducing task. Figure 7 illustrates a Text Simplification pipeline that uses word embeddings to extract candidate terms and Text Paraphrasing to generate simplified texts (Maddela et al., 2021).

A neural simplification model with global attention and input feeding was proposed in (Nisioi, Štajner, Ponzetto, & Dinu, 2017). This approach uses a Sequence to Sequence (Seq2Seq) Neural Network to model text simplification in order to simultaneously perform lexical simplification and content reduction. Human evaluation showed that the Seq2Seq-based systems can significantly outperform the best phrase-based and syntax-based Machine Translation approaches.

A system that performs transformations, not only at the lexical and syntactic levels but also on the discourse level, was proposed in (Štajner & Glavaš, 2017). The proposed system pipeline combines event-based simplification with lexical simplification. The event-based simplification module is based on an event extraction system, i.e., EvGraph, and uses only 11 rules to perform sentence splitting and deletion of irrelevant sentences or sentence parts. The lexical simplification module uses LightLS which leverages word embeddings trained on a large (standard English) corpus, thus not requiring any parallel or comparable Text Simplification corpora. This approach leads to significantly more content reduction within a sentence and within a text, managing to even delete whole sentences.

Some systems try to solve the text simplification task using a paraphrasing approach (Xu, Napoles, Pavlick, Chen, & Callison-Burch, 2016). Thus, given an input text, the task is to rewrite it, with the aim that the output should be simpler than the input, while preserving as much of the meaning of the input as possible, and maintaining the well-formedness of the text. To achieve these, the authors modified four key components of a syntax-based

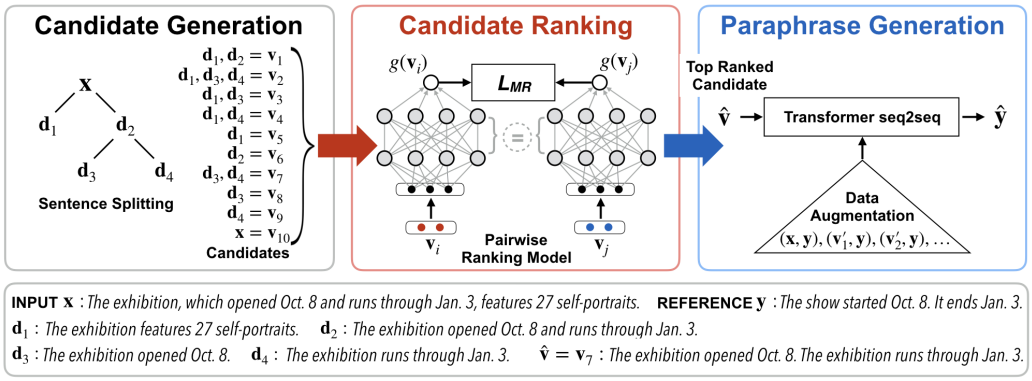

Figure 7: Illustration of a Controllable Text Simplification with Explicit Paraphrasing Pipeline after (Maddela et al., 2021).

machine translation framework using: 1) two novel simplification-specific tunable metrics; 2) large-scale paraphrase rules automatically derived from bilingual parallel corpora, 3) rich rule-level simplification features; and 4) multiple reference simplifications collected via crowd-sourcing for tuning and evaluation.

In Mallinson, Sennrich, and Lapata (2017), the authors propose a neural model that is able to automatically extract paraphrases from bilingual corpora to find meaning-equivalent phrases in a single language by pivoting over a shared translation in another language. The method represents paraphrases in a continuous space and then either estimates the degree of semantic relatedness between text segments of arbitrary length, or generates candidate paraphrases for the source input. The evaluation showed that the proposed neural approach outperforms the conventional phrase-based pivoting approaches. The proposed model is tested on a multilingual dataset that contains texts in English, French, German, and Czech. The model is evaluated for two distinct tasks: paraphrase similarity and sentence-level paraphrase generation. Furthermore, the model employs a novel Neural Machine Translation approach to train paraphrases which exploits a one-to-one NMT architecture: the source English sentence is translated into k candidate foreign sentences and then back-translated into English.

In (Chu, Otani, & Nakashima, 2018), the authors propose a model for extracting visually grounded paraphrases, i.e., different phrasal expressions describing the same visual concept in an image. Thus, given an image and all the entities in the corresponding captions, the task is to cluster the entities to their corresponding visual concepts, represented as image regions. The method applies different unsupervised similarity computation methods in combination with a supervised neural network method using both textual and visual features to explicitly model the similarity of an image segment and a potentially describing sentence. Experiments showed that the proposed neural network approach outperformed the other methods in the literature.

Multimodality. Text Simplification and Paraphrasing become multimodal when the input data is different than the output data, e.g., the input is a video while the output is

a textual description. For example, in the medical domain, datasets contain medical imagery, medical reports, etc. In this case, Text Simplification techniques are used to create a simplified textual description that the patients can understand. Another application of Multimodal Paraphrasing is in the sensemaking and sensegiving analysis of events presented using different mediums. Thus, verbal text, images, and other visual artifacts constitute a key resource for in this type of analysis (Höllerer, Jancsary, & Grafström, 2018).

Multimodal Text Simplification and Paraphrasing is usually encountered when using a dataset that is composed of heterogeneous data. Thus, multimodality increases the difficulty of the problems related to the complexity metric definitions, because different modalities add more aspects to subjective assessment of the information comprehension. Paraphrase is more broadly defined and methods for multimodal text paraphrase generation can be used for text simplification to decrease complexity.

Such is the case of (Chu et al., 2018), in which the authors tried to extract different phrasal expressions describing the same visual concept in an image provided as input. Another approach used for paraphrasing is to combine visual features extracted from medical images with textual data from the associated reports (Li, Liang, Hu, & Xing, 2019). Furthermore, the authors of (Liu, Tang, Wan, & Guo, 2019b) also use visual features generated from images to select textual paraphrases and then give captions to images.

Multilinguality. In the context of of Text Simplification and Paraphrasing, multilingualism is usually present in a way that a proposed language-independent approach is experimentally applied for datasets in different languages. Such is the case of (Mallinson et al., 2017), in which the authors tried to recognize and to generate paraphrases in the context of neural machine translation for English, French, German and Czech. Similarly, Sjöblom, Creutz, and Aulamo (2018) performed automatic paraphrase and semantic similarity detection on subtitle data for German, English, Finnish, French, Russian, and Swedish. A benchmark for multilingual paraphrasing is PAWS-X (Yang, Zhang, Tar, & Baldridge, 2019) that can be used to evaluate paraphrase pairs in French, Spanish, German, Chinese, Japanese, and Korean. But, the majority of systems that use a multilingual approach only employ two languages, e.g., Chinese and English (Li et al., 2019), Korean and English (Park, Oh, Choi, & Gweon, 2019), Japanese and English (Kajiwara & Komachi, 2016).

Learning Strategies. For the problems of Text Simplification and Paraphrasing, multitask learning represents the use of the same input data in order to train a model to solve different tasks on that data. The multitask learning approach injects linguistic-based inductive biases in order to improve simplification and paraphrasing quality. Thus, multitask learning tackles multiple tasks (e.g., translation, part-of-speech tagging, and named entity identification) using shared components of the same model.

One approach used to extract paraphrases for an image is to employ multitask (Chu et al., 2018) training by performing coreference resolution, i.e., to find the expressions that refer to the same entity in a text, and phrase localization, i.e., to find an image region that corresponds to a given phrase in a caption.

Training on several tasks is used to improve performance on each individual task. In (Xu et al., 2016), the authors adapt a syntax-based machine translation framework for paraphrasing to perform text simplification.

Controllability. Explicit control, in the context of Text Simplification and Paraphrasing can be implemented by adding some specific metadata to the training process, e.g., special tokens to represent specific grammatical attributes (Martin, Éric de la Clergerie, Sagot, & Bordes, 2020), domain specific dictionaries (Nassar, Ananda-Rajah, & Haffari, 2019), GloVe word embeddings to find semantically most similar candidates for any input word (Štajner & Glavaš, 2017).

To that end, Kajiwara and Komachi (2016) automatically built a monolingual parallel corpus for Text Simplification considering sentence similarity based on word embeddings. In order to determine how semantically similar sentences are, Sjöblom et al. (2018) experimented with sentence encoding models that take as input a single sentence and produce a vector representing the semantics of the sentence. Similarly, Nishihara, Kajiwara, and Arase (2019) considered levels of both sentences and words, in order to control both the lexical and syntactic complexity and achieve an aggressive rewriting for the Text Simplification problem. In (Nisioi et al., 2017), controllability is achieved by minimizing the vocabulary. Furthermore, Word2Vec skip-gram model is employed to add context.

In conclusion, Text Simplification is used to reduce the linguistic complexity (e.g., remove redundant information, split large sentences into smaller sentences, add explanations for domain-specific terms) while retraining the original information and meaning. Text Paraphrasing is used to create approximate equivalent sentences. Usually, Text Simplification uses Text Paraphrasing to replace hard to understand words with simpler, more common, words, and it employs multimodality to explain, using a simpler and easy to understand language, different kinds of data, e.g., images, videos. For both tasks, multilingual approaches are used for proposing language-independent methods, while, multitask learning is used to tackle multiple tasks at the same time, e.g., translation, part-of-speech tagging, named entity identification. To achieve an aggressive rewriting for both Text Simplification and Paraphrasing tasks, controllability is used to control both lexical and syntactic complexity.

6.6 Question Answering and Generation

6.6.1 Question Answering (QA)

Within the literature, there has been extensive research on question answering. Most of the works on question answering (QA) cast the problem considering different formats such as span-selection, Yes/No, multiple choice, etc. In that regard, the models do not directly produce an answer using natural language generation schemes but rather select the most likely answer from the existing answer set.

Multimodality. In the recent years, there has been a surge of interest in the area of Visual Question Answering (VQA). The task of VQA is to search for visual clues in the image related to the question, making this an inherently multimodal task.

Recent extensive work on VQA was triggered by the availability of large image and video question answering datasets. The COCO-QA (Ren, Kiros, & Zemel, 2015a) large-scale object detection, segmentation, and captioning dataset was used as training data by Wu et al. (2019). Samples from the VQA v2.0 dataset[20] (Antol, Agrawal, Lu, Mitchell,

20. VQA v2.0 dataset, `https://visualqa.org/`

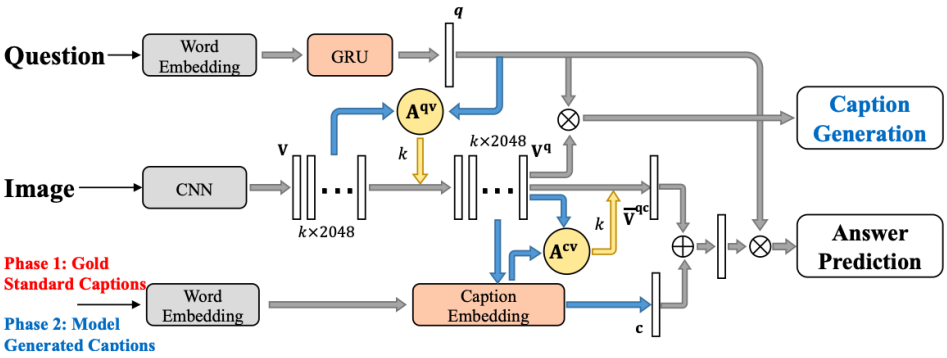

Figure 8: Illustration of the Visual Question Answering, neural caption generation to aid answer prediction after (Wu et al., 2019)

Batra, Zitnick, & Parikh, 2015) containing open-ended questions about images were utilized as training instances for visual question answering by Wu et al. (2019), Huang, Huang, Guo, Qiao, and Zhu (2019), Wu and Mooney (2019). Figure 8 illustrates a schematic overview of neural question answer generation by generating question relevant captions (Wu et al., 2019).

In order to provide salient image regions with clear boundaries, the Visual Genome dataset[21] (Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, & Fei-Fei, 2017) was included as a part of traning procedure by various researchers (Wu et al., 2019; Huang et al., 2019; Kim & Bansal, 2019). For the specific application of Textbook Question Answering (TQA), Kembhavi, Seo, Schwenk, Choi, Farhadi, and Hajishirzi (2017) developed the TQA dataset[22] drawn from middle school science curricula. This dataset was later used to face the challenge of multimodal contexts comprehensions in complicated input data contained in text books (Kim, Kim, & Kwak, 2019).

As a starting point, various authors relied on the Top-Down and Bottom-Up attention model[23] (Anderson et al., 2018) that was further enhanced by introducing a non-uniform image region based grounding, modules and information. In order to discover the most relevant image regions for the question, Huang et al. (2019) tried to identify instances of objects with more general bounding boxes belonging to certain categories, using the existing Bottom-up model. Kim and Bansal (2019) adopted this model to weight visual features from the obtained, salient areas. Motivated by the need to create an explainable VQA system with transparent reasoning, Wu and Mooney (2019) tried to avoid emphasis on irrelevant portions of the image by enhancing the Bottom-up model.

Learning Strategies. Very recently there has been an interest in adopting pre-trained autoregressive models for question answering while incorporating natural language generation for answering questions. Nishida, Saito, Nishida, Shinoda, Otsuka, Asano, and Tomita

21. Visual Genome dataset, `https://visualgenome.org/`
22. TQA dataset, `http://textbookqa.org`
23. Top-Down and Bottom-Up attention model, `https://github.com/peteanderson80/bottom-up-attention`

(2019), proposed a multi-style abstractive summarization model that generates answers based on textual evidence. Dunietz, Burnham, Bharadwaj, Rambow, Chu-Carroll, and Ferrucci (2020), proposed a framework to test questions answering models' ability to comprehend using a large Transformer-based language model pretrained on short narratives. Bi et al. (2020), proposed an autoencoding and autoregressive language model with pre-training on a large corpus for natural language generation tested against question answering task. Khashabi, Min, Khot, Sabharwal, Tafjord, Clark, and Hajishirzi (2020), used a generative text-to-text neural model to build a single pre-trained question answering model. Chen, Stanovsky, Singh, and Gardner (2020), proposed a dataset for training and evaluating generative question answering metrics and poses question answering as a language generation problem and proposed a Transformer-based benchmark model to mimic human judgments for the generative question answering problem.

6.6.2 Question Generation (QG)

Automatic QG is motivated by the need to have a mechanism able to construct syntactically sound, semantically correct and relevant questions from various modalities.

Multilinguality. The application of QG was tackled from the cross-lingual perspective (Kumar, Joshi, Mukherjee, Ramakrishnan, & Jyothi, 2019). The authors reused an available large QG dataset in a secondary language to learn a QG model for a primary language for Hindi/English and Chinese/English language pairs. Another line of work in multi-lingual setup leverages pre-training to improve the model performance for question generation in a cross-lingual setting, in which pre-training encourages the model to represent different languages (English, Chinese and French) in the same space (Chi, Dong, Wei, Wang, Mao, & Huang, 2020).

Multimodality. QG is extended to a multimodal setup in which the problem is formulated as asking natural and engaging questions when shown an image (Mostafazadeh, Misra, Devlin, Mitchell, He, & Vanderwende, 2016). (Jain, Zhang, & Schwing, 2017) tackled with the problem of generating a diverse set of natural questions for a single input image, where they investigated diverse question generating problem in the context of creativity. Shijie, Lizhen, Shaodi, Zhenglu, and Jiawan (2017) addressed the problem of automatically generating visually grounded questions with the goal of generating different types of target questions for a single input image. (Liu, Xiang, Hospedales, Yang, & Sun, 2018) posed the question generation problem as a multimodal dynamic inference process and proposed the inverse visual question answering task, where the goal is to generate a set of questions using textual answers as input along with a single image, and the target is a natural question sharply conditioned on the answer. (Sarrouti, Abacha, & Demner-Fushman, 2020) explored visual question generation task in the medical domain with a model to generate visual questions for radiology images. Figure 9 illustrates a schematic overview of neural question generation from a visual input (Shijie et al., 2017). Recent work on question generation has focused on various aspect of generating questions such as informativeness and usefulness such as (Krishna, Bernstein, & Fei-Fei, 2019) where authors formulate the question generation task as a goal-driven information maximization problem and generate visual questions that maximize the likelihood of receiving an answer. Patro, Kumar, Kurmi, and Namboodi-

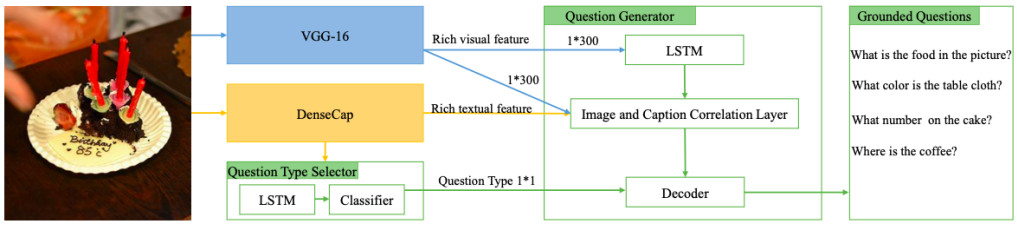

Figure 9: Illustration of the Neural Question Generation, generation questions for a visual input after (Shijie et al., 2017)

iri (2018) used the difference between relevant and irrelevant exemplars in a multimodal question generation setup to generate natural and engaging questions.

Learning Strategies. Another line of work approached question generation task in a multi-task setup, where a model is trained for QG jointly with other tasks. In particular, Wang, Yuan, and Trischler (2017) showed that the QG task can be used to improve QA task performance when jointly trained for a model that has partial extractive and abstractive generation capacity. (Shah, Chen, Rohrbach, & Parikh, 2019) exploited cycle consistency in question answering and question generation to improve model robustness and performance for both question answering and question generation tasks on VQA v2.0 dataset. Li, Duan, Zhou, Chu, Ouyang, Wang, and Zhou (2018) proposed a dual learning framework and showed that both question answering and question generation task performance can be improved when models are trained jointly.

Controllability. The question generation problem has also been investigated in setups in which the QG model is optimized towards generating human-like questions by controlling the model's outputs with another modality, that is the type of questions such as "who", "where", "when", "why", "which", "what", "how", "yes-no", and "other" framed as style, and also a textual hint framed as clue by Liu, Wei, Niu, Chen, and He (2020).

In conclusion, the current efforts in Question Answering are mainly in answer span extraction rather than natural language generation. Although neural Question Generation (QG) received some interest lately from the community, there are still many challenges and opportunities in adapting neural NLG models to produce useful and informative questions as there are not so many works that deal with the automatic generation of questions from a given context (e.g. a paragraph of text, an image).

6.7 Dialogue Generation

Dialog Generation is a fundamental component of artificial intelligence and, in general terms, it uses different Machine Learning algorithms in order to automatically generate a response given a textual or visual post by another agent, e.g., a user. Traditionally in the literature, dialogue systems can be divided into two subcategories: task-oriented dialogue systems (TOD) and chatbots (chit-chat) (Chen, Liu, Yin, & Tang, 2017) with the ability to open domain conversation or open domain dialog (ODD). Currently, it was also proposed

the fusion between task-oriented dialog systems (TOD) and open domain dialog (ODD) approaches to augment responses of TOD with the diversity of ODD systems. Sun et al. (Sun, Moon, Crook, Roller, Silvert, Liu, Wang, Liu, Cho, & Cardie, 2021) described three models of response generation with chit-chat addition on the base of end-to-end and code-switcher approaches. Their solution shows better human evaluation rates and similar performance measurements for tasks on the dataset presented in the article.

A relatively new multimodal task from the field of Dialog generation is the Visual Dialog Generation (VDG) task. The VDG task is widely considered to be an important part of human - artificial agent interaction. It is defined as a task where a machine holds a dialog with humans about a visual content (Das, Kottur, Gupta, Singh, Yadav, Moura, Parikh, & Batra, 2017). In the literature, the VDG task is not restricted to image data only, it can be grounded on a video, including both audio and visual information (Le, Sahoo, Chen, & Hoi, 2019).

In (Palaskar, Sanabria, & Metze, 2020), the authors propose a transfer learning approach to tackle the task of generating a textual answer in response to a textual question on multi-media content by developing a framework that uses hierarchical attention to fuse contributions from different modalities. The framework makes use of hierarchical attention trained on standard video features and a specialized form of extra-textual knowledge, i.e., answer-relevant "context" that includes past dialog turns. Furthermore, the framework manages to successfully combine different modalities, i.e., text summarization trained on multiple sources, to generate videos with commentaries. The experimental results show that the model manages to achieve state-of-the-art performance for both automatic and human evaluations.

In (Qian & Yu, 2019), the authors propose a domain adaptive dialog generation method based on meta-learning (DAML) that generates dialog for a new domain by utilizing multiple data sources from other rich-resource domains. DAML is an improved seq2seq encoder-decoder network that utilizes the two-stage CopyNet to achieve optimal initialization using dialog data from multiple domains. Thus, the model efficiently manages to generate new domain dialog by applying two-step gradient updates to learn general features. The experimental results show that DAML outperforms the state-of-the-art in Entity F1 compared with zero-shot baseline, ZSDG, as well as other transfer learning methods.

Multimodality. The Multimodality dimension in dialog generation is obtained by adding various types of interaction channels between the human user and the Multimodal Dialogue System, e.g., visual (image, video), verbal (speech, tone), vision (gaze, emotion, gesture, posture), physiological signal (blood pressure, pulse). In the literature, there are many examples of multimodality grounded on usage of images as visual content for one more input channel to the human - artificial agent interaction. In (Das et al., 2017), the authors proposed a visual dialog generation task. Authors of (De Vries, Strub, Chandar, Pietquin, Larochelle, & Courville, 2017) used image content augmented with a sequence of questions to locate an unknown object in the picture. A different method for solving of the VDA task is proposed in (Zhang, Ghosh, Heck, Walsh, Zhang, Zhang, & Kuo, 2019). Authors point out that training of generative models by the maximum likelihood estimation method (MLE) cause providing of the frequent and generic responses. In their work they proposed weighted likelihood estimation (WLE) which assigns different weights to each training sam-

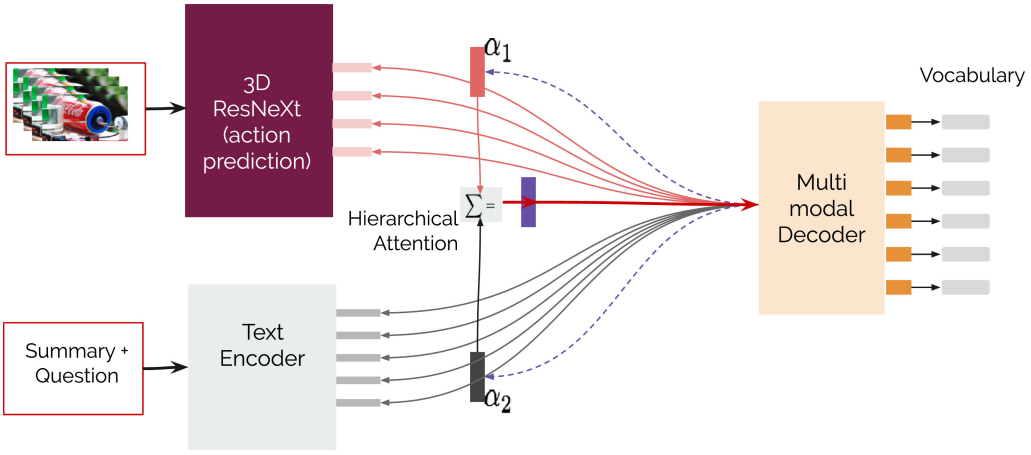

Figure 10: Illustration of Text-and-Video Dialog Generation models with Hierarchical Attention after (Palaskar et al., 2020).

ple determined by positive and negative response. This approach allows to increase level of answer diversity for the generative models. Mostafazadeh and co-workers (Mostafazadeh, Brockett, Dolan, Galley, Gao, Spithourakis, & Vanderwende, 2017) introduced task of Image Grounded Conversation(IGC) where the visual context used as a topic for a dialog and released a dataset with multi-turn conversations about images. In article Hori, Alamri, Wang, Wichern, Hori, Cherian, Marks, Cartillier, Lopes, Das, et al. (2019), the authors propose the addition of video input as a source for human - artificial agent interaction (Audio Visual Scene-Aware Dialog). In this case, video is represented as a sequence of the images, augmented with the dialog history. The task is to answer the question using this input data. Article Palaskar et al. (2020) uses hierarchical attention trained on standard video features and a specialized form of extra-textual knowledge to generate videos with accompanying commentary. The graphical representation of the text-video dialog generation model with hierarchical attention is given in Figure 10. Shuster, Humeau, Bordes, and Weston (2020) propose a multimodal architecture for grounded dialogue and an evaluation dataset, i.e., Image-Chat, for the chit-chat conversations. They evaluate retrieval and generative models with sub-components for the different modalities of the input (style, dialog history, and image). From the experimental results described in this paper, we can conclude that retrieval models show better performance than generative ones.

Multilinguality. The term multilinguality is generally understood to imply that the dialog is initiated and/or conducted in different languages. Most studies in the area of VDG tend to focus on a single language rather than on multilingual approaches. But the different language support is widely investigated in the classic, text-only dialog generation systems. In (Chen, Qiu, Fu, Liu, & Yan, 2019a), authors introduced a novel multilingual dialogue system tested on Chinese and English conversation corpora. This approach is based on the Seq2Seq framework augmented with improved shared-private memory which learns language features and improves cross-lingual transfer. Experiments on cross-lingual transfer learning for multilingual task oriented dialog were performed by Schuster, Gupta, Shah, and

Lewis (2019). In this study, the authors compared methods based on the translation of the training data, cross-lingual pre-trained embedding, and multilingual machine translation encoders as contextual word representations.

Learning Strategies. This dimension can be defined from multiple perspectives, e.g., from the point of view of: (i) the training process (aka Multi-task training) – the use of the same input dataset(s) to train a model to solve multiple tasks (e.g., translation, part-of-speech tagging, named entity identification) where models use a sequence to sequence model with shared components, showing that training on several tasks improves performance on each individual task, (ii) dialogue generation systems, e.g., a system whose primary task is dialogue generation, in order to deliver and analyze sentiments during dialogue processing, (iii) the human component which refers to the situation when a human user is busy with a primary task(s), such as driving, eye-tracking, military, and police situations, and provide dialogue (usually speech) as a secondary task.

In (Palaskar et al., 2020), a framework that uses summarization trained on multiple sources is used to generate commentaries for videos. Another multi-task approach is to learn new dialogues for resource-poor domains by training models that accurately extract meta features from resource-rich domains (Qian & Yu, 2019).

Controllability. Explicit Control offers mechanisms that incorporate a desired invariance into the learned representation. A recent review of the literature on VDG found that researches has tended to focus on the quality of the response instead of controllability. The majority of Dialogue Generation solutions based on generative models do not offer explicit control. Colombo, Witon, Modi, Kennedy, and Kapadia (2019) present an affect-driven dialogue system, which generates emotional responses in a controlled manner using a continuous representation of emotions. Madotto, Ishii, Lin, Dathathri, and Fung (2020) propose the adaptation of the Plug and play language models (PPLM) (Dathathri et al., 2020) to the generation of the conversation with different styles and topics with a large pretrained model. Plug and play conversational model (PPCM) includes a combination of the usage of the residual adapters (Houlsby, Giurgiu, Jastrzebski, Morrone, de Laroussilhe, Gesmundo, Attariyan, & Gelly, 2019) and attribute classifiers to guide the conversation generation without the modifications of the parameter of the pretrained language model. Smith, Gonzalez-Rico, Dinan, and Boureau (2020) compare the performance of the retrieve-and-style-transfer method (RnST) with PPLM and conditioned generator on inputs appended with style tags on 217 styles. This research shows that the PPLM-style approach has lower performance on larger style spaces, and condition generation without retrieval has stronger style control than RnTS.

Dialog Generation systems, task-oriented or chit-chat, are attracting more and more attention in many communities from academia and industry. From the point of view of the analyzed dimensions, we can draw the following conclusions. The Multimodality dimension in dialog generation is obtained by adding various types of interaction channels between the human user and the Multimodal Dialogue System. The multilinguality dimension generally implies that the dialog is initiated and/or conducted in different languages. From this point of view, most studies tend to focus on a single language. Dialog Generation systems can use multiple learning strategies for the training process, the human component, or the system

Table 1: The maturity levels of the NLG applications with respect to the four dimensions explored in the survey.

	Multilinguality	**Multimodality**	**Learning Strategies**	**Controllability**
Machine translation	★★★★☆	★★★☆☆	★★★☆☆	★★☆☆☆
Description generation	★★☆☆☆	★★★★☆	★★★☆☆	★★★☆☆
Automatic Speech Recognition	★★☆☆☆	★★★☆☆	★★☆☆☆	☆☆☆☆☆
Abstractive Summarization	★☆☆☆☆	★★☆☆☆	★★★☆☆	★★★☆☆
Text Simplification	★★★☆☆	★★☆☆☆	★★☆☆☆	★★★★☆
Question Answering and Generation	★★☆☆☆	★★★☆☆	★★★☆☆	★☆☆☆☆
Dialogue Generation	★★☆☆☆	★★★★☆	★★★☆☆	★☆☆☆☆

as a whole. As for the final dimension, the majority of the Dialogue Generation solutions do not offer explicit control.

7. Conclusion and Further Insights

In this paper, a survey of the state of the art in Neural Natural Language Generation (NNLG) was provided, addressing and analyzing it from a multidimensional perspective, specifically multimodality, multilinguality, learning strategies, and controllability. These dimensions were discussed in the context of relevant NNLG tasks and applications.

Research into multilinguality and multimodality has made advancements in recent years, especially through multilingual and multimodal Transformers benefiting from multi-task and transfer learning strategies. Yet, there are still *general* challenges to be addressed, regarding the redundancy–complementarity aspect of both multilingual and multimodal tasks: in modalities and in languages, there is information that is more efficiently expressed in one modality/language, than in the other modality/language, making translation from one to another difficult, while also making it challenging to integrate languages/modalities. Controllability of text generated by neural models is another *critical* research direction, as these strategies have the potential to reduce the biases in training data or to make use of supporting facts in the context.There are also *particular* problems, task-specific challenges to be addressed, that are worth highlighting in the context of the applications dealt with in this survey.

Table 1 shows, in an intuitive and illustrative manner, the maturity levels of NLG applications with respect to the four dimensions explored in this survey. More stars means that the dimension is studied in more detail, whereas less stars indicate the need for more research in that particular direction.

Concerning Machine Translation (MT), data availability is still a concern in those cases where there is little availability of labelled training data by definition (e.g., unsupervised or weakly supervised or semi-supervised settings). Upcoming research lines that have started and could become important in the next years include using images and/or video to ground simultaneous machine translation models, as well as MT models grounded on images and/or video that make little or no use of parallel data. Finally, we should see many improvements on non-autoregressive MT models in the next years as they close the performance gap with

autoregressive Neural MT model, as it has been shown in (Zhou & Keung, 2020) or (Gu & Kong, 2021).

Description generation and abstractive summarization share some common problems and future directions. On the one hand, statistical and neural models can lead to problems related to the veracity of the generated text, which may be prone to content/object hallucinations, i.e. the generation of descriptions or facts that are not fully supported by the input, for which reason more faithful approaches mitigating this effect would be desired. On the other hand, the resulting text, especially when it has to be generated from multiple sources of information (e.g., multi-sentence descriptions, multi-document summaries) often lacks coherence; hence approaches that tackle e.g. the possible repetitiveness of descriptions and their inconsistencies need to be developed. In addition, as interpretability and explainability plays an increasingly important role, researching new models with increased explainability (that go beyond saliency maps) also seems to be a promising direction for these tasks. For both tasks —description generation and abstractive summarization—, although there exist some language-independent methods, there is scarcity of resources and datasets for languages other than English or Chinese.

In the case of Automatic Speech Recognition (ASR), although much of the work on multimodal ASR has reported an increase in word error rate (WER), it should be noted that there is a marginal benefit of integrating additional modalities. Some work suggests that visual information is useful in conditions where acoustic speech is corrupted. It was shown, that the multimodal ASR models outperform the unimodal baseline model on the full unmasked test set. However, these models do not improve robustness to noise. The issue of better visually grounded adaptation techniques is still to be solved. As regards the controllable speech recognition, it can be said that all ASR systems are controllable, because the input speech influences the output text. However, when we refer to speech recognition system that enable additional output control beyond the text alone, a review of the literature revealed that there are no scientific articles dealing with this issue.

Regarding Text Simplification and Paraphrasing, it is worth stressing that text simplification uses text paraphrasing to replace hard-to-understand words with simpler, more common words, and multimodality is used to explain different kinds of data, e.g., images, videos, etc., also using these multimodal elements to support the understanding of a text. Concerning multilinguality, language-independent methods are mainly proposed for this task. Moreover, from the analysis conducted, it was shown that when doing Text Simplification, other tasks could be also tackled at the same time, e.g., translation, part-of-speech tagging, and named entity identification, etc. Thus, multitask learning is also relevant for this task. Finally, controllability is used to control both lexical and syntactic complexity in order to achieve an aggressive rewriting for Text Simplification and Paraphrasing.

As far as Question Generation and Visual Question Generation are concerned, the current efforts in question answering are mainly in answer span extraction rather than natural language generation. Lately, with the growing interest in autoregressive pretrained models such as GPT-2 or T5, there have been some recent improvements in using generative models while answering a given question. Furthermore, to steer the works on generative reading comprehension and visual questions answering tasks, we need new benchmark Question Answering (QA) datasets with open-ended answers. In line with that, developing proper automatic evaluation metrics for these kind of open-ended QA tasks would be of importance

in the near future. Neural Question generation is a research topic which has received some interest lately from the community. In the literature, there are not so many works that deal with automatic generation of questions from a given context (e.g. a paragraph of text, an image). Thus, there are still many challenges and opportunities for adapting neural NLG models to produce useful and informative questions. How to generate questions in a cross-lingual setting or how to control the style of the questions to be generated are two open problems that deserve further attention in the question generation domain as well.

Dialog generation is a central component in building real-world conversational agents. In the context of this task, the Multimodality dimension is obtained by adding various types of interaction channels between the human user and the Multimodal Dialogue System, e.g., visual (image, video), verbal (speech, tone), vision (gaze, emotion, gesture, posture), physiological signal (blood pressure, pulse). Although support for different languages is widely investigated in the classic, text-only dialog generation systems, most studies in the area of visual-based dialog generation tend to focus on a single language rather than on multilingual approaches. As the majority of the Dialogue Generation solutions do not offer explicit control, this would be an open research line to be further investigated in this field.

From the previous main take-away messages, there is no doubt that NNLG has a direct applicability in many relevant tasks for which progress has improved in the recent years. However, concerning the dimensions studied in this survey, there is still some room for improvement that opens future research lines to be investigated with the purpose of allowing these applications go in the direction towards integrating an holistic perspective with respect to multilinguality, multimodality, learning strategies and controllability.

Acknowledgments

This work has been partially supported by the European Commission ICT COST Action "Multi-task, Multilingual, Multi-modal Language Generation" (CA18231). AE was supported by BAGEP 2021 Award of the Science Academy. EE was supported in part by TUBA GEBIP 2018 Award. IC has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 838188. EL is partly funded by Generalitat Valenciana and the Spanish Government throught projects PROMETEU/2018/089 and RTI2018-094649-B-I00, respectively. SMI is partly funded by UNIRI project uniri-drustv-18-20. GB is partly supported by the Ministry of Innovation and the National Research, Development and Innovation Office within the framework of the Hungarian Artificial Intelligence National Laboratory Programme. COT is partially funded by the Romanian Ministry of European Investments and Projects through the Competitiveness Operational Program (POC) project "HOLOTRAIN" (grant no. 29/221_ap2/07.04.2020, SMIS code: 129077) and by the German Academic Exchange Service (DAAD) through the project "AWAKEN: content-Aware and netWork-Aware faKE News mitigation" (grant no. 91809005). ESA is partially funded by the German Academic Exchange Service (DAAD) through the project "Deep-Learning Anomaly Detection for Human and Automated Users Behavior" (grant no. 91809358).

References

Aafaq, N., Mian, A., Liu, W., Gilani, S. Z., & Shah, M. (2020). Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys*, *52*(6), 1–37.

Afouras, T., Chung, J., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep audio-visual speech recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13.

Agarwal, S., Bui, T., Lee, J.-Y., Konstas, I., & Rieser, V. (2020). History for visual dialog: Do we really need it?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8182–8197.

Aharoni, R., Johnson, M., & Firat, O. (2019). Massively multilingual neural machine translation. 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)*, pp. 3874–3884, Minneapolis, Minnesota.

Aksenov, D., Moreno-Schneider, J., Bourgonje, P., Schwarzenberg, R., Hennig, L., & Rehm, G. (2020). Abstractive text summarization based on language model conditioning and locality modeling. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 6680–6689.

Alikhani, M., Sharma, P., Li, S., Soricut, R., & Stone, M. (2020). Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6525–6535.

Alumäe, T., Tsakalidis, S., & Schwartz, R. M. (2016). Improved multilingual training of stacked neural network acoustic models for low resource languages. In *Interspeech*, pp. 3883–3887.

Amplayo, R. K., & Lapata, M. (2021). Informative and controllable opinion summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2662–2672.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433.

Aries, A., Zegour, D. E., & Hidouci, W. (2019). Automatic text summarization: What has been done and what has to be done. *CoRR*, *abs/1904.00688*.

Arivazhagan, N., Cherry, C., Macherey, W., Chiu, C.-C., Yavuz, S., Pang, R., Li, W., & Raffel, C. (2019). Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1313–1323.

Artetxe, M., Labaka, G., & Agirre, E. (2019). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics, pp. 194–203, Florence, Italy. Association for Computational Linguistics.

Artetxe, M., Labaka, G., Agirre, E., & Cho, K. (2018). Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations, ICLR 2015*, San Diego, California.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4945–4949. IEEE.

Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, *41*(2), 423–443.

Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., & Frank, S. (2018). Findings of the Third Shared Task on Multimodal Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 304–323, Belgium, Brussels. Association for Computational Linguistics.

Bastings, J., Aziz, W., Titov, I., & Sima'an, K. (2019). Modeling latent sentence structure in neural machine translation..

Bateman, J., & Zock, M. (2003). Natural language generation. In *The Oxford Handbook of Computational Linguistics 2nd edition*. Oxford University Press.

Bateman, J. A., & Licheng, C. M. I. M. M. (1999). Multilingual natural language generation for multilingual software: A functional linguistic approach. *Applied Artificial Intelligence*, *13*(6), 607–639.

Baumel, T., & Elhadad, M. (2019). *A Survey of Neural Models for Abstractive Summarization*, chap. Chapter 6, pp. 175–199. WSPC.

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, *7*, 49–72.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., & Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Int. Res.*, *55*(1), 409–442.

Bernsen, N. O. (2008). Multimodality theory. In *Multimodal User Interfaces*, pp. 5–29. Springer, Berlin, Heidelberg.

Bhagat, R., & Hovy, E. (2013). What is a paraphrase?. *Computational Linguistics*, *39*(3), 463–472.

Bi, B., Li, C., Wu, C., Yan, M., Wang, W., Huang, S., Huang, F., & Si, L. (2020). PALM: Pre-training an autoencoding&autoregressive language model for context-conditioned generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8681–8691.

Bikel, D., & Zitouni, I. (2012). *Multilingual Natural Language Processing Applications: From Theory to Practice*. IBM Press.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, *5*, 135–146.

Bor-Chun Chen, Y.-Y. C., & Chen, F. (2017). Video to text summary: Joint video summarization and captioning with recurrent neural networks. In Tae-Kyun Kim, Stefanos Zafeiriou, G. B., & Mikolajczyk, K. (Eds.), *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 118.1–118.14.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642.

Braga, O., Makino, T., Siohan, O., & Liao, H. (2020). End-to-end multi-person audio/visual automatic speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6994–6998. IEEE.

Bugliarello, E., Cotterell, R., Okazaki, N., & Elliott, D. (2021). Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. *Transactions of the Association for Computational Linguistics*, *9*, 978–994.

Bukhari, D., Wang, Y., & Wang, H. (2017). Multilingual convolutional, long short-term memory, deep neural networks for low resource speech recognition. *Procedia Computer Science*, *107*, 842–847.

Caglayan, O., Sanabria, R., Palaskar, S., Barraul, L., & Metze, F. (2019). Multimodal grounding for sequence-to-sequence speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8648–8652.

Caglayan, O., Ive, J., Haralampieva, V., Madhyastha, P., Barrault, L., & Specia, L. (2020). Simultaneous machine translation with visual context. In *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2350–2361.

Caglayan, O., Madhyastha, P., Specia, L., & Barrault, L. (2019). Probing the need for visual context in multimodal machine translation. 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)*, pp. 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Calixto, I., & Liu, Q. (2019). An error analysis for image-based multi-modal neural machine translation. *Machine Translation*, *33*(1–2), 155–177.

Calixto, I., Rios, M., & Aziz, W. (2019). Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6392–6405, Florence, Italy. Association for Computational Linguistics.

Calixto, I., Stein, D., Matusov, E., Castilho, S., & Way, A. (2017a). Human Evaluation of Multi-modal Neural Machine Translation: A Case-Study on E-Commerce Listing Titles. In *Proceedings of the Sixth Workshop on Vision and Language*, pp. 31–37, Valencia, Spain.

Calixto, I., Stein, D., Matusov, E., Lohar, P., Castilho, S., & Way, A. (2017b). Using Images to Improve Machine-Translating E-Commerce Product Listings.. In *Proceedings of*

the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 637–643, Valencia, Spain.

Cao, J., Gan, Z., Cheng, Y., Yu, L., Chen, Y.-C., & Liu, J. (2020). Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. *CoRR*, *abs/2006.14799*.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4964. IEEE.

Chandu, K., Prabhumoye, S., Salakhutdinov, R., & Black, A. W. (2019). "my way of telling a story": Persona based grounded story generation. In *Proceedings of the Second Workshop on Storytelling*, pp. 11–21.

Chen, A., Stanovsky, G., Singh, S., & Gardner, M. (2020). MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6521–6532.

Chen, C., Qiu, L., Fu, Z., Liu, J., & Yan, R. (2019a). Multilingual dialogue generation with shared-private memory. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 42–54. Springer.

Chen, C., Pan, Z. F., Liu, M., & Sun, M. (2019b). Unsupervised stylish image description generation via domain layer norm. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 8151–8158.

Chen, H., Xie, W., Vedaldi, A., & Zisserman, A. (2020). Vggsound: A large-scale audio-visual dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 721–725.

Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, *19*(2), 25–35.

Chen, J., & Zhuge, H. (2018). Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4046–4056.

Chen, W., Chen, J., Su, Y., Chen, Z., & Wang, W. Y. (2020). Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7929–7942.

Chen, Y., Ma, Y., Mao, X., & Li, Q. (2019). Multi-task learning for abstractive and extractive summarization. *Data Science and Engineering*, *4*(1), 14–23.

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). Uniter: Learning universal image-text representations. In *ICLR*.

Chen, Y., Liu, Y., & Li, V. (2018). Zero-resource neural machine translation with multi-agent communication game. In *AAAI Conference on Artificial Intelligence*.

Chi, Z., Dong, L., Wei, F., Wang, W., Mao, X.-L., & Huang, H. (2020). Cross-lingual natural language generation via pre-training.. In *AAAI*, pp. 7570–7577.

Chiu, C.-C., & Raffel, C. (2018). Monotonic chunkwise attention. In *International Conference on Learning Representations*.

Cho, J., Baskar, M. K., Li, R., Wiesner, M., Mallidi, S. H., Yalta, N., Karafiát, M., Watanabe, S., & Hori, T. (2018). Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 521–527.

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014a). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111, Doha, Qatar. Association for Computational Linguistics.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014b). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar.

Chu, C., Otani, M., & Nakashima, Y. (2018). iParaphrasing: Extracting visually grounded paraphrases via an image. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3479–3492.

Chung, H. W., Garrette, D., Tan, K. C., & Riesa, J. (2020). Improving multilingual models with language-clustered vocabularies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4536–4546.

Collell, G., & Moens, M. F. (2018). Do neural network cross-modal mappings really bridge modalities?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 462–468.

Colombo, P., Witon, W., Modi, A., Kennedy, J., & Kapadia, M. (2019). Affect-driven dialog generation. In *NAACL-HLT (1)*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451.

Cornia, M., Baraldi, L., & Cucchiara, R. (2019). Show, control and tell: A framework for generating controllable and grounded captions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Crego, J. M., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., Brunelle, P., Coquard, A., Deng, Y., et al. (2016). Systran's pure neural machine translation systems. *CoRR, abs/1610.05540*.

Cui, J., Kingsbury, B., Ramabhadran, B., Saon, G., Sercu, T., Audhkhasi, K., Sethy, A., Nussbaum-Thom, M., & Rosenberg, A. (2017). Knowledge distillation across ensembles of multilingual models for low-resource languages. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4825–4829.

Currey, A., & Heafield, K. (2019). Incorporating source syntax into transformer-based neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pp. 24–33, Florence, Italy. Association for Computational Linguistics.

Cygert, S., Szwoch, G., Zaporowski, S., & Czyzewski, A. (2018). Vocalic segments classification assisted by mouth motion capture. In *IEEE International Conference on Human System Interaction (HSI)*, pp. 318–324.

Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J., & Szykulski, M. (2017). An audio-visual corpus for multimodal automatic speech recognition. *Journal of Intelligent Information Systems*, *49*, 167–192.

Dabre, R., Chu, C., & Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. *ACM Comput. Surv.*, *53*(5).

Dalmia, S., Sanabria, R., Metze, F., & Black, A. W. (2018). Sequence-based multi-lingual low resource speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4909–4913.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., & Batra, D. (2017). Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 326–335.

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., & Liu, R. (2020). Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., & Courville, A. (2017). Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5503–5512.

Della Pietra, S., Epstein, M., Roukos, S., & Ward, T. (1997). Fertility models for statistical natural language understanding. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 168–173.

Deshpande, A., Aneja, J., Wang, L., Schwing, A. G., & Forsyth, D. (2019). Fast, diverse and accurate image captioning guided by part-of-speech. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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)*, pp. 4171–4186.

DGT (2018). Management plan 2018. In *Directorate-General for Translation, European Commission*.

Dinu, G., Mathur, P., Federico, M., & Al-Onaizan, Y. (2019). Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3063–3068, Florence, Italy. Association for Computational Linguistics.

Dong, D., Wu, H., He, W., Yu, D., & Wang, H. (2015). Multi-Task Learning for Multiple Language Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1723–1732, Beijing, China.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., & Hon, H.-W. (2019). Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32*, pp. 13063–13075.

Du, C., Li, H., Lu, Y., Wang, L., & Qian, Y. (2021). Data augmentation for end-to-end code-switching speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 194–200. IEEE.

Dunietz, J., Burnham, G., Bharadwaj, A., Rambow, O., Chu-Carroll, J., & Ferrucci, D. (2020). To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7839–7859.

Dutta Chowdhury, K., & Elliott, D. (2019). Understanding the effect of textual adversaries in multimodal machine translation. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pp. 35–40, Hong Kong, China. Association for Computational Linguistics.

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 489–500, Brussels, Belgium. Association for Computational Linguistics.

Eikema, B., & Aziz, W. (2019). Auto-encoding variational neural machine translation. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pp. 124–141, Florence, Italy. Association for Computational Linguistics.

Elliott, D. (2018). Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

Elliott, D., Frank, S., Barrault, L., Bougares, F., & Specia, L. (2017). Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pp. 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Fan, A., & Gardent, C. (2020). Multilingual AMR-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2889–2901.

Fan, A., Grangier, D., & Auli, M. (2018). Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 45–54.

Fernandez-Lopez, A., & Sukno, F. M. (2018). Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*, *78*, 53–72.

Firat, O., Cho, K., & Bengio, Y. (2016a). Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 866–875, San Diego, California.

Firat, O., Sankaran, B., Al-onaizan, Y., Yarman Vural, F. T., & Cho, K. (2016b). Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 268–277, Austin, Texas. Association for Computational Linguistics.

Fisch, A., Lee, K., Chang, M. W., Clark, J. H., & Barzilay, R. (2020). CAPWAP: Captioning with a Purpose. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Freitag, M., & Firat, O. (2020). Complete multilingual neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 550–560, Online. Association for Computational Linguistics.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 457–468.

Garbacea, C., & Mei, Q. (2020). Neural language generation: Formulation, methods, and evaluation. *CoRR*, *abs/2007.15780*.

Garcia, X., Foret, P., Sellam, T., & Parikh, A. (2020). A multilingual view of unsupervised machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3160–3170, Online. Association for Computational Linguistics.

Gardent, C., Shimorina, A., Narayan, S., & Perez-Beltrachini, L. (2017). The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pp. 124–133.

Garmash, E., & Monz, C. (2016). Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1409–1418, Osaka, Japan. The COLING 2016 Organizing Committee.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, *61*(1), 65–170.

Goldberg, Y., Hirst, G., Liu, Y., & Zhang, M. (2018). Neural network methods for natural language processing yoav goldberg (bar ilan university)morgan & claypool (synthesis lectures on human language technologies, edited by graeme hirst, volume 37), 2017, xxii+287 pp; paperback, ISBN 9781627052986, $74.95; ebook, ISBN 9781627052955, $59.96; doi: 10.2200/s00762ed1v01y201703hlt037. *Comput. Linguistics*, *44*(1).

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., & Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., & Weinberger, K. Q. (Eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2672–2680.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913.

Graves, A. (2012). Sequence transduction with recurrent neural networks. In *arXiv preprint arXiv:1211.3711*.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376.

Grega, M., Smaïli, K., Leszczuk, M., González-Gallardo, C.-E., Torres-Moreno, J.-M., Pontes, E. L., Fohr, D., Mella, O., Menacer, M., & Jouvet, D. (2018). An integrated amis prototype for automated summarization and translation of newscasts and reports. In *International Conference on Multimedia and Network Information System*, pp. 415–423. Springer.

Gu, J., Bradbury, J., Xiong, C., Li, V. O., & Socher, R. (2018). Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Gu, J., & Kong, X. (2021). Fully non-autoregressive neural machine translation: Tricks of the trade. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 120–133, Online. Association for Computational Linguistics.

Gu, J., Joty, S., Cai, J., & Wang, G. (2018). Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Guo, H., Pasunuru, R., & Bansal, M. (2018). Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 687–697.

Guo, W., Wang, J., & Wang, S. (2019). Deep multimodal representation learning: A survey. *IEEE Access*, *7*, 63373–63394.

Gupta, A., Miao, Y., Neves, L., & Metze, F. (2017). Visual features for context-aware speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5020–5024.

Gupta, S., & Gupta, S. K. (2019). Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, *121*, 49 – 65.

Hasler, E., de Gispert, A., Iglesias, G., & Byrne, B. (2018). Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 506–512, New Orleans, Louisiana. Association for Computational Linguistics.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.-Y., & Ma, W.-Y. (2016). Dual learning for machine translation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, p. 820–828, Red Hook, NY, USA. Curran Associates Inc.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, X., & Deng, L. (2017). Deep learning for image-to-text generation: A technical overview. *IEEE Signal Processing Magazine*, *34*(6), 109–116.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2545–2568.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*, 1735–80.

Hokamp, C., & Liu, Q. (2017). Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Hori, C., Alamri, H., Wang, J., Wichern, G., Hori, T., Cherian, A., Marks, T. K., Cartillier, V., Lopes, R. G., Das, A., et al. (2019). End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2352–2356. IEEE.

Hotelling, H. (1992). Relations between two sets of variates. In *Breakthroughs in statistics*, pp. 162–190. Springer.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In Chaudhuri, K., & Salakhutdinov, R. (Eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR.

Hrga, I., & Ivašić-Kos, M. (2019). Deep image captioning: An overview. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 995–1000.

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., & Xing, E. P. (2017). Toward controlled generation of text. In Precup, D., & Teh, Y. W. (Eds.), *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of *Proceedings of Machine Learning Research*, pp. 1587–1596.

Huang, P., Huang, J., Guo, Y., Qiao, M., & Zhu, Y. (2019). Multi-grained attention with object-level grounding for visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3595–3600.

Höllerer, M. A., Jancsary, D., & Grafström, M. (2018). 'a picture is worth a thousand words': Multimodal sensemaking of the global financial crisis. *Organization Studies*, *39*(5-6), 617–644.

Imankulova, A., Kaneko, M., Hirasawa, T., & Komachi, M. (2020). Towards multimodal simultaneous neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 540–549, Online. Association for Computational Linguistics.

Iqbal, T., & Qureshi, S. (2020). The survey: Text generation models in deep learning. In *Journal of King Saud University - Computer and Information Sciences*.

Jaffe, A. (2017). Generating image descriptions using multilingual data. In *Proceedings of the Second Conference on Machine Translation*, pp. 458–464.

Jain, U., Zhang, Z., & Schwing, A. G. (2017). Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6485–6494.

Jangra, A., Jatowt, A., Hasanuzzaman, M., & Saha, S. (2020). Text-image-video summary generation using joint integer linear programming. In Jose, J. M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M. J., & Martins, F. (Eds.), *Advances in Information Retrieval*, pp. 190–198.

Jin, H., Cao, Y., Wang, T., Xing, X., & Wan, X. (2020). Recent advances of neural text generation: Core tasks, datasets, models and challenges. *Science China Technological Sciences*, *63*(10), 1990–2010.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, 339–351.

Joshi, V., Zhao, R., Mehta, R. R., Kumar, K., & Li, J. (2020). Transfer learning approaches for streaming end-to-end speech recognition system. In *Proc. Interspeech*.

Kaiser, L., Bengio, S., Roy, A., Vaswani, A., Parmar, N., Uszkoreit, J., & Shazeer, N. (2018). Fast decoding in sequence models using discrete latent variables. In Dy, J., & Krause, A. (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 2390–2399, Stockholmsmässan, Stockholm Sweden. PMLR.

Kajiwara, T., & Komachi, M. (2016). Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1147–1158.

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pp. 119–126, Sofia, Bulgaria. Association for Computational Linguistics.

Kannan, A., Datta, A., Sainath, T. N., Weinstein, E., Ramabhadran, B., Wu, Y., Bapna, A., Chen, Z., & Lee, S. (2019). Large-scale multilingual speech recognition with a streaming end-to-end model. In *Proc. Interspeech*, pp. 2130–2134.

Kawaler, M., & Czyzewski, A. (2019). Database of speech and facial expressions recorded with optimized face motion capture settings. *Journal of Intelligent Information Systems*, *53*, 1–24.

Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., & Hajishirzi, H. (2017). Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5376–5384.

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). CTRL: A conditional transformer language model for controllable generation. *CoRR*, *abs/1909.05858*.

Khashabi, D., Min, S., Khot, T., Sabharwal, A., Tafjord, O., Clark, P., & Hajishirzi, H. (2020). UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1896–1907.

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP)*, pp. 36–45.

Kikuchi, Y., Neubig, G., Sasano, R., Takamura, H., & Okumura, M. (2016). Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1328–1338.

Kim, D., Kim, S., & Kwak, N. (2019). Textbook question answering with multi-modal context graph understanding and self-supervised open-set comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3568–3584.

Kim, H., & Bansal, M. (2019). Improving visual question answering by referring to generated paragraph captions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3606–3612.

Kim, S., Hori, T., & Watanabe, S. (2017). Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4835–4839.

Kim, S., & Seltzer, M. L. (2018). Towards language-universal end-to-end speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4914–4918.

Kim, Y., Graca, M., & Ney, H. (2020). When and why is unsupervised neural machine translation useless?. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 35–44, Lisboa, Portugal. European Association for Machine Translation.

Kiros, R., Salakhutdinov, R., & Zemel, R. (2014). Multimodal neural language models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, p. II–595–II–603.

Krause, B., Gotmare, A. D., McCann, B., Keskar, N. S., Joty, S. R., Socher, R., & Rajani, N. F. (2020). GeDi: Generative discriminator guided sequence generation. *CoRR*, *abs/2009.06367*.

Krishna, R., Bernstein, M., & Fei-Fei, L. (2019). Information maximizing visual question generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2008–2018.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., & Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*, 32–73.

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71.

Kumar, V., Joshi, N., Mukherjee, A., Ramakrishnan, G., & Jyothi, P. (2019). Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4863–4872.

Laenen, K., & Moens, M.-F. (2019). Multimodal neural machine translation of fashion e-commerce descriptions. In Kalbaska, N., Sádaba, T., Cominelli, F., & Cantoni, L. (Eds.), *Fashion Communication in the Digital Age*, pp. 46–57, Cham. Springer International Publishing.

Lala, C., & Specia, L. (2018). Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Lan, W., Li, X., & Dong, J. (2017). Fluency-guided cross-lingual image captioning. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, p. 1549–1557.

Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4483–4499.

Le, H., Sahoo, D., Chen, N., & Hoi, S. (2019). Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5612–5623.

Lee, J., Mansimov, E., & Cho, K. (2018). Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1173–1182, Brussels, Belgium. Association for Computational Linguistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (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*, pp. 7871–7880.

Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. (2019). Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 6666–6673.

Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D., & Zhou, M. (2020). Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training.. In *AAAI*, pp. 11336–11344.

Li, H., Zhu, J., Liu, T., Zhang, J., & Zong, C. (2018). Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 4152–4158.

Li, H., Zhu, J., Ma, C., Zhang, J., & Zong, C. (2017). Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1092–1102.

Li, H., Zhu, J., Zhang, J., & Zong, C. (2018). Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1430–1441.

Li, K., Li, J., Ye, G., Zhao, R., & Gong, Y. (2019). Towards code-switching asr for end-to-end ctc models. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6076–6080. IEEE.

Li, L., & Gong, B. (2019). End-to-end video captioning with multitask reinforcement learning. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pp. 339–348.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., & Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In Vedaldi, A., Bischof, H., Brox, T., & Frahm, J.-M. (Eds.), *Computer Vision – ECCV 2020*, pp. 121–137, Cham.

Li, Y., Duan, N., Zhou, B., Chu, X., Ouyang, W., Wang, X., & Zhou, M. (2018). Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6116–6124.

Libovický, J., & Helcl, J. (2018). End-to-end non-autoregressive neural machine translation with connectionist temporal classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3016–3021, Brussels, Belgium. Association for Computational Linguistics.

Libovický, J., Palaskar, S., Gella, S., & Metze, F. (2018). Multimodal abstractive summarization for open-domain videos. In *Visually Grounded Interaction and Language (ViGIL)*, pp. 1–8.

Lin, H., & Ng, V. (2019). Abstractive summarization: A survey of the state of the art. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 9815–9822.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., & Tuytelaars, T. (Eds.), *Computer Vision – ECCV 2014*, pp. 740–755.

Lin, Z., & Wan, X. (2021). Pushing paraphrase away from original sentence: A multi-round paraphrase generation approach. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1548–1557, Online. Association for Computational Linguistics.

Liu, B., Wei, H., Niu, D., Chen, H., & He, Y. (2020). Asking questions the human way: Scalable question-answer generation from text corpus. In *Proceedings of The Web Conference 2020*, pp. 2032–2043.

Liu, D., Wan, X., Xu, J., & Zhang, P. (2018). Multilingual speech recognition training and adaptation with language-specific gate units. In *11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 86–90.

Liu, D., Xu, J., & Zhang, P. (2020). End-to-end multilingual speech recognition system with language supervision training. *IEICE Transactions on Information and Systems*, *103*, 1427–1430.

Liu, F., Xiang, T., Hospedales, T. M., Yang, W., & Sun, C. (2018). ivqa: Inverse visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8611–8619.

Liu, F., Liu, Y., Ren, X., He, X., & Sun, X. (2019a). Aligning visual regions and textual concepts for semantic-grounded image representations. In *NeurIPS*, pp. 6847–6857.

Liu, L., Tang, J., Wan, X., & Guo, Z. (2019b). Generating diverse and descriptive image captions using visual paraphrases. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4240–4249.

Liu, W., Li, L., Huang, Z., & Liu, Y. (2019c). Multi-lingual Wikipedia summarization and title generation on low resource corpus. In *Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources*, pp. 17–25.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, *8*, 726–742.

Liu, Y., Luo, Z., & Zhu, K. (2018). Controlling length in abstractive summarization using a convolutional neural network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4110–4119.

Long, Y., Li, Y., Zhang, Q., Wei, S., Ye, H., & Yang, J. (2020). Acoustic data augmentation for mandarin-english code-switching speech recognition. *Applied Acoustics*, *161*, 107175.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pp. 13–23.

Lu, J., Goswami, V., Rohrbach, M., Parikh, D., & Lee, S. (2020). 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10437–10446.

Lu, Y., Keung, P., Ladhak, F., Bhardwaj, V., Zhang, S., & Sun, J. (2018). A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 84–92, Brussels, Belgium. Association for Computational Linguistics.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., & Kaiser, L. (2016). Multi-Task Sequence to Sequence Learning. In *Proceedings of the International Conference on Learning Representations (ICLR), 2016*, San Juan, Puerto Rico.

Luong, T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation.. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1412–1421, Lisbon, Portugal.

Ma, M., Huang, L., Xiong, H., Zheng, R., Liu, K., Zheng, B., Zhang, C., He, Z., Liu, H., Li, X., Wu, H., & Wang, H. (2019). STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3025–3036, Florence, Italy. Association for Computational Linguistics.

Ma, X., Pino, J. M., Cross, J., Puzon, L., & Gu, J. (2020). Monotonic multihead attention. In *International Conference on Learning Representations*.

Maddela, M., Alva-Manchego, F., & Xu, W. (2021). Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3536–3553.

Madotto, A., Ishii, E., Lin, Z., Dathathri, S., & Fung, P. (2020). Plug-and-play conversational models. In Cohn, T., He, Y., & Liu, Y. (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, Vol. EMNLP 2020 of *Findings of ACL*, pp. 2422–2433. Association for Computational Linguistics.

Makino, T., Liao, H., Assael, Y., Shillingford, B., Garcia, B., Braga, O., & Siohan, O. (2019). Recurrent neural network transducer for audio-visual speech recognition. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pp. 905–912. IEEE.

Mallinson, J., Sennrich, R., & Lapata, M. (2017). Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 881–893.

Marie, B., Wang, R., Fujita, A., Utiyama, M., & Sumita, E. (2018). NICT's neural and statistical machine translation systems for the WMT18 news translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 449–455, Belgium, Brussels. Association for Computational Linguistics.

Martin, L., Éric de la Clergerie, Sagot, B., & Bordes, A. (2020). Controllable sentence simplification. In *Conference on Language Resources and Evaluation*, pp. 4689–4698.

Massiceti, D., Dokania, P. K., Siddharth, N., & Torr, P. H. S. (2018). Visual dialogue without vision or dialogue. *CoRR*, *abs/1812.06417*.

Matassoni, M., Gretter, R., Falavigna, D., & Giuliani, D. (2018). Non-native children speech recognition through transfer learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6229–6233.

Melaku, S., Di Gangi, M., & Federico, M. (2019). Controlling the output length of neural machine translation. In *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT)*.

Meng, Z., Li, J., Chen, Z., Zhao, Y., Mazalov, V., Gang, Y., & Juang, B. H. (2018). Speaker-invariant training via adversarial learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5969–5973.

Merboldt, A., Zeyer, A., Schlüter, R., & Ney, H. (2019). An analysis of local monotonic attention variants.. In *INTERSPEECH*, pp. 1398–1402.

Miao, Y., & Metze, F. (2016). Open-domain audio-visual speech recognition: A deep learning approach. In *Interspeech*, pp. 3414–3418.

Miiller, M., Stiiker, S., & Waibel, A. (2018). Multilingual adaptation of RNN based ASR systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5219–5223.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Ghahramani, Z., & Weinberger, K. Q. (Eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 3111–3119.

Miyazaki, T., & Shimizu, N. (2016). Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1780–1790.

Mogadala, A., Kalimuthu, M., & Klakow, D. (2021). Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *J. Artif. Int. Res.*, *71*, 1183 – 1317.

Mostafazadeh, N., Brockett, C., Dolan, B., Galley, M., Gao, J., Spithourakis, G., & Vanderwende, L. (2017). Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 462–472.

Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., & Vanderwende, L. (2016). Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1802–1813.

Mulcaire, P., Kasai, J., & Smith, N. A. (2019). Low-resource parsing with crosslingual contextualized representations. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 304–315.

Murray, K., & Chiang, D. (2018). Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 212–223, Brussels, Belgium.

Nakayama, H., & Nishida, N. (2017). Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot. *Machine Translation*, *31*(1-2), 49–64.

Nan, L., Radev, D., Zhang, R., Rau, A., Sivaprasad, A., Hsieh, C., Tang, X., Vyas, A., Verma, N., Krishna, P., Liu, Y., Irwanto, N., Pan, J., Rahman, F., Zaidi, A., Mutuma, M., Tarabar, Y., Gupta, A., Yu, T., Tan, Y. C., Lin, X. V., Xiong, C., Socher, R., & Rajani, N. F. (2021). DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 432–447, Online. Association for Computational Linguistics.

Nassar, I., Ananda-Rajah, M., & Haffari, G. (2019). Neural versus non-neural text simplification: A case study. In *Australasian Language Technology Association*, pp. 172–177.

Nishida, K., Saito, I., Nishida, K., Shinoda, K., Otsuka, A., Asano, H., & Tomita, J. (2019). Multi-style generative reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2273–2284.

Nishihara, D., Kajiwara, T., & Arase, Y. (2019). Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 260–266.

Nishimura, Y., Sudoh, K., Neubig, G., & Nakamura, S. (2020). Multi-source neural machine translation with missing data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 569–580.

Nisioi, S., Štajner, S., Ponzetto, S. P., & Dinu, L. P. (2017). Exploring neural text simplification models. In *Annual Meeting of the Association for Computational Linguistics*.

Orbach, E., & Goldberg, Y. (2020). Facts2Story: Controlling text generation by key facts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 2329–2345, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Palaskar, S., Sanabria, R., & Metze, F. (2018). End-to-end multimodal speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5774–5778.

Palaskar, S., Libovický, J., Gella, S., & Metze, F. (2019). Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6587–6596.

Palaskar, S., Sanabria, R., & Metze, F. (2020). Transfer learning for multimodal dialog. *Computer Speech & Language*, *64*, 101093.

Parikh, A., Wang, X., Gehrmann, S., Faruqui, M., Dhingra, B., Yang, D., & Das, D. (2020). ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1173–1186.

Park, C. C., Kim, B., & Kim, G. (2017). Attend to You: Personalized Image Captioning with Context Sequence Memory Networks. In *CVPR*.

Park, H., Oh, K.-J., Choi, H.-J., & Gweon, G. (2019). Constructing a paraphrase database for agglutinative languages. *Data & Knowledge Engineering*, *123*, 101604.

Pasunuru, R., Guo, H., & Bansal, M. (2017). Towards improving abstractive summarization via entailment generation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 27–32.

Patil, C., & Patwardhan, M. (2020). Visual question generation: The state of the art. *ACM Comput. Surv.*, *53*(3).

Patro, B. N., Kumar, S., Kurmi, V. K., & Namboodiri, V. (2018). Multimodal differential network for visual question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4002–4012.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In Moschitti, A., Pang, B., & Daelemans, W. (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In Walker, M. A., Ji, H., & Stent, A. (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 2227–2237.

Petridis, S., Stafylakis, T., Ma, P., Tzimiropoulos, G., & Pantic, M. (2018). Audio-visual speech recognition with a hybrid ctc/attention architecture. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 513–520.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Pironkov, G., Dupont, S., & Dutoit, T. (2016). Multi-task learning for speech recognition: an overview. In *ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 27–29.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649.

Popović, B., Pakoci, E., & Pekar, D. (2019). Transfer leaming in automatic speech recognition for Serbian. In *27th Telecommunications Forum (TELFOR)*, pp. 1–4.

Post, M., & Vilar, D. (2018). Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Prabhumoye, S., Black, A. W., & Salakhutdinov, R. (2020). Exploring controllable text generation techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 1–14.

Qian, K., & Yu, Z. (2019). Domain adaptive dialog generation via meta learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2639–2649.

Radford, A., & Narasimhan, K. (2018). Improving language understanding by generative pre-training. Tech. rep., OpenAI.

Raunak, V., Choe, S. K., Lu, Q., Xu, Y., & Metze, F. (2019). On leveraging the visual modality for neural machine translation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 147–151, Tokyo, Japan. Association for Computational Linguistics.

Reed, S. E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., & Lee, H. (2016). Learning what and where to draw. In *Advances in neural information processing systems*, pp. 217–225.

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., & Pinkal, M. (2013). Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, *1*, 25–36.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge university press.

Ren, M., Kiros, R., & Zemel, R. (2015a). Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems 28*, pp. 2953–2961.

Ren, S., He, K., Girshick, R., & Sun, J. (2015b). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99.

Ren, S., Zhang, Z., Liu, S., Zhou, M., & Ma, S. (2019). Unsupervised neural machine translation with smt as posterior regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 241–248.

Rishes, E., Lukin, S. M., Elson, D. K., & Walker, M. A. (2013). Generating different story tellings from semantic representations of narrative. In Koenitz, H., Sezen, T. I., Ferri, G., Haahr, M., Sezen, D., & Çatak, G. (Eds.), *Interactive Storytelling*, pp. 192–204.

Rönnqvist, S., Kanerva, J., Salakoski, T., & Ginter, F. (2019). Is multilingual BERT fluent in language generation?. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pp. 29–36.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *CoRR*, *abs/1706.05098*.

Ruder, S., Vulić, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, *65*(1), 569–630.

Saggion, H. (2017). Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1–137.

Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., & Metze, F. (2018). How2: A large-scale dataset for multimodal language understanding. In *Visually Grounded Interaction and Language (ViGIL), Montreal; Canada, December 2018. Neural Information Processing Society (NeurIPS).*, arXiv. arxiv.org. 32nd Annual Conference on Neural Information Processing Systems, NeurIPS ; Conference date: 02-12-2018 Through 08-12-2018.

Santhanam, S., & Shaikh, S. (2019). A survey of natural language generation techniques with a focus on dialogue systems - past, present and future directions. *CoRR*, *abs/1906.00500*.

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L., Roomi, B., & Hall, B. (2017). English conversational telephone speech recognition by humans and machines. In *arXiv preprint arXiv:1703.02136*.

Sargin, M. E., Yemez, Y., Erzin, E., & Tekalp, A. M. (2007). Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Transactions on Multimedia*, *9*(7), 1396–1403.

Sarrouti, M., Abacha, A. B., & Demner-Fushman, D. (2020). Visual question generation from radiology images. In *Proceedings of the First Workshop on Advances in Language and Vision Research*, pp. 12–18.

Schuster, M., & Nakajima, K. (2012). Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5149–5152. IEEE.

Schuster, S., Gupta, S., Shah, R., & Lewis, M. (2019). Cross-lingual transfer learning for multilingual task oriented dialog. 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)*, pp. 3795–3805.

Schwenk, H., & Douze, M. (2017). Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*.

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1083.

Seide, F., Li, G., & Yu, D. (2011). Conversational speech transcription using context-dependent deep neural networks. In *Twelfth annual conference of the international speech communication association*.

Sennrich, R., & Haddow, B. (2016). Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pp. 83–91, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., & Birch, A. (2016a). Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the As-*

sociation for Computational Linguistics (Volume 1: Long Papers), pp. 86–96, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., & Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Sercu, T., Puhrsch, C., Kingsbury, B., & LeCun, Y. (2016). Very deep multilingual convolutional neural networks for lvcsr. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4955–4959.

Sercu, T., Saon, G., Cui, J., Cui, X., Ramabhadran, B., Kingsbury, B., & Sethy, A. (2017). Network architectures for multilingual speech representation learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5295–5299.

Serdyuk, D., Audhkhasi, K., Brakel, P., Ramabhadran, B., Thomas, S., & Bengio, Y. (2016). Invariant representations for noisy speech recognition. In *arXiv preprint arXiv:1612.01928*.

Shah, H., & Barber, D. (2018). Generative neural machine translation. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 31*, pp. 1346–1355. Curran Associates, Inc.

Shah, M., Chen, X., Rohrbach, M., & Parikh, D. (2019). Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6649–6658.

Shahidi, H., Li, M., & Lin, J. (2020). Two birds, one stone: A simple, unified model for text generation from structured and unstructured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3864–3870.

Shekhar, R., Takmaz, E., Fernández, R., & Bernardi, R. (2019). Evaluating the representational hub of language and vision models. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pp. 211–222.

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2020). Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3239–3254.

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. 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)*, pp. 3407–3412.

Shi, T., Keneshloo, Y., Ramakrishnan, N., & Reddy, C. K. (2021). Neural abstractive text summarization with sequence-to-sequence models. *ACM/IMS Trans. Data Sci.*, *2*(1).

Shijie, Z., Lizhen, Q., Shaodi, Y., Zhenglu, Y., & Jiawan, Z. (2017). Automatic generation of grounded visual questions. In *Proceedings of the 26th IJCAI*, pp. 4235–4243.

Shinohara, Y. (2016). Adversarial multi-task learning of deep neural networks for robust speech recognition. In *Proc. Interspeech*, pp. 2369–2372.

Shivakumar, P. G., & Georgiou, P. (2020). Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Computer Speech and Language*, *63*, 101077.

Shreve, G. M. (2006). Integration of translation and summarization processes in summary translation. *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association*, *1*(1), 87–109.

Shu, L., Papangelis, A., Wang, Y.-C., Tur, G., Xu, H., Feizollahi, Z., Liu, B., & Molino, P. (2020). Controllable text generation with focused variation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3805–3817, Online. Association for Computational Linguistics.

Shuster, K., Humeau, S., Bordes, A., & Weston, J. (2020). Image-chat: Engaging grounded conversations. In Jurafsky, D., Chai, J., Schluter, N., & Tetreault, J. R. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 2414–2429. Association for Computational Linguistics.

Shuster, K., Humeau, S., Hu, H., Bordes, A., & Weston, J. (2019). Engaging image captioning via personality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sigurdsson, G. A., Alayrac, J.-B., Nematzadeh, A., Smaira, L., Malinowski, M., Carreira, J., Blunsom, P., & Zisserman, A. (2020). Visual grounding in video for unsupervised word translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y., & LeCun, Y. (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sjöblom, E., Creutz, M., & Aulamo, M. (2018). Paraphrase detection on noisy subtitles in six languages. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pp. 64–73.

Smith, E. M., Gonzalez-Rico, D., Dinan, E., & Boureau, Y. (2020). Controlling style in generated dialogue. *CoRR*, *abs/2009.10855*.

Song, S., Huang, H., & Ruan, T. (2019). Abstractive text summarization using lstm-cnn based deep learning. *Multimedia Tools Appl.*, *78*(1), 857–875.

Specia, L., Frank, S., Sima'an, K., & Elliott, D. (2016). A Shared Task on Multimodal Machine Translation and Crosslingual Image Description. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016*, pp. 543–553, Berlin, Germany.

Srinivasan, T., Sanabria, R., & Metze, F. (2019). Analyzing utility of visual context in multimodal speech recognition under noisy conditions. In *arXiv preprint arXiv:1907.00477*.

Štajner, S., & Glavaš, G. (2017). Leveraging event-based semantics for automated text simplification. *Expert Systems with Applications*, *82*, 383–395.

Sterpu, G., Saam, C., & Harte, N. (2020). Should we Hard-Code the Recurrence Concept or Learn it Instead ? Exploring the Transformer Architecture for Audio-Visual Speech Recognition. In *Proc. Interspeech 2020*, pp. 3506–3509.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2019). Vl-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Su, Y., Vandyke, D., Wang, S., Fang, Y., & Collier, N. (2021). Plan-then-generate: Controlled data-to-text generation via planning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.

Su, Y., Fan, K., Bach, N., Kuo, C.-C. J., & Huang, F. (2019). Unsupervised multi-modal neural machine translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10482–10491.

Sulubacak, U., Caglayan, O., Grönroos, S.-A., Rouhe, A., Elliott, D., Specia, L., & Tiedemann, J. (2020). Multimodal machine translation through visuals and speech. *Machine Translation*, *34*.

Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). VideoBERT: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7464–7473.

Sun, K., Moon, S., Crook, P., Roller, S., Silvert, B., Liu, B., Wang, Z., Liu, H., Cho, E., & Cardie, C. (2021). Adding chit-chat to enhance task-oriented dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1570–1583.

Sun, S., Yeh, C.-F., Ostendorf, M., Hwang, M.-Y., & Xie, L. (2018). Training Augmentation with Adversarial Examples for Robust Speech Recognition. In *Proc. Interspeech*, pp. 2404–2408.

Susanto, R. H., Chollampatt, S., & Tan, L. (2020). Lexically constrained neural machine translation with Levenshtein transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3536–3543, Online. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, pp. 3104–3112, Montréal, Canada.

Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. 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)*, pp. 5103–5114.

Tang, Z., Li, L., & Wang, D. (2016). Multi-task recurrent model for true multilingual speech recognition. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*.

Tanti, M., Gatt, A., & Camilleri, K. (2018). Where to put the image in an image caption generator. *Natural Language Engineering*, *24*(3).

Tao, F., & Busso, C. (2018a). Aligning audiovisual features for audiovisual speech recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*.

Tao, F., & Busso, C. (2018b). Gating neural network for large vocabulary audiovisual speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*, 1290–1302.

Thomas, S., Audhkhasi, K., Cui, J., Kingsbury, B., & Ramabhadran, B. (2016). Multilingual data selection for low resource speech recognition. In *Interspeech*, pp. 3853–3857.

Tong, R., Wang, L., & Ma, B. (2017a). Transfer learning for children's speech recognition. In *International Conference on Asian Language Processing (IALP)*, pp. 36–39.

Tong, S., Garner, P. N., & Bourlard, H. (2017b). An investigation of deep neural networks for multilingual speech recognition training and adaptation. In *Interspeech*.

Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E., & Rao, K. (2018). Multilingual speech recognition with a single end-to-end model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4904–4908.

Toshniwal, S., Tang, H., Lu, L., & Livescu, K. (2017). Multitask learning with low-level auxiliary tasks for encoder-decoder based speech recognition. In *Proc. Interspeech*.

Toyama, J., Misono, M., Suzuki, M., Nakayama, K., & Matsuo, Y. (2016). Neural machine translation with latent semantic of image and text. *CoRR*, *abs/1611.08459*.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.

van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech and Language*, *67*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., & Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 30.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to sequence - video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wan, X., Li, H., & Xiao, J. (2010). Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 917–926.

Wang, B., Wang, C., Zhang, Q., Su, Y., Wang, Y., & Xu, Y. (2020). Cross-lingual image caption generation based on visual attention model. *IEEE Access*, *8*, 104543–104554.

Wang, L., Li, Y., Huang, J., & Lazebnik, S. (2018). Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 394–407.

Wang, T., Yuan, X., & Trischler, A. (2017). A joint model for question answering and question generation. *CoRR*, *abs/1706.01450*.

Wang, X., Suhara, Y., Nuno, N., Li, Y., Li, J., Carmeli, N., Angelidis, S., Kandogann, E., & Tan, W.-C. (2020). Extremereader: An interactive explorer for customizable and explainable review summarization. In *Companion Proceedings of the Web Conference 2020*, pp. 176–180.

Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.-F., & Wang, W. Y. (2019). VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Wang, Y., Zhang, J., Zhai, F., Xu, J., & Zong, C. (2018). Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2955–2960. Association for Computational Linguistics.

Wang, Z., Wang, X., An, B., Yu, D., & Chen, C. (2020). Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1072–1086.

Wei, L., Zhang, J., Hou, J., & Dai, L. (2020). Attentive fusion enhanced audio-visual encoding for transformer based robust speech recognition. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., & Grave, E. (2020). CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4003–4012.

Wiriyathammabhum, P., Summers-Stay, D., Fermüller, C., & Aloimonos, Y. (2016). Computer vision and natural language processing: Recent approaches in multimedia and robotics. *ACM Comput. Surv.*, *49*(4).

Wiseman, S., Shieber, S., & Rush, A. (2018). Learning neural templates for text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3174–3187, Brussels, Belgium. Association for Computational Linguistics.

Wu, J., Hu, Z., & Mooney, R. (2019). Generating question relevant captions to aid visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3585–3594.

Wu, J., & Mooney, R. (2019). Faithful multimodal explanation for visual question answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 103–112.

Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT?. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pp. 120–130.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation..

Wu, Z., Ive, J., Wang, J., Madhyastha, P., & Specia, L. (2019). Predicting actions to help predict translations. In *The How2 Challenge: New Tasks for Vision & Language, ICML 2019 Workshop*, Long Beach, CA, USA.

Xia, C., Zhang, C., Nguyen, H., Zhang, J., & Yu, P. S. (2020a). CG-BERT: conditional text generation with BERT for generalized few-shot intent detection. *CoRR*, *abs/2004.01881*.

Xia, P., Wu, S., & Van Durme, B. (2020b). Which* bert? a survey organizing contextualized encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7516–7533.

Xu, J., Hou, J., Song, Y., Guo, W., & Dai, L. (2019). Knowledge distillation from multilingual and monolingual teachers for end-to-end multilingual speech recognition. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 844–849.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In Bach, F., & Blei, D. (Eds.), *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37 of *Proceedings of Machine Learning Research*, pp. 2048–2057.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, *4*, 401–415.

Xu, W., Li, C., Lee, M., & Zhang, C. (2020). Multi-task learning for abstractive text summarization with key information guide network. *EURASIP J. Adv. Signal Process.*, *2020*(1), 16.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (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*, pp. 483–498.

Yang, X., Liu, Y., Xie, D., Wang, X., & Balasubramanian, N. (2019). Latent part-of-speech sequences for neural machine translation. 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)*, pp. 780–790, Hong Kong, China. Association for Computational Linguistics.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., hsuan Sung, Y., Strope, B., & Kurzweil, R. (2020). Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Yang, Y., Zhang, Y., Tar, C., & Baldridge, J. (2019). PAWS-x: A cross-lingual adversarial dataset for paraphrase identification. 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)*.

Yang, Z., Chen, W., Wang, F., & Xu, B. (2018). Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46–55, Melbourne, Australia. Association for Computational Linguistics.

Yang, Z., Gao, Y., Wang, W., & Ney, H. (2020). Predicting and using target length in neural machine translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*.

Yi, J., Tao, J., Wen, Z., & Bai, Y. (2018). Adversarial multilingual training for low-resource speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4899–4903.

Yu, Z., Yu, J., Cui, Y., Tao, D., & Tian, Q. (2019). Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6281–6290.

Yue, X., Lee, G., Yilmaz, E., Deng, F., & Li, H. (2019). End-to-end code-switching asr for low-resourced language pairs. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 972–979. IEEE.

Zeng, Z., Khassanov, Y., Pham, V. T., Xu, H., Chng, E. S., & Li, H. (2018). On the end-to-end solution to mandarin-english code-switching speech recognition. In *arXiv preprint arXiv:1811.00241*.

Zhang, B., Xiong, D., Su, J., Duan, H., & Zhang, M. (2016). Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 521–530, Austin, Texas. Association for Computational Linguistics.

Zhang, H., Ghosh, S., Heck, L. P., Walsh, S., Zhang, J., Zhang, J., & Kuo, C.-C. J. (2019). Generative visual dialogue system via weighted likelihood estimation.. In *IJCAI*, pp. 1025–1031.

Zhang, R., Zhang, C., He, Z., Wu, H., & Wang, H. (2020a). Learning adaptive segmentation policy for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2280–2289, Online. Association for Computational Linguistics.

Zhang, Z., Chen, K., Wang, R., Utiyama, M., Sumita, E., Li, Z., & Zhao, H. (2020b). Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.