Chapter 2
State-of-the-Art in Language Technology and Language-centric Artificial Intelligence

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Jon Ander Campos, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernáez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Ander Salaberria, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa

Abstract This chapter landscapes the field of Language Technology (LT) and language-centric AI by assembling a comprehensive state-of-the-art of basic and applied research in the area. It sketches all recent advances in AI, including the most recent deep learning neural technologies. The chapter brings to light not only where language-centric AI as a whole stands, but also where the required resources should be allocated to place European LT at the forefront of the AI revolution. We identify key research areas and gaps that need to be addressed to ensure LT can overcome the current inequalities.[1]

1 Introduction

Interest in the computational processing of human languages led to the establishment of specialised fields known as Computational Linguistics (CL), Natural Language Processing (NLP) and Language Technology (LT). CL is more informed by linguis-

Rodrigo Agerri · Eneko Agirre · Itziar Aldabe · Nora Aranberri · Jose Maria Arriola · Aitziber Atutxa · Gorka Azkune · Jon Ander Campos · Arantza Casillas · Ainara Estarrona · Aritz Farwell · Iakes Goenaga · Josu Goikoetxea · Koldo Gojenola · Inma Hernaez · Mikel Iruskieta · Gorka Labaka · Oier Lopez de Lacalle · Eva Navas · Maite Oronoz · Arantxa Otegi · Alicia Pérez · Olatz Perez de Viñaspre · German Rigau · Ander Salaberria · Jon Sanchez · Ibon Saratxaga · Aitor Soroa
University of the Basque Country, Spain,
rodrigo.agerri@ehu.eus, e.agirre@ehu.eus, itziar.aldabe@ehu.eus, nora.aranberri@ehu.eus, josemaria.arriola@ehu.eus, aitziber.atutxa@ehu.eus, gorka.azkune@ehu.eus, jonander.campos@ehu.eus, arantza.casillas@ehu.eus, ainara.estarrona@ehu.eus, aritz.farwell@ehu.eus, iakes.goenaga@ehu.eus, josu.goikoetxea@ehu.eus, koldo.gojenola@ehu.eus, inma.hernaez@ehu.eus, mikel.iruskieta@ehu.eus, gorka.labaka@ehu.eus, oier.lopezdelacalle@ehu.eus, eva.navas@ehu.eus, maite.oronoz@ehu.eus, arantza.otegi@ehu.eus, alicia.perez@ehu.eus, olatz.perezdevinaspre@ehu.eus, german.rigau@ehu.eus, ander.salaberria@ehu.eus, jon.sanchez@ehu.eus, ibon.saratxaga@ehu.eus, a.soroa@ehu.eus

[1] This chapter is an abridged version of Agerri et al. (2021).

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_2

tics and NLP by computer science, LT is a more neutral term. In practice, these communities work closely together, sharing the same publishing venues and conferences, combining methods and approaches inspired by both, and together making up language-centric AI. In this chapter we treat them interchangeably.

Over the years, LT has developed different methods to make the information contained in written and spoken language explicit or to generate or synthesise written or spoken language. Despite the inherent difficulties in many of the tasks performed, current LT support allows many advanced applications which were unthinkable only a few years ago. LT is present in our daily lives, for example, through search engines, recommendation systems, virtual assistants, chatbots, text editors, text predictors, automatic translation systems, automatic subtitling, automatic summarisation and inclusive technology. Its recent accelerated development promises even more encouraging and exciting results in the near future.

This state-of-the-art in LT and language-centric AI begins with a brief historical account in Section 2 on the development of the field from its inception through the current deep learning era. The following three sections are neural language models (Section 3), research areas (Section 4) and LT beyond language (Section 5). They offer a survey that maps today's LT and language-centric AI landscape. Finally, a discussion and various conclusions are outlined in Section 6.

2 Language Technology: Historical Overview

2.1 A Brief History

The 1950s mark the beginning of Language Technology as a discipline. In the middle of the 20th century, Alan Turing proposed his famous test, which defines a criterion to determine whether a machine can be considered intelligent (Turing 1950). A few years later, Noam Chomsky laid the foundations to formalise, specify and automate linguistic rules with his generative grammar (Chomsky 1957). For a long period of time, the horizon defined by Turing and the instrument provided by Chomsky influenced the majority of NLP research.

The early years of LT were closely linked to Machine Translation (MT), a well-defined task, and also relevant from a political and strategic point of view. In the 1950s it was believed that a high-quality automatic translator would be available soon. By the mid-1960s, however, the Automatic Language Processing Advisory Committee (ALPAC) report revealed the true difficulty of the task and NLP in general. The following two decades were heavily influenced by Chomsky's ideas, with increasingly complex systems of handwritten rules. At the end of the 1980s, a revolution began which irreversibly changed the field of NLP. This change was driven mainly by four factors: 1. the clear definition of individual NLP tasks and corresponding rigorous evaluation methods; 2. the availability of relatively large amounts of data; 3. machines that could process these large amounts of data; and 4. the gradual introduction of more robust approaches based on statistical methods and machine learning (ML), that would pave the way for subsequent major developments.

Since the 1990s, NLP has moved forward with new resources, tools and applications. An effort was made to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc., from which WordNet (Miller 1992) is one of the main results. Data-driven systems displaced rule-based systems, leading to the almost ubiquitous presence of ML components in NLP systems. In the 2010s we observed a radical technological shift in NLP. Collobert et al. (2011) presented a multilayer neural network (NN) adjusted by backpropagation that solved various sequential labeling problems. Word embeddings gained particular relevance due to their role in the incorporation of pre-trained external knowledge into neural architectures (Mikolov et al. 2013). Large volumes of unannotated texts, together with progress in self-supervised ML and the rise of high-performance hardware (Graphics Processing Units, GPU), enabled highly effective deep learning systems to be developed across a range of application areas. These and other breakthroughs helped launch today's Deep Learning Era.

2.2 The Deep Learning Era

Today, LT is moving away from a methodology in which a pipeline of multiple modules is utilised to implement solutions to architectures based on complex neural networks trained on vast amounts of data. Four research trends are converging: 1. mature deep neural network technology, 2. large amounts of multilingual data, 3. increased High Performance Computing (HPC) power, and 4. the application of simple but effective self-learning approaches (Devlin et al. 2019; Yinhan Liu et al. 2020). These advancements have produced a new state-of-the-art through systems that are claimed to obtain human-level performance in laboratory benchmarks on difficult language understanding tasks. As a result, various large IT enterprises have started deploying large language models (LLMs) in production.

Despite their notable capabilities, however, LLMs have certain drawbacks that will require interdisciplinary collaboration and research to resolve. First, we have no clear understanding of how they work, when they fail, or what emergent properties they present. Indeed, some authors call these models "foundation models" to underscore their critically central yet incomplete character (Bommasani et al. 2021). Second, the systems are very sensitive to phrasing and typos, are not robust enough, and perform inconsistently (Ribeiro et al. 2019). Third, these models are expensive to train, which means that only a limited number of organisations can currently afford their development (Ahmed and Wahed 2020). Fourth, large NLP datasets used to train these models have been 'filtered' to remove targeted minorities (Dodge et al. 2021). In addition, LLMs can sometimes produce unpredictable and factually inaccurate text or even recreate private information. Finally, computing large pre-trained models comes with a substantial carbon footprint (Strubell et al. 2019).

The implications of LLMs may extend to questions of language-centred AI sovereignty. Given the impact of LT in everyone's daily lives, many LT practitioners are particularly concerned by the need for digital language equality (DLE) across all aspects of our societies. As expected, only a small number of the world's more than 6,000 languages are represented in the rapidly evolving LT field. This disproportionate representation is further exacerbated by systematic inequalities in LT across the world's languages (Joshi et al. 2020). Interestingly, the application of zero-shot to few-shot transfer learning with multilingual pre-trained language models, prompt learning and self-supervised systems opens a path to leverage LT for less-developed languages. However, the development of these new LT systems will require resources along with carefully designed evaluation benchmarks and annotated datasets for every language and domain of application.

Forecasting the future of LT and language-centric AI is a challenge. It is, nevertheless, safe to assume that many more advances will be achieved utilising pre-trained language models and that they will substantially impact society. Future users are likely to discover novel applications and wield them positively or negatively. In either case, as Bender et al. (2021) argue, it is important to understand the current limitations of LLMs, which they refer to as "stochastic parrots". Focusing on state-of-the-art results exclusively with the help of leaderboards, without encouraging deeper understanding of the mechanisms by which they are attained, can give rise to misleading conclusions. These, in turn, may direct resources away from efforts that would facilitate long-term progress towards multilingual, efficient, accurate, explainable, ethical and unbiased language understanding and communication.

3 Neural Language Models

LT is undergoing a paradigm shift with the rise of neural language models that are trained on broad data at scale and are adaptable to a wide range of monolingual and multilingual downstream tasks (Devlin et al. 2019; Yinhan Liu et al. 2020). These models are based on standard self-supervised deep learning and transfer learning, but their scale results in emergent and surprising capabilities. One of the advantages is their ability to alleviate the feature engineering problem by using low-dimensional and dense vectors (distributed representation) to implicitly represent the language examples (Collobert et al. 2011). In self-supervised learning, the language model is derived automatically from large volumes of unannotated language data (text or voice). There has been considerable progress in self-supervised learning since word embeddings associated word vectors with context-independent vectors.

With transfer learning, the learning process starts from patterns that have been learned when solving a different problem, i. e., leveraging previous learning to avoid starting from scratch. Within deep learning, pre-training is the dominant approach to transfer learning: the objective is to pre-train a deep Transformer model on large amounts of data and then reuse this pre-trained language model by fine-tuning it on small amounts of (usually annotated) task-specific data. Recent work has shown that pre-trained language models can robustly perform tasks in a few-shot or even zero-shot fashion when given an adequate task description in its natural language prompt (Brown et al. 2020). Unlike traditional supervised learning, which trains a model to take in an input and predict an output, prompt-based learning or in-context learning is based on exploiting pre-trained language models to solve a task using text directly. This framework is very promising since some NLP tasks can be solved in a fully unsupervised fashion by providing a pre-trained language model with task descriptions in natural language (Raffel et al. 2020). Surprisingly, fine-tuning pre-trained language models on a collection of tasks described via instructions (or prompts) substantially boosts zero-shot performance on unseen tasks (Wei et al. 2021).

Multilingual Large Language Models (MLLMs) such as mBERT (Devlin et al. 2019), XLM-R (Conneau et al. 2020), mBART (Yinhan Liu et al. 2020), mT5 (Xue et al. 2021), etc. have emerged as viable options for bringing the power of pre-training to a large number of languages. For example, mBERT is pre-trained on Wikipedia corpora in 104 languages. mBERT can generalise cross-lingual knowledge in zero-shot scenarios. This indicates that even with the same structure of BERT, using multilingual data can enable the model to learn cross-lingual representations. The surprisingly good performance of MLLMs in cross-lingual transfer as well as bilingual tasks suggests that these language models are learning universal patterns (Doddapaneni et al. 2021). Thus, one of the main motivations of training MLLMs is to enable transfer from high-resource languages to low-resource languages.

New types of processing pipelines and toolkits have arisen in recent years due to the fast-growing collection of efficient tools. Libraries that are built with NN components are increasingly common, including pre-trained models that perform multilingual NLP tasks. Neural language models are adaptable to a wide spectrum of monolingual and multilingual tasks. These models are currently often considered black boxes, in that their inner mechanisms are not clearly understood. Nonetheless, Transformer architectures may present an opportunity to offer advances to the broader LT community if certain obstacles can be successfully overcome. One is the question of the resources needed to design the best-performing neural language models, currently done almost exclusively at large IT companies. Another is the problem of stereotypes, prejudices and personal information within the corpora used to train the models. The predominance of English as the default language in NLP can be successfully addressed if there is sufficient will and coordination. The continued consolidation of large infrastructures will help determine how this is accomplished in the near future. Their successful implementation would mark a crucial first step towards the development, proliferation and management of language resources for *all* European languages. This capability would, in turn, enable Europe's languages to enjoy full and equal access to digital language technology.

4 Research Areas

Section 4 introduces some of the more prominent research areas in the field: Language Resources (Section 4.1), Text Analysis (Section 4.2), Speech Processing (Section 4.3), Machine Translation (Speech 4.4), Information Extraction and Retrieval (Section 4.5), NLG and Summarisation (Section 4.6) as well as HCI (Section 4.7).

4.1 Language Resources

The term Language Resource (LR) refers to a set of speech or written data and descriptions in machine readable form. These are utilised for building, improving or evaluating text- and speech-based algorithms or systems. They also serve as resources for the software localisation and language services industries, language studies, digital publishing, international transactions, subject-area specialists and end users. Although no widely standardised typology of LRs exists, they are usually classified as: 1. Data (i. e., corpora and lexical/conceptual resources); 2. Tools/Services (i. e., linguistic annotations; tools for creating annotations; search and retrieval applications; applications for automatic annotation) and 3. Metadata and vocabularies (i. e., vocabularies or repositories of linguistic terminology; language metadata). In this section we will focus on the first two categories.

A main objective of the LR community is the development of infrastructures and platforms for presenting and disseminating LRs. There are numerous repositories in which resources for each language are documented. Among the major European catalogues are European Language Grid (ELG, Rehm 2023),[2] ELRC-SHARE, [3] European Language Resources Association (ELRA), [4] Common Language Resources and Technology Infrastructure (CLARIN)[5] and META-SHARE.[6] The Linguistic Data Consortium,[7] which operates outside of Europe, should also be highlighted.

In addition, there are several relevant multilingual public domain initiatives. Among these are the Common Voice Project,[8] designed to encourage the development of ASR systems; the M-AILABS Speech Dataset,[9] for text-to-speech synthesis; the Ryerson Audio-Visual Database of Emotional Speech and Song,[10] for research

[2] https://www.european-language-grid.eu

[3] http://www.elrc-share.eu

[4] http://catalogue.elra.info

[5] https://www.clarin.eu/content/language-resources

[6] http://www.meta-share.org

[7] https://catalog.ldc.upenn.edu

[8] https://commonvoice.mozilla.org

[9] https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/

[10] https://zenodo.org/record/1188976

on emotional multimedia content; and LibriVox,[11] an audiobook repository that can be used in different research fields and applications.

A cursory glance at these repositories not only gives us an idea of the amount of resources available for Europe's languages, but also reveals the clear inequality between official and minority languages. Moreover, although the four European languages with the most resources are English, French, German and Spanish, English is far ahead of the rest, with more than twice as many resources as the next language (see Figure 1, p. 50). At the same time, the languages without official status trail significantly behind in terms of LR development, demonstrating the critical impact that official status has on the extent of available resources.

4.2 Text Analysis

Text Analysis (TA) aims to extract relevant information from large amounts of unstructured text in order to enable data-driven approaches to manage textual content. In other words, its purpose is to create structured data out of unstructured text content by identifying entities, facts and relationships that are buried in the textual data. TA employs a variety of methodologies to process text. It is crucial for establishing "who did what, where and when," a technology that has proven to be key for applications such as Information Extraction, Question Answering, Summarisation and nearly every linguistic processing task involving semantic interpretation, including Opinion Mining and Aspect-based Sentiment Analysis (ABSA).

The best results for TA tasks are generally obtained by means of supervised, corpus-based approaches. In most cases, manually annotating text for every single specific need is extremely time-consuming and not affordable in terms of human resources and economic costs. To make the problem more manageable, TA is addressed in several tasks that are typically performed in order to preprocess the text to extract relevant information. The most common tasks currently available in state-of-the-art NLP tools and pipelines include Part-of-Speech (POS) tagging, Lemmatisation, Word Sense Disambiguation (WSD), Named Entity Recognition (NER), Named Entity Disambiguation (NED) or Entity Linking (EL), Parsing, Coreference Resolution, Semantic Role Labelling (SRL), Temporal Processing, ABSA and, more recently, Open Information Extraction (OIE).

Today, all these tasks are addressed in an end-to-end manner, i. e., even for a traditionally complex task such as Coreference Resolution (Pradhan et al. 2012), current state-of-the-art systems are based on an approach in which no extra linguistic annotations are required. These systems typically employ LLMs. Similarly, most state-of-the-art TA toolkits, including AllenNLP and Trankit, among others (Gardner et al. 2018; M. V. Nguyen et al. 2021), use a highly multilingual end-to-end approach. Avoiding intermediate tasks has helped to mitigate the common cascading errors problem that was pervasive in more traditional TA pipelines. As a consequence, the

[11] https://librivox.org

appearance of end-to-end systems has helped bring about a significant jump in performance across every TA task.

4.3 Speech Processing

Speech processing aims at allowing humans to communicate with digital devices through voice. This entails developing machines that understand and generate not only oral messages, but also all the additional information that we can extract from the voice, like who is speaking, their age, their personality, their mood, etc. Some of the main areas in speech technology are text-to-speech synthesis (TTS), automatic speech recognition (ASR) and speaker recognition (SR).

TTS attempts to produce the oral signal that corresponds to an input text with an intelligibility, naturalness and quality similar to a natural speech signal. Statistical parametric speech synthesis techniques generate speech by means of statistical models trained to learn the relation between linguistic labels derived from text and acoustic parameters extracted from speech by means of a vocoder. HMM (Hidden Markov Models) and more recently DNN (Deep Neural Networks) have been used as statistical frameworks. Various architectures have been tested, such as feed-forward networks (Qian et al. 2014), recurrent networks (Y. Fan et al. 2014) and WaveNet (Oord et al. 2016). Among the criteria used for training, the most common is minimum generation error (Z. Wu and King 2016), although recently new methods based on Generative Adversarial Networks (GAN, Saito et al. 2017) have been proposed with excellent results in terms of naturalness of the produced voice.

ASR, producing a transcription from a speech signal, has been long sought after. The intrinsic difficulty of the task has required a step-by-step effort, with increasingly ambitious objectives. Only in the last two decades has this technology jumped from the laboratory to production. The first commercial systems were based on statistical models, i. e., HMMs (Juang and Rabiner 2005; Gales and Young 2008). While this technology was the standard during the first decade of the century, in the 2010s, the increase in computing power and the ever-growing availability of training data allowed for the introduction of DNN techniques for ASR.

More recently, end-to-end or fully differentiable architectures have appeared that aim to simplify a training process that is capable of exploiting the available data. In these systems, a DNN maps the acoustic signal in the input directly to the textual output. Thus, the neural network models the acoustic information, the time evolution and some linguistic information, learning everything jointly. New architectures, in the form of Transformers (Gulati et al. 2020; Xie Chen et al. 2021) and teacher-student schemes (Z. Zhang et al. 2020; Jing Liu et al. 2021), have been applied to ASR with great success. Recently, Whisper, a Transformer sequence-to-sequence model trained on very large amounts of data that can perform several tasks such as multilingual ASR, translation and language identification, has been developed by OpenAI (Radford et al. 2022) showing the potential of weakly supervised systems.

A similar evolution has taken place in the area of SR. Part of the widespread emergence of biometric identification techniques, exemplified by the now commonplace ability to unlock a smartphone with a fingerprint or an iris, speaker recognition involves the automatic identification of people based on their voice. Nowadays, the classical systems have been outperformed by end-to-end neural network based systems, which are being improved using widespread databases (Nagrani et al. 2017) and enforcing research (Nagrani et al. 2020), obtaining better recognition rates by means of new network architectures and techniques (Safari et al. 2020; H. Zhang et al. 2020; R. Wang et al. 2022).

4.4 Machine Translation

Machine Translation (MT) is the automatic translation from one natural language into another. Since its first implementation (Weaver 1955) it has remained a key application in LT/NLP. While a number of approaches and architectures have been proposed and tested over the years, Neural MT (NMT) has become the most popular paradigm for MT development both within the research community (Vaswani et al. 2018; Yinhan Liu et al. 2020; Zhu et al. 2020; Sun et al. 2022) and for large-scale production systems (Y. Wu et al. 2016). This is due to the good results achieved by NMT systems, which attain state-of-the-art results for many language pairs (Akhbardeh et al. 2021; Adelani et al. 2022; Min 2023). NMT systems use distributed representations of the languages involved, which enables end-to-end training of systems. If we compare them with classical statistical MT models (Koehn et al. 2003), we see that they do not require word aligners, translation rule extractors, and other feature extractors; the *embed – encode – attend – decode* paradigm is the most common NMT approach (Vaswani et al. 2017; You et al. 2020; Dione et al. 2022).

Thanks to current advances in NMT it is common to find systems that can easily incorporate multiple languages simultaneously. We refer to these types of systems as Multilingual NMT (MNMT) systems. The principal goal of an MNMT system is to translate between as many languages as possible by optimising the linguistic resources available. MNMT models (Aharoni et al. 2019; B. Zhang et al. 2020; Emezue and Dossou 2022; Siddhant et al. 2022) are interesting for several reasons. On the one hand, they can address translations among all the languages involved within a single model, which significantly reduces training time and facilitates deployment of production systems. On the other hand, by reducing operational costs, multilingual models achieve better results than bilingual models for low- and zero-resource language pairs: training is performed jointly and this generates a positive transfer of knowledge from high(er)-resource languages (Aharoni et al. 2019; Arivazhagan et al. 2019). This phenomenon is known as translation knowledge transfer or transfer learning (Zoph et al. 2016; T. Q. Nguyen and Chiang 2017; Hujon et al. 2023).

For instance, A. Fan et al. (2021) have created several MNMT models by building a large-scale many-to-many dataset for 100 languages. They significantly reduce the complexity of this task, employing automatic building of parallel corpora (Artetxe and Schwenk 2019; Schwenk et al. 2021) with a novel data mining strategy that exploits language similarity in order to avoid mining all directions. The method allows for direct translation between 100 languages without using English as a pivot and it performs as well as bilingual models on many competitive benchmarks. Additionally, they take advantage of backtranslation to improve the quality of their model on zero-shot and low-resource language pairs.

4.5 Information Extraction and Information Retrieval

Deep learning has had a tremendous impact on Information Retrieval (IR) and Information Extraction (IE). The goal of IR is to meet the information needs of users by providing them with documents or text snippets that contain answers to their queries. IR is a mature technology that enabled the development of search engines. The area has been dominated by classic methods based on vector space models that use manually created sparse representations such as TF-IDF or BM25 (Robertson and Zaragoza 2009), but recent approaches that depend on dense vectors and deep learning have shown promising results (Karpukhin et al. 2020; Izacard and Grave 2021). Dense representations are often combined with Question Answering (QA) to develop systems that are able to directly answer specific questions posed by users, either by pointing at text snippets that answer the questions (Karpukhin et al. 2020; Izacard and Grave 2021) or by generating the appropriate answers themselves (P. Lewis et al. 2021).

IE aims to extract structured information from text. Typically, IE systems recognise the main events described in a text, as well as the entities that participate in those events. Modern techniques mostly focus on two challenges: learning textual semantic representations for events in event extraction (both at sentence and document level) and acquiring or augmenting labeled instances for model training (K. Liu et al. 2020). Regarding the former, early approaches relied on manually coded lexical, syntactic and kernel-based features (Ahn 2006). With the development of deep learning, however, researchers have employed neural networks, including CNNs (Y. Chen et al. 2015), RNNs (T. H. Nguyen and Grishman 2016) and Transformers (Yang et al. 2019). Data augmentation has been typically performed by using methods such as distant supervision or employing data from other languages to improve IE on the target language, which is especially useful when the target language is under-resourced. Deep learning techniques utilised in NMT (Jian Liu et al. 2018) and pre-trained multilingual LLMs (Jian Liu et al. 2019) have also helped in this task.

Another important task within IE is Relation Extraction (RE), whose goal is to predict the semantic relationship between two entities, if any. The best results on RE are obtained by fine-tuning LLMs, which are supplied with a classification head. One of the most pressing problems in RE is the scarcity of manually annotated examples in real-world applications, particularly when there is a domain and language shift. In recent years, new methods have emerged that only require a few-shot or zero-shot examples. Prompt-based learning, e. g., uses task and label verbalisations that can be designed manually or learned automatically (Schick and Schütze 2021) as an alternative to fine-tuning. In these methods, the inputs are augmented with prompts and the LM objective is used in learning and inference. This paradigm shift has allowed IE tasks to be framed as a QA problem (Sulem et al. 2022) or as a constrained text generation problem (S. Li et al. 2021) using prompts, questions or templates.

4.6 Natural Language Generation and Summarisation

Natural Language Generation (NLG) has become one of the most important and challenging tasks in NLP (Gehrmann et al. 2021). NLG automatically generates understandable texts, typically using a non-linguistic or textual representation of information as input (Reiter and Dale 1997; Gatt and Krahmer 2018; Junyi Li et al. 2021a). Applications that generate new texts from existing text include MT from one language to another (see Section 4.4), fusion and summarisation, simplification, text correction, paraphrase generation, question generation, etc. With the recent resurgence of deep learning, new ways to solve text generation tasks based on different neural architectures have arisen (Junyi Li et al. 2021b). One advantage of these neural models is that they enable end-to-end learning of semantic mappings from input to output in text generation. Existing datasets for most supervised text generation tasks are small (except MT). Therefore, researchers have proposed various methods to solve text generation tasks based on LLMs. Transformer models such as T5 (Raffel et al. 2020) and BART (M. Lewis et al. 2020) or a single Transformer decoder block such as GPT (Brown et al. 2020) are currently standard architectures for generating high quality text.

Due to the rapid growth of information generated daily online (Gambhir and Gupta 2017), there is a growing need for automatic summarisation techniques that produce short texts from one or more sources efficiently and precisely. Several extractive approaches have been developed for automatic summary generation that implement a number of machine learning and optimisation techniques (J. Xu and Durrett 2019). Abstractive methods are more complex as they require NLU capabilities. Abstractive summarisation produces an abstract with words and phrases that are based on concepts that occur in the source document (Du et al. 2021). Both approaches can now be modeled using Transformers (Yang Liu and Lapata 2019).

4.7 Human-Computer Interaction

The demand for technologies that enable users to interact with machines at any time utilising text and speech has grown, motivating the use of dialogue systems. Such systems allow the user to converse with computers using natural language and include Siri, Google Assistant, Amazon Alexa, and ChatGPT, among others. Dialogue systems can be divided into three groups: task-oriented systems, conversational agents (also known as chatbots) and interactive QA systems.

The distinguishing features of task-oriented dialogue systems are that they are designed to perform a concrete task in a specific domain and that their dialogue flow is defined and structured beforehand. For example, such systems are used to book a table at a restaurant, call someone or check the weather forecast. The classical implementation of this type of system follows a pipeline architecture based on three modules: the NLU module, the dialogue manager and the NLG module. While classical dialogue systems trained and evaluated these modules separately, more recent systems rely on end-to-end trainable architectures based on neural networks (Bordes et al. 2017; Hosseini-Asl et al. 2020).

Conversational agents enable engaging open-domain conversations, often by emulating the personality of a human (S. Zhang et al. 2018). The Alexa prize,[12] for instance, focused on building agents that could hold a human in conversation as long as possible. These kinds of agents are typically trained in conversations mined from social media using end-to-end neural architectures (Roller et al. 2021).

Interactive QA systems try to respond to user questions by extracting answers from either documents (Rajpurkar et al. 2018) or knowledge bases (T. Yu et al. 2018). In order to be able to have meaningful interactions, interactive QA systems have a simple dialogue management procedure taking previous questions and answers into account (Choi et al. 2018). The core technology is commonly based on LLMs (Qiu et al. 2020) where some mechanism is included to add context representation (Vakulenko et al. 2021).

5 Language Technology beyond Language

Knowledge about our surrounding world is required to properly understand natural language utterances (Bender and Koller 2020). That knowledge is known as world knowledge and many authors argue that it is a key ingredient to achieve human-level NLU (Storks et al. 2019). One of the ways to acquire this knowledge is to explore the visual world together with the textual world (Elu et al. 2021). CNNs have been the standard architecture for generating representations for images (LeCun and Bengio 1995) during the last decade. Recently, self-attention-based Transformer models (Vaswani et al. 2017) have emerged as an alternative architecture, leading to exciting progress on a number of vision tasks (Khan et al. 2021). Compared to previous approaches, Transformers allow multiple modalities to be processed (e. g., images, videos, text and speech) using similar processing blocks and demonstrate excellent scalability properties. Encoder-decoder models in particular have been gaining traction recently due to their versatility on solving different generative tasks (Junnan Li et al. 2022; Xi Chen et al. 2022).

[12] https://developer.amazon.com/alexaprize

Regarding downstream tasks, caption generation is a typical visio-linguistic task, where a textual description of an image must be generated. The first approaches to solve this problem combined CNNs with RNNs in an encoder-decoder architecture (Vinyals et al. 2015). Further improvements were achieved when attention was included (K. Xu et al. 2015) and some researchers have proposed utilising object-based attention instead of spatial attention (Anderson et al. 2018). Although it is not currently clear which attention mechanism is better, the quality of the text generated by these models is high as measured by metrics such as BLEU (Papineni et al. 2002) and METEOR (Banerjee and Lavie 2005)

Visual generation, in contrast to caption generation, requires an image to be generated from a textual description. One of this task's most significant challenges is to develop automatic metrics to evaluate the quality of the generated images and their coherence with the input text. The first effective approaches were based on Generative Adversarial Networks (Goodfellow et al. 2014) and Variational Autoencoders (Kingma and Welling 2013). Cho et al. (2020) demonstrate that multimodal Transformers can also generate impressive images from textual input. Nevertheless, novel advancements in diffusion models (Sohl-Dickstein et al. 2015; Ho et al. 2020) have defined the current state-of-the-art in image generation (Ramesh et al. 2022). These models learn to iteratively reconstruct noisy images and, recently, their size and computational cost has been reduced as diffusion can be now applied in a reduced latent space instead of an image's pixel space (Rombach et al. 2022).

Another typical task is Visual Question Answering (VQA), where given an image and a question about the contents of that image, the right textual answer must be found. There are many VQA datasets in the literature (Antol et al. 2015; Johnson et al. 2017). Some demand leveraging external knowledge to infer an answer and, thus, they are known as knowledge-based VQA tasks (P. Wang et al. 2017a,b; Marino et al. 2019). These VQA tasks demand skills to understand the content of an image and how it is referred to in the textual question, as well as reasoning capabilities to infer the correct answer. Multimodal Transformers, such as OFA (P. Wang et al. 2022) and PaLI (Xi Chen et al. 2022), define the state-of-the-art in several of these tasks.

Visual Referring Expressions are one of the multimodal tasks that may be considered an extension of a text-only NLP task, i. e., referring expressions (Krahmer and Deemter 2012) in NLG systems. Its objective is to ground a natural language expression to objects in a visual input. There are several approaches to solve this task (Golland et al. 2010; Kazemzadeh et al. 2014). The most recent ones use attention mechanisms to merge both modalities (L. Yu et al. 2018) or are based on multimodal Transformers (Ding et al. 2022).

A natural extension of textual entailment, Visual Entailment is an inference task for predicting whether an image semantically entails a text. Vu et al. (2018) initially proposed a visually-grounded version of the textual entailment task, where an image is augmented to include a textual premise and hypothesis. However, Xie et al. (2019) propose visual entailment, where the premise is an image and the hypothesis is textual. As an alternative to entailment, there are other grounding tasks that classify whether an image and its caption match (Suhr et al. 2018; F. Liu et al. 2022) or tasks that measure the similarity between sentences with visual cues, such as vSTS (Lopez de Lacalle et al. 2020).

Multimodal MT (MMT) seeks to translate natural language sentences that describe visual content in a source language into a target language by taking the visual content as an additional input to the source language sentences (Elliott et al. 2017; Barrault et al. 2018). Different approaches have been proposed to handle MMT, although attention models that associate textual and visual elements with multimodal attention mechanisms are the most common (Huang et al. 2016; Calixto et al. 2017).

6 Conclusions

Language tools and resources have increased and improved since the end of the last century, a process further catalysed by the advent of deep learning and LLMs over the past decade. Indeed, we find ourselves today in the midst of a significant paradigm shift in LT and language-centric AI. This revolution has brought noteworthy advances to the field along with the promise of substantial breakthroughs in the coming years. However, this transformative technology poses problems, from a research advancement, environmental, and ethical perspective. Furthermore, it has also laid bare the acute digital inequality that exists between languages. In fact, as emphasised in this chapter, many sophisticated NLP systems are unintentionally exacerbating this imbalance due to their reliance on vast quantities of data derived mostly from English-language sources. Other languages lag far behind English in terms of digital presence and even the latter would benefit from greater support. Moreover, the striking asymmetry between official and non-official European languages with respect to available digital resources is concerning. The unfortunate truth is that DLE in Europe is failing to keep pace with the newfound and rapidly evolving changes in LT. One need look no further than what is happening today across the diverse topography of state-of-the-art LT and language-centric AI for confirmation of the current linguistic unevenness. The paradox at the heart of LT's recent advances is evident in almost every LT discipline. Our ability to reproduce ever better synthetic voices has improved sharply for well-resourced languages, but dependence on large volumes of high-quality recordings effectively undermines attempts to do the same for low-resource languages. Multilingual NMT systems return demonstrably improved results for low- and zero-resource language pairs, but insufficient model capacity continues to haunt transfer learning because large multilingual datasets are required, forcing researchers to rely on English as the best resourced language.

Nonetheless, we believe this time of technological transition represents an opportunity to achieve full DLE in Europe. There are ample reasons for optimism. Recent research in the field has considered the implementation of cross-lingual transfer learning and multilingual language models for low-resource languages, an example of how the state-of-the-art in LT could benefit from better digital support for low-resource languages.

Forecasting the future of LT and language-centric AI is a challenge. Just a few years ago, nobody would have predicted the recent breakthroughs that have resulted in systems able to deal with unseen tasks or maintaining natural conversations. It is, however, safe to predict that even more advances will be achieved in all LT research areas and domains in the near future. Despite claims of human parity in many LT tasks, *Natural Language Understanding is still an open research problem* far from being solved since all current approaches have severe limitations. Interestingly, the application of zero-shot to few-shot transfer learning with multilingual LLMs and self-supervised systems opens up the way to leverage LT for less developed languages. However, the development of these new LT systems would not be possible without sufficient resources (experts, data, HPC facilities, etc.) as well as the creation of carefully designed and constructed evaluation benchmarks and annotated datasets for every language and domain of application. Focusing on state-of-the-art results exclusively with the help of leaderboards without encouraging deeper understanding of the mechanisms by which they are achieved can generate misleading conclusions, and direct resources away from efforts that would facilitate long-term progress towards multilingual, efficient, accurate, explainable, ethical and unbiased language understanding and communication, to create transparent digital language equality in Europe in all aspects of society, from government to business to citizen.

References

Adelani, David, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R. Costa-jussà, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Alexandre Mourachko, Safiyyah Saleem, Holger Schwenk, and Guillaume Wenzek (2022). "Findings of the WMT'22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages". In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 773–800. https://aclanthology.org/2022.wmt-1.72.

Agerri, Rodrigo, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernaez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa (2021). *Deliverable D1.2 Report on the State of the Art in Language Technology and Language-centric AI*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/LT-state-of-the-art.pdf.

Aharoni, Roee, Melvin Johnson, and Orhan Firat (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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3874–3884. DOI: 10.18653/v1/N19-1388. https://aclanthology.org/N19-1388.

Ahmed, Nur and Muntasir Wahed (2020). "The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research". In: *CoRR* abs/2010.15581. https://arxiv.org/abs/2010.15581.

Ahn, David (2006). “The stages of event extraction”. In: *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*. Sydney, Australia: Association for Computational Linguistics, pp. 1–8. https://aclanthology.org/W06-0901.

Akhbardeh, Farhad, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri (2021). “Findings of the 2021 Conference on Machine Translation (WMT21)”. In: *Proceedings of the Sixth Conference on Machine Translation*. Online: Association for Computational Linguistics, pp. 1–88. https://aclanthology.org/2021.wmt-1.1.

Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang (2018). “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, pp. 6077–6086. DOI: 10.1109/CVPR.2018.00636. http://openaccess.thecvf.com/content%5C_cvpr%5C_2018/html/Anderson%5C_Bottom-Up%5C_and%5C_Top-Down%5C_CVPR%5C_2018%5C_paper.html.

Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh (2015). “VQA: Visual Question Answering”. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, pp. 2425–2433. DOI: 10.1109/ICCV.2015.279. https://doi.org/10.1109/ICCV.2015.279.

Arivazhagan, Naveen, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey (2019). “The missing ingredient in zero-shot neural machine translation”. In: *arXiv preprint arXiv:1903.07091*. https://arxiv.org/abs/1903.07091.

Artetxe, Mikel and Holger Schwenk (2019). “Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond”. In: *Transactions of the Association for Computational Linguistics* 7, pp. 597–610. DOI: 10.1162/tacl_a_00288. https://aclanthology.org/Q19-1038.

Banerjee, Satanjeev and Alon Lavie (2005). “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments”. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 65–72. https://aclanthology.org/W05-0909.

Barrault, Loïc, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank (2018). “Findings of the Third Shared Task on Multimodal Machine Translation”. In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, pp. 304–323. DOI: 10.18653/v1/W18-6402. https://aclanthology.org/W18-6402.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada, pp. 610–623.

Bender, Emily M. and Alexander Koller (2020). “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5185–5198. https://aclanthology.org/2020.acl-main.463.

Bommasani, Rishi et al. (2021). *On the Opportunities and Risks of Foundation Models*. arXiv: 2108.07258 `[cs.LG]`. https://arxiv.org/abs/2108.07258.

Bordes, Antoine, Y-Lan Boureau, and Jason Weston (2017). “Learning End-to-End Goal-Oriented Dialog”. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon,* *France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=S1Bb3D5gg.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). "Language Models are Few-Shot Learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Calixto, Iacer, Qun Liu, and Nick Campbell (2017). "Doubly-Attentive Decoder for Multi-modal Neural Machine Translation". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1913–1924. DOI: 10.18653/v1/P17-1175. https://aclanthology.org/P17-1175.

Chen, Xi, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut (2022). "Pali: A jointly-scaled multilingual language-image model". In: *arXiv preprint arXiv:2209.06794*.

Chen, Xie, Yu Wu, Zhenghao Wang, Shujie Liu, and Jinyu Li (2021). "Developing real-time streaming transformer transducer for speech recognition on large-scale dataset". In: *ICASSP*. IEEE, pp. 5904–5908.

Chen, Yubo, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao (2015). "Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks". 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)*. Beijing, China: Association for Computational Linguistics, pp. 167–176. DOI: 10.3115/v1/P15-1017. https://aclanthology.org/P15-1017.

Cho, Jaemin, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi (2020). "X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8785–8805. DOI: 10.18653/v1/2020.emnlp-main.707. https://aclanthology.org/2020.emnlp-main.707.

Choi, Eunsol, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer (2018). "QuAC: Question Answering in Context". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2174–2184. DOI: 10.18653/v1/D18-1241. https://aclanthology.org/D18-1241.

Chomsky, Noam (1957). *Syntactic structures.* The Hague: Mouton.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (2011). "Natural Language Processing (Almost) from Scratch". In: *Journal of Machine Learning Research* 12, pp. 2493–2537.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. https://aclanthology.org/2020.acl-main.747.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

Ding, Henghui, Chang Liu, Suchen Wang, and Xudong Jiang (2022). “VLT: Vision-Language Transformer and Query Generation for Referring Segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dione, Cheikh M Bamba, Alla Lo, Elhadji Mamadou Nguer, and Sileye Ba (2022). “Low-resource Neural Machine Translation: Benchmarking State-of-the-art Transformer for Wolof<-> French”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6654–6661.

Doddapaneni, Sumanth, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra (2021). “A primer on pretrained multilingual language models”. In: *arXiv preprint arXiv:2107.00676*. https://arxiv.org/abs/2107.00676.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner (2021). “Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus”. In: *arXiv preprint arXiv:2104.08758*.

Du, Zhengxiao, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang (2021). “All NLP Tasks Are Generation Tasks: A General Pretraining Framework”. In: *arXiv preprint arXiv:2103.10360*. https://arxiv.org/abs/2103.10360.

Elliott, Desmond, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia (2017). “Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description”. In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 215–233. DOI: 10.18653/v1/W17-4718. https://aclanthology.org/W17-4718.

Elu, Aitzol, Gorka Azkune, Oier Lopez de Lacalle, Ignacio Arganda-Carreras, Aitor Soroa, and Eneko Agirre (2021). “Inferring spatial relations from textual descriptions of images”. In: *Pattern Recognition* 113, p. 107847.

Emezue, Chris C and Bonaventure FP Dossou (2022). “MMTAfrica: Multilingual machine translation for African languages”. In: *arXiv preprint arXiv:2204.04306*.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin (2021). “Beyond english-centric multilingual machine translation”. In: *Journal of Machine Learning Research* 22.107, pp. 1–48.

Fan, Yuchen, Yao Qian, Feng-Long Xie, and Frank K Soong (2014). “TTS synthesis with bidirectional LSTM based recurrent neural networks”. In: *Fifteenth annual conference of the international speech communication association*.

Gales, Mark and Steve Young (2008). *The application of hidden Markov models in speech recognition*. Now Publishers Inc.

Gambhir, Mahak and Vishal Gupta (2017). “Recent automatic text summarization techniques: a survey”. In: *Artificial Intelligence Review* 47.1, pp. 1–66.

Gardner, Matt, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer (2018). “AllenNLP: A Deep Semantic Natural Language Processing Platform”. In: *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Melbourne, Australia: Association for Computational Linguistics, pp. 1–6. DOI: 10.18653/v1/W18-2501. https://aclanthology.org/W18-2501.

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

Gehrmann, Sebastian, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou (2021). "The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics". In: *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Online: Association for Computational Linguistics, pp. 96–120. DOI: 10.18653/v1/2021.gem-1.10. https://aclanthology.org/2021.gem-1.10.

Golland, Dave, Percy Liang, and Dan Klein (2010). "A Game-Theoretic Approach to Generating Spatial Descriptions". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, pp. 410–419. https://aclanthology.org/D10-1040.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Ed. by Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, pp. 2672–2680. https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

Gulati, Anmol, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang (2020). "Conformer: Convolution-augmented Transformer for Speech Recognition". In: *Interspeech*, pp. 5036–5040.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33, pp. 6840–6851.

Hosseini-Asl, Ehsan, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher (2020). "A simple language model for task-oriented dialogue". In: *Advances in Neural Information Processing Systems* 33, pp. 20179–20191.

Huang, Po-Yao, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer (2016). "Attention-based Multimodal Neural Machine Translation". In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, pp. 639–645. DOI: 10.18653/v1/W16-2360. https://aclanthology.org/W16-2360.

Hujon, Aiusha V, Thoudam Doren Singh, and Khwairakpam Amitab (2023). "Transfer Learning Based Neural Machine Translation of English-Khasi on Low-Resource Settings". In: *Procedia Computer Science* 218, pp. 1–8.

Izacard, Gautier and Edouard Grave (2021). "Distilling Knowledge from Reader to Retriever for Question Answering". In: *International Conference on Learning Representations*. https://openreview.net/forum?id=NTEz-6wysdb.

Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick (2017). "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, pp. 1988–1997. DOI: 10.1109/CVPR.2017.215. https://doi.org/10.1109/CVPR.2017.215.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6282–6293. https://aclanthology.org/2020.acl-main.560.

Juang, Biing-Hwang and Lawrence R Rabiner (2005). "Automatic speech recognition–a brief history of the technology development". In: *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara* 1, p. 67.

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih (2020). "Dense Passage Retrieval for Open-Domain Question Answering". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Pro-*

cessing (EMNLP). Online: Association for Computational Linguistics, pp. 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550. https://aclanthology.org/2020.emnlp-main.550.

Kazemzadeh, Sahar, Vicente Ordonez, Mark Matten, and Tamara Berg (2014). "ReferItGame: Referring to Objects in Photographs of Natural Scenes". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 787–798. DOI: 10.3115/v1/D14-1086. https://aclanthology.org/D14-1086.

Khan, Salman, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah (2021). *Transformers in Vision: A Survey*. arXiv: 2101.01169 [cs.CV]. https://arxiv.org/abs/2101.01169.

Kingma, Diederik P and Max Welling (2013). "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114*.

Koehn, Philipp, Franz J. Och, and Daniel Marcu (2003). "Statistical Phrase-Based Translation". In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 127–133. https://aclanthology.org/N03-1017.

Krahmer, Emiel and Kees van Deemter (2012). "Computational Generation of Referring Expressions: A Survey". In: *Computational Linguistics* 38.1, pp. 173–218. DOI: 10.1162/COLI_a_00088. https://aclanthology.org/J12-1006.

LeCun, Yann and Yoshua Bengio (1995). "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10, p. 1995.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (2020). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. https://aclanthology.org/2020.acl-main.703.

Lewis, Patrick, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel (2021). *PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them*. arXiv: 2102.07033 [cs.CL].

Li, Junnan, Dongxu Li, Caiming Xiong, and Steven Hoi (2022). "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation". In: *International Conference on Machine Learning*. PMLR, pp. 12888–12900.

Li, Junyi, Tianyi Tang, Gaole He, Jinhao Jiang, Xiaoxuan Hu, Puzhao Xie, Zhipeng Chen, Zhuohao Yu, Wayne Xin Zhao, and Ji-Rong Wen (2021a). "TextBox: A Unified, Modularized, and Extensible Framework for Text Generation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 30–39. DOI: 10.18653/v1/2021.acl-demo.4. https://aclanthology.org/2021.acl-demo.4.

Li, Junyi, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen (2021b). "Pretrained Language Model for Text Generation: A Survey". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Zhi-Hua Zhou. Survey Track. International Joint Conferences on Artificial Intelligence Organization, pp. 4492–4499. DOI: 10.24963/ijcai.2021/612. https://doi.org/10.24963/ijcai.2021/612.

Li, Sha, Heng Ji, and Jiawei Han (2021). "Document-Level Event Argument Extraction by Conditional Generation". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 894–908. https://aclanthology.org/2021.naacl-main.69.

Liu, Fangyu, Guy Emerson, and Nigel Collier (2022). "Visual spatial reasoning". In: *arXiv preprint arXiv:2205.00363*.

Liu, Jian, Yubo Chen, Kang Liu, and Jun Zhao (2018). "Event Detection via Gated Multilingual Attention Mechanism". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial*

Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, pp. 4865–4872. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16371.

Liu, Jian, Yubo Chen, Kang Liu, and Jun Zhao (2019). "Neural Cross-Lingual Event Detection with Minimal Parallel Resources". 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 738–748. DOI: 10.18653/v1/D19-1068. https://aclanthology.org/D19-1068.

Liu, Jing, Rupak Vignesh Swaminathan, Sree Hari Krishnan Parthasarathi, Chunchuan Lyu, Athanasios Mouchtaris, and Siegfried Kunzmann (2021). "Exploiting Large-scale Teacher-Student Training for On-device Acoustic Models". In: *Proc. International Conference on Text, Speech and Dialogue (TSD)*.

Liu, Kang, Yubo Chen, Jian Liu, Xinyu Zuo, and Jun Zhao (2020). "Extracting Events and Their Relations from Texts: A Survey on Recent Research Progress and Challenges". In: *AI Open* 1, pp. 22–39. ISSN: 2666-6510. DOI: https://doi.org/10.1016/j.aiopen.2021.02.004. https://www.sciencedirect.com/science/article/pii/S266665102100005X.

Liu, Yang and Mirella Lapata (2019). "Text Summarization with Pretrained Encoders". 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 3730–3740. DOI: 10.18653/v1/D19-1387. https://aclanthology.org/D19-1387.

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer (2020). "Multilingual Denoising Pre-training for Neural Machine Translation". In: *Transactions of the Association for Computational Linguistics* 8, pp. 726–742. DOI: 10.1162/tacl_a_00343. https://aclanthology.org/2020.tacl-1.47.

Lopez de Lacalle, Oier, Ander Salaberria, Aitor Soroa, Gorka Azkune, and Eneko Agirre (2020). "Evaluating Multimodal Representations on Visual Semantic Textual Similarity". In: *Proceedings of the Twenty-third European Conference on Artificial Intelligence, ECAI 2020, June 8-12, 2020, Santiago Compostela, Spain*.

Marino, Kenneth, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi (2019). "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, pp. 3195–3204. DOI: 10.1109/CVPR.2019.00331. http://openaccess.thecvf.com/content%5C_CVPR%5C_2019/html/Marino%5C_OK-VQA%5C_A%5C_Visual%5C_Question%5C_Answering%5C_Benchmark%5C_Requiring%5C_External%5C_Knowledge%5C_CVPR%5C_2019%5C_paper.html.

Mikolov, Tomás, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *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*. Ed. by Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, pp. 3111–3119. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Miller, George A. (1992). "WordNet: A Lexical Database for English". In: *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. https://aclanthology.org/H92-1116.

Min, Zeping (2023). "Attention Link: An Efficient Attention-Based Low Resource Machine Translation Architecture". In: *arXiv preprint arXiv:2302.00340*.

Nagrani, Arsha, Joon Son Chung, Jaesung Huh, Andrew Brown, Ernesto Coto, Weidi Xie, Mitchell McLaren, Douglas A Reynolds, and Andrew Zisserman (2020). "Voxsrc 2020: The second voxceleb speaker recognition challenge". In: *arXiv preprint arXiv:2012.06867*. https://arxiv.org/abs/2012.06867.

Nagrani, Arsha, Joon Son Chung, and Andrew Zisserman (2017). "VoxCeleb: A Large-Scale Speaker Identification Dataset". In: *Interspeech*, pp. 2616–2620.

Nguyen, Minh Van, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen (2021). "Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. ACL, pp. 80–90. DOI: 10.18653/v1/2021.eacl-demos.10. https://aclanthology.org/2021.eacl-demos.10.

Nguyen, Thien Huu and Ralph Grishman (2016). "Modeling Skip-Grams for Event Detection with Convolutional Neural Networks". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 886–891. DOI: 10.18653/v1/D16-1085. https://aclanthology.org/D16-1085.

Nguyen, Toan Q. and David Chiang (2017). "Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 296–301. https://aclanthology.org/I17-2050.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2016). "Wavenet: A generative model for raw audio". In: *arXiv preprint arXiv:1609.03499*. https://arxiv.org/abs/1609.03499.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. https://aclanthology.org/P02-1040.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang (2012). "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes". In: *Proceedings of CoNLL*, pp. 1–40. https://www.aclweb.org/anthology/W12-4501.

Qian, Yao, Yuchen Fan, Wenping Hu, and Frank K Soong (2014). "On the training aspects of deep neural network (DNN) for parametric TTS synthesis". In: *ICASSP*. IEEE, pp. 3829–3833.

Qiu, Xipeng, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang (2020). "Pretrained models for natural language processing: A survey". In: *Science China Technological Sciences* 63.10, pp. 1872–1897.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever (2022). "Robust speech recognition via large-scale weak supervision". In: *arXiv preprint arXiv:-2212.04356*.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of Machine Learning Research* 21.1, pp. 5485–5551.

Rajpurkar, Pranav, Robin Jia, and Percy Liang (2018). "Know What You Don't Know: Unanswerable Questions for SQuAD". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 784–789. https://aclanthology.org/P18-2124.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen (2022). "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125*.

Rehm, Georg, ed. (2023). *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Cham, Switzerland: Springer.

Reiter, Ehud and Robert Dale (1997). "Building applied natural language generation systems". In: *Natural Language Engineering* 3.1, pp. 57–87.

Ribeiro, Marco Tulio, Carlos Guestrin, and Sameer Singh (2019). "Are Red Roses Red? Evaluating Consistency of Question-Answering Models". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 6174–6184. DOI: 10.18653/v1/P19-1621. https://aclanthology.org/P19-1621.

Robertson, Stephen and Hugo Zaragoza (2009). “The Probabilistic Relevance Framework: BM25 and Beyond”. In: *Found. Trends Inf. Retr.* 3.4, pp. 333–389. ISSN: 1554-0669. DOI: 10.1561/1500000019. https://doi.org/10.1561/1500000019.

Roller, Stephen, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston (2021). “Recipes for Building an Open-Domain Chatbot”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 300–325. DOI: 10.18653/v1/2021.eacl-main.24. https://aclanthology.org/2021.eacl-main.24.

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer (2022). “High-resolution image synthesis with latent diffusion models”. In: *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695.

Safari, Pooyan, Miquel India, and Javier Hernando (2020). “Self-attention encoding and pooling for speaker recognition”. In: *Interspeech*, pp. 941–945.

Saito, Yuki, Shinnosuke Takamichi, and Hiroshi Saruwatari (2017). “Statistical parametric speech synthesis incorporating generative adversarial networks”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.1, pp. 84–96.

Schick, Timo and Hinrich Schütze (2021). “It’s Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 2339–2352. DOI: 10.18653/v1/2021.naacl-main.185. https://aclanthology.org/2021.naacl-main.185.

Schwenk, Holger, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan (2021). “CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 6490–6500. DOI: 10.18653/v1/2021.acl-long.507. https://aclanthology.org/2021.acl-long.507.

Siddhant, Aditya, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia (2022). “Towards the Next 1000 Languages in Multilingual Machine Translation: Exploring the Synergy Between Supervised and Self-Supervised Learning”. In: arXiv: 2201.03110 [cs.CL].

Sohl-Dickstein, Jascha, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli (2015). “Deep unsupervised learning using nonequilibrium thermodynamics”. In: *International Conference on Machine Learning*. PMLR, pp. 2256–2265.

Storks, Shane, Qiaozi Gao, and Joyce Y Chai (2019). “Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches”. In: *arXiv preprint arXiv:1904.01172*, pp. 1–60.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2019). “Energy and Policy Considerations for Deep Learning in NLP”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3645–3650. DOI: 10.18653/v1/P19-1355. https://aclanthology.org/P19-1355.

Suhr, Alane, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi (2018). “A corpus for reasoning about natural language grounded in photographs”. In: *arXiv preprint arXiv:1811.00491*.

Sulem, Elior, Jamaal Hay, and Dan Roth (2022). “Yes, No or IDK: The Challenge of Unanswerable Yes/No Questions”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 1075–1085. https://aclanthology.org/2022.naacl-main.79.

Sun, Zewei, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li (2022). “Rethinking document-level neural machine translation”. In: *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3537–3548.

Turing, Alan M. (1950). “Computing Machinery and Intelligence”. In: *Mind* LIX.236, pp. 433–460. ISSN: 0026-4423. eprint: https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf. https://doi.org/10.1093/mind/LIX.236.433.

Vakulenko, Svitlana, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha (2021). “Question rewriting for conversational question answering”. In: *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 355–363.

Vaswani, Ashish, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit (2018). “Tensor2Tensor for Neural Machine Translation”. In: *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*. Boston, MA: Association for Machine Translation in the Americas, pp. 193–199. https://aclanthology.org/W18-1819.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). “Attention is all you need”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010.

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan (2015). “Show and tell: A neural image caption generator”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, pp. 3156–3164. DOI: 10.1109/CVPR.2015.7298935. https://doi.org/10.1109/CVPR.2015.7298935.

Vu, Hoa Trong, Claudio Greco, Aliia Erofeeva, Somayeh Jafaritazehjan, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt (2018). “Grounded Textual Entailment”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2354–2368. https://aclanthology.org/C18-1199.

Wang, Peng, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel (2017a). “Explicit Knowledge-based Reasoning for Visual Question Answering”. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. Ed. by Carles Sierra. ijcai.org, pp. 1290–1296. DOI: 10.24963/ijcai.2017/179. https://doi.org/10.24963/ijcai.2017/179.

Wang, Peng, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel (2017b). “Fvqa: Fact-based visual question answering”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.10, pp. 2413–2427.

Wang, Peng, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang (2022). “Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework”. In: *International Conference on Machine Learning*. PMLR, pp. 23318–23340.

Wang, Rui, Junyi Ao, Long Zhou, Shujie Liu, Zhihua Wei, Tom Ko, Qing Li, and Yu Zhang (2022). “Multi-view self-attention based transformer for speaker recognition”. In: *ICASSP*. IEEE, pp. 6732–6736.

Weaver, Warren (1955). “Translation”. In: *Machine translation of languages* 14.15-23, p. 10.

Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le (2021). “Finetuned Language Models Are Zero-Shot Learners”. In: *arXiv preprint arXiv:2109.01652*. arXiv: 2109.01652 [cs.CL]. https://arxiv.org/abs/2109.01652.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean (2016). “Google’s neural machine translation system: Bridging the gap between human and machine translation”. In: *arXiv preprint arXiv:1609.08144*. https://arxiv.org/abs/1609.08144.

Wu, Zhizheng and Simon King (2016). “Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.7, pp. 1255–1265.

Xie, Ning, Farley Lai, Derek Doran, and Asim Kadav (2019). “Visual entailment: A novel task for fine-grained image understanding”. In: *arXiv preprint arXiv:1901.06706*. https://arxiv.org/abs/1901.06706.

Xu, Jiacheng and Greg Durrett (2019). “Neural Extractive Text Summarization with Syntactic Compression”. 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 3292–3303. DOI: 10.18653/v1/D19-1324. https://aclanthology.org/D19-1324.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio (2015). “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”. In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. Ed. by Francis R. Bach and David M. Blei. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 2048–2057. http://proceedings.mlr.press/v37/xuc15.html.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (2021). “mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41. https://aclanthology.org/2021.naacl-main.41.

Yang, Sen, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li (2019). “Exploring Pre-trained Language Models for Event Extraction and Generation”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5284–5294. DOI: 10.18653/v1/P19-1522. https://aclanthology.org/P19-1522.

You, Weiqiu, Simeng Sun, and Mohit Iyyer (2020). “Hard-Coded Gaussian Attention for Neural Machine Translation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7689–7700. DOI: 10.18653/v1/2020.acl-main.687. https://aclanthology.org/2020.acl-main.687.

Yu, Licheng, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg (2018). “MAttNet: Modular Attention Network for Referring Expression Comprehension”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, pp. 1307–1315. DOI: 10.1109/CVPR.2018.00142. http://openaccess.thecvf.com/content%5C_cvpr%5C_2018/html/Yu%5C_MAttNet%5C_Modular%5C_Attention%5C_CVPR%5C_2018%5C_paper.html.

Yu, Tao, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev (2018). “Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3911–3921. DOI: 10.18653/v1/D18-1425. https://aclanthology.org/D18-1425.

Zhang, Biao, Philip Williams, Ivan Titov, and Rico Sennrich (2020). “Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault. Association for Computational Linguistics, pp. 1628–1639. https://doi.org/10.18653/v1/2020.acl-main.148.

Zhang, Hanyi, Longbiao Wang, Yunchun Zhang, Meng Liu, Kong Aik Lee, and Jianguo Wei (2020). “Adversarial Separation Network for Speaker Recognition.” In: *Interspeech*, pp. 951–955.

Zhang, Saizheng, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston (2018). “Personalizing Dialogue Agents: I have a dog, do you have pets too?” In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2204–2213. DOI: 10.18653/v1/P18-1205. https://aclanthology.org/P18-1205.

Zhang, Ziqiang, Yan Song, Jian-shu Zhang, Ian McLoughlin, and Li-Rong Dai (2020). "Semi-Supervised End-to-End ASR via Teacher-Student Learning with Conditional Posterior Distribution". In: *Interspeech*, pp. 3580–3584.

Zhu, Jinhua, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu (2020). "Incorporating BERT into Neural Machine Translation". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=Hyl7ygStwB.

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight (2016). "Transfer Learning for Low-Resource Neural Machine Translation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1568–1575. DOI: 10.18653/v1/D16-1163. https://aclanthology.org/D16-1163.