EUROPEAN
LANGUAGE
EQUALITY

D1.2

Report on the state of the art in Language Technology and Language-centric AI

Authors	Rodrigo Agerri, 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, Aitor Soroa (HiTZ Center, University of the Basque Country UPV/EHU)
Dissemination level	Public
Date	30-09-2021

About this document

Project	European Language Equality (ELE)
Grant agreement no.	LC-01641480 – 101018166 ELE
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2021, 18 months
Deliverable number	D1.2
Deliverable title	Report on the state of the art in LT and Language-centric AI
Type	Report
Number of pages	71
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 30-09-2021– Actual: 30-09-2021
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.2 Language Technologies and Language-centric AI – State of the Art
Authors	Rodrigo Agerri, 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, Aitor Soroa (HiTZ Center, University of the Basque Country UPV/EHU)
Reviewers	Federico Gaspari (DCU), Sarah McGuinness (DCU), Tereza Vojtechova (CUNI), Andy Way (DCU)
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu © 2021 ELE Consortium

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
5	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
6	CROSSLANG NV	CRSLNG	BE
7	European Federation of National Institutes for Language	EFNIL	LU
8	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR
9	European Civil Society Platform for Multilingualism	ECSPM	DK
10	CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium	CLARIN	NL
11	Universiteit Leiden (University of Leiden)	ULEI	NL
12	Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH)	ERSCM	DE
13	Stichting LIBER (Association of European Research Libraries)	LIBER	NL
14	Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.)	WMD	DE
15	Tilde SIA	TILDE	LV
16	Evaluations and Language Resources Distribution Agency	ELDA	FR
17	Expert System Iberia SL	EXPSYS	ES
18	HENSOLDT Analytics GmbH	HENS	AT
19	Xcelerator Machine Translations Ltd. (KantanMT)	KNTN	IE
20	PANGEANIC-B. I. Europa SLU	PAN	ES
21	Semantic Web Company GmbH	SWC	AT
22	SIRMA AI EAD (Ontotext)	ONTO	BG
23	SAP SE	SAP	DE
24	Universität Wien (University of Vienna)	UVIE	AT
25	Universiteit Antwerpen (University of Antwerp)	UANTW	BE
26	Institute for Bulgarian Language "Prof. Lyubomir Andreychin"	IBL	BG
27	Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences)	FFZG	HR
28	Københavns Universitet (University of Copenhagen)	UCPH	DK
29	Tartu Ulikool (University of Tartu)	UTART	EE
30	Helsingin Yliopisto (University of Helsinki)	UHEL	FI
31	Centre National de la Recherche Scientifique	CNRS	FR
32	Nyelvtudományi Kutatóközpont (Research Institute for Linguistics)	NYTK	HU
33	Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies)	SAM	IS
34	Fondazione Bruno Kessler	FBK	IT
35	Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia)	IMCS	LV
36	Lietuvių Kalbos Institutas (Institute of the Lithuanian Language)	LKI	LT
37	Luxembourg Institute of Science and Technology	LIST	LU
38	Università ta Malta (University of Malta)	UM	MT
39	Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute)	INT	NL
40	Språkrådet (Language Council of Norway)	LCNOR	NO
41	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polish Academy of Sciences)	IPIPAN	PL
42	Universidade de Lisboa, Faculdade de Ciências (University of Lisbon, Faculty of Science)	FCULisbon	PT
43	Institutul de Cercetări Pentru Inteligență Artificială (Romanian Academy)	ICIA	RO
44	University of Cyprus, French and European Studies	UCY	CY
45	Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied (Slovak Academy of Sciences)	JULS	SK
46	Institut Jožef Stefan (Jozef Stefan Institute)	JSI	SI
47	Centro Nacional de Supercomputación (Barcelona Supercomputing Center)	BSC	ES
48	Kungliga Tekniska högskolan (Royal Institute of Technology)	KTH	SE
49	Universität Zürich (University of Zurich)	UZH	CH
50	University of Sheffield	USFD	UK
51	Universidad de Vigo (University of Vigo)	UVIGO	ES
52	Bangor University	BNGR	UK

Contents

List of Figures

List of Tables

List of Acronyms

ABSA	Aspect-based Sentiment Analysis
AI	Artificial Intelligence
ALPAC	Automatic Language Processing Advisory
ASR	Automatic Speech Recognition
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
CLEF	Conference and Labs of the Evaluation Forum / Cross-Language Evaluation Forum
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
DNN	Deep Neural Networks
EHR	Electronic Health Records
EL	Entity Linking
ELE	European Language Equality (this project)
ELG	European Language Grid (EU project, 2019-2022)
ELRA	European Language Resource Association
ESFRI	European Strategy Forum on Research Infrastructures
GCN	Graph Convolution Networks
GMM	Gaussian Mixture Model
GPU	Graphical Processing Unit
HCI	Human-Computer Interactions
HLT	Human Language Technology
HMM	Hidden Markov Models
HPC	High Performance Computing
ICALL	Intelligent Computer-Assisted Language Learning
ICD	International Classification of Diseases
ICT	Information and Communications Technology
IE	Information Extraction
IR	Information Retrieval
LDC	Linguistic Data Consortium
LM	Language Model
LR	Language Resources/Resource
LT	Language Technology
MIMIC	Medical Information Mart for Intensive Care
ML	Machine Learning
MLLM	Multilingual Language Models
MMLM	Multilingual Masked Language Modeling

MMT	Multimodal Machine Translation
MNMT	Multilingual Neural Machine Translation
MT	Machine Translation
MUC	Message Understanding Conference
NED	Named Entity Disambiguation
NER	Named Entity Recognition
NMT	Neural Machine Translation
NLG	Natural Language Generation
NLI	Natural Language Inference
NLM	National Library of Medicine
NLP	Natural Language Processing
NLU	Natural Language Understanding
OIE	Open Information Extraction
POS	Part-of-Speech
RE	Relation Extraction
RNN	Recurrent Neural Network
SemEval	International Workshop on Semantic Evaluation
SMM4H	Social Media Mining for Health shared tasks
SNOMED-CT	Standarized Nomenclature of Medicine - Clinical Terms
SOTA	State-of-the-art
SR	Speaker Recognition
SRL	Semantic Role Labelling
TA	Text Analysis
TAC	Text Analysis Conference
TTS	Text to Speech Synthesis
UMLS	Unified Medical Language System
VLO	Virtual Language Observatory
VQA	Visual Question Answering
vSTS	Visual Semantic Textual Similarity
WSD	Word Sense Disambiguation

Abstract

D1.2 reports on the current state of the art in the field of Language Technology (LT) and language-centric Artificial Intelligence (AI). The main purpose of this deliverable is to landscape the field of LT and language-centric AI by assembling a comprehensive report of the state of the art of basic and applied research in the area. Given the multidisciplinarity of the field, this state of the art also reviews various scientific fields and areas involved (linguistics, computational linguistics, AI, computer science, etc.) and sketches all recent advances in AI, including the most recent deep learning neural technologies as well as the most advanced pretrained language models. In doing so, we map the relevant technologies onto a meaningful multidimensional structure that depicts the different areas involved, the methodologies and approaches applied, the modalities addressed (text, speech, sign), the communicative tasks, subtasks and application areas, including but not limited to Machine Translation, Speech Processing, Interactive and Dialogue Systems, Text Analytics, etc., domain sectors such as health, legal, media, education, tourism, etc. and the level of LT development. Special attention is paid to innovative solutions to less-resourced languages since LT is an important factor for language development in minority language communities. The final purpose of this exercise is to bring to light not only where Language-centric AI as a whole stands in 2020/2021, but also – and most importantly – where the required resources should be allocated to place European LT at the forefront of the AI revolution and in order to make real progress by 2030 instead of just small incremental improvements. We identify key research areas and gaps in research that need to be addressed to ensure LT can overcome the current LT inequality.

1 Introduction

Interest in the computational processing of human languages (machine translation, dialogue systems, etc.) coincided with the emergence of AI and, due to its increasing importance, the discipline has been established as specialized fields known as *Computational Linguistics* (CL), *Natural Language Processing* (NLP) or Language Technology. While there are differences in focus and orientation, since CL is more informed by linguistics 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 report we treat them interchangeably as long as it is not otherwise explicitly stated.

LT is concerned with studying and developing systems capable of processing human language. The field has developed, over the years, 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 difficulty of many of the tasks performed, current LT support allows many advanced applications which have been 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 summaries, inclusive technology, etc. Its rapid development in recent years predicts even more encouraging and also exciting results in the near future.

This report on the 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 four parts that follow this initial historical overview are frameworks, research areas, domain sectors and LT beyond language. They offer a survey that maps today's LT and language-centric AI landscape. Section 3 is devoted to existing LT frameworks and discusses processing pipelines and toolkits, language models, benchmarking and infrastructures. It highlights recent advances in these areas, including the shift to neural networks and components, a motif that runs throughout the report. Section 4 consists of seven sections devoted to LT resources, Text Analysis (TA), speech processing, Machine Translation (MT), Information Extraction (IE) and Information Retrieval (IR), Natural Language Generation (NLG) and summarization, and Human-Computer Interactions (HCI). Section 5 focuses on LT in large domain sectors including medicine, education, etc. Section 6 surveys recent developments in language-centered multimodal AI. Finally, some discussion and conclusions are outlined in sections 7 and 8 respectively.

2 Historical Overview

Today, many people use LT on a daily basis, especially online forms, often oblivious they are doing so. LT is an important but frequently invisible component of applications as diverse as, for example, search engines, spell-checkers, Machine Translation (MT) systems, recommender systems, virtual assistants, transcription tools, voice synthesizers and many others. This section presents a very brief historical view in section 2.1 and in section 2.2 the current revolution that is happening thanks to the new deep learning era.

2.1 A very brief historical view

The 1950s mark the beginning of LT 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 with his generative grammar laid the foundations to formalise, specify and automate linguistic rules (Chomsky, 1957). For a long time, the horizon defined by Turing and the instrument provided by Chomsky influenced the vast majority of NLP research.

The early years of LT were closely linked to MT, a well-defined task, and also relevant from a political and strategic point of view. In the 1950s it was believed that a quality automatic translator would be available soon. After several years of effort, in the mid-1960s the Automatic Language Processing Advisory Committee (ALPAC) report, issued by a panel of leading US experts acting in an advisory capacity to the US government, revealed the true difficulty of the task and, in general, of NLP (Pierce and Carroll, 1966). The ALPAC report had a devastating impact on R&D&I funding for the field. From then on, the NLP community turned towards more specific and realistic objectives. The 1970s and 1980s 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 and 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. Also noteworthy from this period was the effort to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc., of which WordNet (Miller, 1992) is one of the main results. Gradually, data-based systems have been displacing rule-based systems, and today it is difficult to conceive of an NLP system that does not have some component based on ML. In the 2010s we observed a radical technological change in NLP. Collobert et al. (2011) presented a multilayer neural network adjusted by backpropagation which was able to solve various sequential labeling problems. The success of this approach lies in the ability of these networks to learn continuous vector representations of the words (or word embeddings) using unlabelled data (for parameter initialisation) and using labelled data (for fine-tuning the parameters) to solve the task at hand. Word embeddings have played a very relevant role in recent years as they allow the incorporation of pretrained external *knowledge* in the neural architecture (Mikolov et al., 2013b; Pennington et al., 2014; Mikolov et al., 2018).

The availability of large volumes of unannotated texts together with the progress in self-supervised Machine Learning and the development of high-performance hardware (in the form of Graphical Processing Units, GPUs) enabled the development of very effective deep learning systems across a range of application areas.

2.2 The Deep Learning era

In recent years, the LT community has witnessed the emergence of powerful new deep learning techniques and tools that are revolutionizing the approach to LT tasks. We are gradually moving from a methodology in which a pipeline of multiple modules was the typical way to implement LT solutions, to architectures based on complex neural networks trained with vast amounts of text data. For instance, the *AI Index Report 2021*[1] highlights the rapid progress in NLP, vision and robotics thanks to deep learning and deep reinforcement learning techniques. In fact, the *Artificial Intelligence: A European Perspective* report[2] establishes that the success in these areas of AI has been possible because of the confluence of four different research trends: 1) mature deep neural network technology, 2) large amounts of data (and for NLP processing large and diverse multilingual textual data), 3) increase in High Performance Computing (HPC) power in the form of GPUs, and 4) application of simple but effective self-learning approaches (Goodfellow et al., 2016; Devlin et al., 2019; Liu et al., 2020b; Torfi et al., 2020; Wolf et al., 2020).

As a result, various IT enterprises have started deploying large pretrained neural language models in production. Google and Microsoft have integrated them in their search engines and companies such as OpenAI have also been developing very large language models. Compared to the previous state of the art, the results are so good that systems are claimed to obtain human-level performance in laboratory benchmarks when testing some difficult English language understanding tasks. However, those systems are not robust enough, very sensitive to phrasing and typos, perform inconsistently (when they are faced with similar input), etc. (Ribeiro et al., 2018, 2019). Additionally, existing laboratory benchmarks and datasets also have a number of inherent and severe problems (Caswell et al., 2021). For instance, the ten most cited AI datasets are riddled with label errors, which is likely to distort our understanding of the field's progress (Northcutt et al., 2021).

Forecasting the future of LT and language-centric AI is a challenge. Five years ago, few would have predicted the recent breakthroughs that have resulted in systems that can translate without parallel corpora (Artetxe et al., 2019), create image captions (Hossain et al., 2019), generate full text claimed to be almost indistinguishable from human prose (Brown et al., 2020), generate theatre play scripts (Rosa et al., 2020) and create pictures from textual descriptions (Ramesh et al., 2021).[3] It is, however, safe to predict that even more advances will be achieved by using pretrained language models. For instance, GPT-3 (Brown et al., 2020), one of the largest dense language models, can be fine-tuned for an excellent performance on specific, narrow tasks with very few examples. GPT-3 has 175 billion parameters and was trained on 570 gigabytes of text, with a cost estimated at more than four million USD.[4] In comparison, its predecessor, GPT-2, was over 100 times smaller, at 1.5 billion parameters.

1 https://aiindex.stanford.edu/report/
2 https://ec.europa.eu/jrc/en/publication/artificial-intelligence-european-perspective
3 https://openai.com/blog/dall-e/
4 https://lambdalabs.com/blog/demystifying-gpt-3/

This increase in scale leads to surprising behaviour: GPT-3 is able to perform tasks it was not explicitly trained on with zero to few training examples (referred to as zero-shot and few-shot learning, respectively). This behaviour was mostly absent in the much smaller GPT-2 model. Furthermore, for some tasks (but not all), GPT-3 outperforms state-of-the-art models explicitly trained to solve those tasks with far more training examples.

It is impressive that a single model can achieve a state-of-the-art or close to a state-of-the-art performance in limited training data regimes. Most models developed until now have been designed for a single task, and thus can be evaluated effectively by a single metric. Despite their impressive capabilities, large pretrained language models do come with some drawbacks. For example, they can generate racist, sexist, and otherwise biased text. Furthermore, they can generate unpredictable and factually inaccurate text or even recreate private information.[5] Combining large language models with symbolic approaches (knowledge bases, knowledge graphs), which are often used in large enterprises because they can be easily edited by human experts, is a non-trivial challenge. Techniques for controlling and steering such outputs to better align with human values are nascent but promising. These models are also very expensive to train, which means that only a limited number of organisations with abundant resources in terms of funding, computing capabilities, LT experts and data can currently afford to develop and deploy such models. A growing concern is that due to unequal access to computing power, only certain firms and elite universities have advantages in modern AI research (Ahmed and Wahed, 2020).

Moreover, computing large pretrained models also comes with a very large carbon footprint. Strubell et al. (2019) recently benchmarked model training and development costs in financial terms and estimated carbon dioxide emissions. While the average human is responsible for an estimated five tons of carbon dioxide per year,[6] the authors trained a big neural architecture and estimated that the training process emitted 284 tons of carbon dioxide. Finally, such language models have an unusually large number of uses, from chatbots to summarization, from computer code generation to search or translation. Future users are likely to discover more applications, and use positively (such as knowledge acquisition from electronic health records) and negatively (such as generating deep fakes), making it difficult to identify and forecast their impact on society. As argued by Bender et al. (2021), it is important to understand the limitations of large pretrained language models, which they call "stochastic parrots" and put their success in context.

3 Frameworks

Authors: Itziar Aldabe, German Rigau

This section presents various existing LT frameworks. More specifically, section 3.1 discusses processing pipelines and toolkits, including new types that have emerged over the last few years. Section 3.2 outlines the paradigm shift in LT, i.e. neural language models. The most important means to evaluate performance of NLP systems are presented in section 3.3 and section 3.4 outlines existing large infrastructures for LT.

3.1 Processing Pipelines and Toolkits

As previously mentioned, NLP has undergone a rapid transformation over the last few years. Architectures based on deep learning have enabled substantial progress for a variety of tasks such as question answering, Machine Translation or Automatic Speech Recognition (ASR).

[5] https://ai.googleblog.com/2020/12/privacy-considerations-in-large.html
[6] https://ourworldindata.org/co2-emissions

In addition, these improvements have been accompanied by libraries that allow end-to-end processing and the integration of NLP tools in higher-level applications.

This fast-growing collection of efficient tools has led to the emergence of new types of processing pipelines and toolkits. Traditionally, when one calls an NLP pipeline the text is first tokenized and then processed in different steps, forming the processing pipeline. This pipeline often includes a tagger, a lemmatizer, a parser, a named entity recognizer etc. Each pipeline module or component returns the processed text and this is then passed to the next module. Thus, the pipeline takes in raw text as input and produces a set of annotations. Well-known examples of this type of multilingual pipeline are: CoreNLP,[7] Freeling,[8] ixa-pipes,[9] GATE,[10] DKPro,[11] Apache UIMA,[12] Stanza,[13] Trankit, [14] and Spark NLP.[15]

Today, it is becoming more common to find libraries that are built with neural network components and pretrained models that also cover multilingual NLP tasks. SpaCy[16] supports more than 60 languages and offers 55 trained pipelines for 17 languages. In its capacity as a production-ready training system it is focused on state-of-the-art speed. UDify[17] (Kondratyuk and Straka, 2019) is a single model that parses Universal Dependencies (UPOS, UFeats, Lemmas, Deps) accepting any of 75 supported languages as input. Flair[18] (Akbik et al., 2019) was designed to work with different types of word embeddings, as well as training and distributing sequence labeling and text classification models. UDPipe (Straka, 2018),[19] which utilizes a neural network with a single joint model for POS tagging, lemmatization and dependency parsing, is trained using only CoNLL-U training data and pretrained word embeddings. Stanza[20] (Qi et al., 2020) features a language-agnostic fully neural pipeline for Text Analysis, including tokenization, multi-word token expansion, lemmatization, part-of-speech and morphological feature tagging, dependency parsing and named entity recognition. Spark NLP[21] (Kocaman and Talby, 2021) is a state-of-the-art Natural Language Processing library built on top of Apache Spark. It provides simple, performant and accurate NLP annotations for machine learning pipelines that scale easily in a distributed environment. Spark NLP comes with 3700+ pretrained pipelines and models in more than 200+ languages. Finally, Trankit[22] (Nguyen et al., 2021) is a multilingual Transformer-based toolkit that supports 56 languages with 90 pretrained pipelines on 90 treebanks of the Universal Dependency v2.5. Several transformer-based models for many languages may be simultaneously loaded into GPU memory to process the raw text inputs of different languages.

3.2 Neural Language Models

LT is undergoing a paradigm shift with the rise of *neural language models*[23] that are trained on broad data at scale and are adaptable to a wide range of monolingual and multilingual

7 https://stanfordnlp.github.io/CoreNLP/
8 http://nlp.lsi.upc.edu/freeling/
9 https://ixa2.si.ehu.eus/ixa-pipes/
10 https://gate.ac.uk/
11 https://dkpro.github.io/
12 https://uima.apache.org/
13 https://stanfordnlp.github.io/stanza/
14 https://github.com/nlp-uoregon/trankit
15 https://github.com/JohnSnowLabs/spark-nlp
16 https://spacy.io
17 https://github.com/Hyperparticle/udify
18 https://github.com/flairNLP/flair
19 https://ufal.mff.cuni.cz/udpipe/2
20 https://stanfordnlp.github.io/stanza/
21 https://github.com/JohnSnowLabs/spark-nlp
22 https://github.com/nlp-uoregon/trankit
23 Also known as Pretrained Language Models (Han et al., 2021)

downstream tasks (Devlin et al., 2019; Qiu et al., 2020; Liu et al., 2020b; Torfi et al., 2020; Wolf et al., 2020; Han et al., 2021; Xue et al., 2021). Though these models are based on standard *self-supervised* deep learning and *transfer learning*, their scale results in new emergent and surprising capabilities, but their effectiveness across so many tasks demands caution, as their defects are inherited by all the adapted models downstream. Moreover, we currently have no clear understanding of how they work, when they fail, and what emergent properties they present. To tackle these questions, much critical interdisciplinary collaboration and research is needed. Thus, some authors call these models *foundation models* to underscore their critically central yet incomplete character (Bommasani et al., 2021).

Most LT systems today are powered by ML where predictive models are trained on known data and used to make predictions on new data. The rise of machine learning within AI and LT started in the 1990s where rather than specifying *how* to solve a task, a learning algorithm induced a model based on a set of *features* representing in the best possible way the training data examples. Thus, complex NLP tasks still require a manually-driven *feature engineering* process to characterise raw data into task useful representations. Around ten years ago, *Deep Learning* (Salakhutdinov, 2014) started gaining traction in LT thanks to mature deep neural network technology, much larger datasets, more computational capacity (notably, the availability of GPUs), and application of simple but effective self-learning objectives (Goodfellow et al., 2016). One of the advantages of these neural language models is their ability to alleviate the *feature engineering* problem by using low-dimensional and dense vectors (aka. *distributed representation*) to implicitly represent the language examples (Collobert et al., 2011). By the end of 2018,[24] the field of NLP observed another relevant disruption with BERT (Devlin et al., 2019). Since then BERT has become a ubiquitous baseline in NLP experiments and inspired a large number of studies and improvements (Rogers et al., 2020).

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* (Turian et al., 2010; Mikolov et al., 2013a; Pennington et al., 2014; Mikolov et al., 2018) associated word vectors with context-independent vectors. Shortly thereafter, self-supervised learning based on autoregressive language modelling (predict the next word given the previous words) (Dai and Le, 2015) became popular. This approach produced language models such as GPT (Radford et al., 2018), ELMo (Peters et al., 2018) and ULMFiT (Howard and Ruder, 2018). The next wave of developments in self-supervised learning — BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2020), BART (Lewis et al., 2020) — quickly followed, embracing the Transformer architecture (Vaswani et al., 2017), incorporating more powerful deep bidirectional encoders of sentences, and scaling up to larger models and datasets. Figure 1 presents the relationship of some of these pre-trained language models in a diagram.[25] For example, BERT (Devlin et al., 2019) applies two training self-supervised tasks namely *Masked Language Model* and *Next Sentence Prediction*. The *Masked Language Model* learns to predict a missing word in a sentence given its surrounding context while the *Next Sentence Prediction* learns to predict if the next sentence will follow the current one or not. Self-supervised tasks are not only more scalable, just depending on unlabelled data, but they are designed to force the model to predict coherent parts of the input. Through self-supervised learning, tremendous amounts of unlabeled textual data can be utilised to capture versatile linguistic knowledge without labour-intensive workloads. This pretrained language model recipe has been replicated across languages leading to many language specific BERTs such as FlauBERT and CamemBERT for French (Le et al., 2020; Martin et al., 2020), Robbert for Dutch (Delobelle et al., 2020), BERTeus for Basque (Agerri et al., 2020), etc.

The idea of *transfer learning* is to take the "knowledge" learned from one task (e.g., pre-

[24] The paper first appeared in http://arxiv.org.
[25] https://github.com/thunlp/PLMpapers

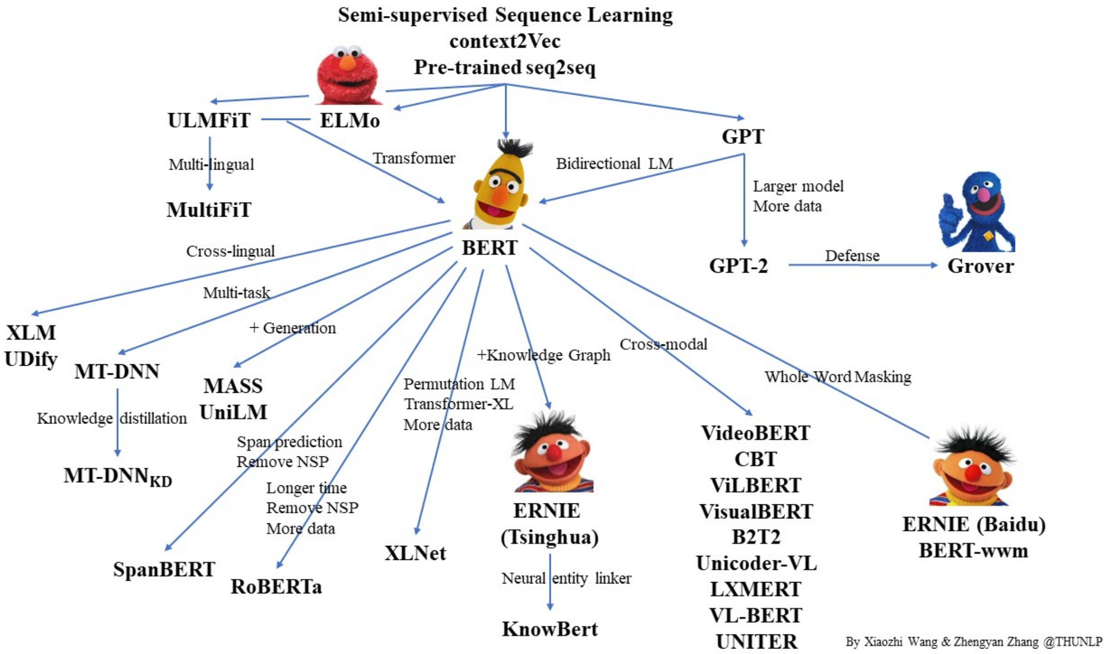

Figure 1: Relationship between some pre-trained language models.

dict the next word given the previous words) and apply it to another task (e.g., summarization). With transfer learning, instead of starting the learning process from scratch, you start from patterns that have been learned when solving a different problem. This way you leverage previous learning and avoid starting from scratch. Within deep learning, pretraining is the dominant approach to *transfer learning*: the objective is to *pretrain* a deep transformer model on large amounts of data and then reuse this pretrained language model by *fine-tuning* it on small amounts of (usually annotated) task-specific data. Thus, transfer learning formalises a two-phase learning framework: a pretraining phase to capture knowledge from one or more source tasks, and a fine-tuning stage to transfer the captured knowledge to many target tasks. Recent work has shown that pretrained language models can robustly perform classification tasks in a few-shot or even in 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* is based on exploiting pretrained language models to solve a task using text directly (Liu et al., 2021b). To use these models to perform prediction tasks, the original input is modified using a template into a textual string prompt that has some missing slots, and then the language model is used to probabilistically fill the missing information to obtain a final string, from which the final output for the task can be derived. This framework looks very promising for a number of reasons: it allows the language model to be pretrained on massive amounts of raw text, and by defining a new prompting function the model is able to perform few-shot or even zero-shot learning, adapting to new scenarios with few or no labeled data. Thus, some NLP tasks can be solved in a fully unsupervised fashion by providing a pretrained language model with “task descriptions” in natural language (Raffel et al., 2020; Schick and Schütze, 2021a). Surprisingly, fine-tuning pretrained 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 Language Models (MLLMs) such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mT5 (Xue et al., 2021), mBART (Liu et al., 2020b), etc. have emerged as a viable option for bringing the power of pretraining to a large number of languages. For example, mBERT (Devlin et al., 2019) is pretrained with the Multilingual Masked Language Modeling (MMLM) task using non-parallel multilingual Wikipedia corpora in 104 languages. mBERT has the ability to generalize 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. An MLLM is pretrained using large amounts of unlabeled data from multiple languages with the hope that low-resource languages may benefit from high-resource languages due to a shared vocabulary and latent language properties. The surprisingly good performance of MLLMs in crosslingual transfer as well as bilingual tasks motivates the hypothesis that MLLMs 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. Thus, of particular interest is the ability of MLLMs to facilitate zero-shot crosslingual transfer from a resource-rich language to a resource-deprived language which does not have any task-specific training data, or to finetune more robust language models by using annotated training data in multiple languages.

In summary, recent progress in LT has been driven by advances in both model architecture and model pretraining. Transformer architectures have facilitated the building of higher-capacity models and pretraining has made it possible to effectively utilise this capacity for a wide variety of tasks. Open-source libraries such as Transformers[26] may open up these advances to a wider LT community. The library consists of carefully engineered state-of-the art Transformer architectures under a unified API and a curated collection of pretrained models (Wolf et al., 2020). Unfortunately, the resources necessary to create the best-performing neural language models are found almost exclusively at US and China technology giants. Moreover, this transformative technology poses problems from a research advancement, environmental, and ethical perspective. For example, models such as GPT-3 are private, anglo-centric, and inaccessible to academic organisations (Floridi and Chiriatti, 2020; Dale, 2021). This situation also promotes a colossal duplication of energy requirements and environmental costs, due to the duplicated training of private models. Finally, there are worrying shortcomings in the text corpora used to train these models, ranging from a lack of representation of populations, to a predominance of harmful stereotypes, and to the inclusion of personal information.

3.3 Benchmarking NLP

As important as it is to develop new rule-based, machine-based or deep learning systems to solve different NLP tasks, it is equally essential to measure the performance of these systems. The most common method to do so is through the use of benchmarks, i.e., according to manually annotated datasets. Well-known examples include datasets for Text Classification (Minaee et al., 2021), Language Modeling (Merity et al., 2017), Image Captioning (Chen et al., 2015a), Machine Translation (Callison-Burch et al., 2009a), Question Answering (Rajpurkar et al., 2016), Automatic Speech Recognition (Panayotov et al., 2015), Document Summarization (Nallapati et al., 2016) and Natural Language Inference (NLI) (Bowman et al., 2015), etc. Leaderboards such as NLP-progress[27], Allen Institute of AI leaderboard,[28] Papers with code,[29] or Kaggle[30] are meant to encourage participation and facilitate evaluation across many different NLP tasks and datasets.

Although measuring performance in this way currently represents the primary means to assess progress in various NLP tasks and models, performance-based evaluation on a shared

[26] https://huggingface.co/
[27] http://nlpprogress.com/
[28] https://leaderboard.allenai.org/
[29] https://paperswithcode.com/area/natural-language-processing
[30] https://www.kaggle.com/datasets?tags=13204-NLP

task is a paradigm that has existed since the Message Understanding Conferences (MUC) in the late 1980s (Hirschman, 1998). For example, the International Workshop on Semantic Evaluation (SemEval) is an ongoing series of evaluations that started in 2007 after running three SensEval evaluation exercises for word sense disambiguation organised under the umbrella of SIGLEX. The 15th edition of SemEval featured tasks ranging from prediction of lexical complexity to learning with disagreements and included several cross-lingual and multimodal tasks.[31] The Text Analysis Conference (TAC), launched in 2008, also hosts a series of evaluation workshops in which they provide large test collections to pursue common evaluation procedures. The TAC 2020 included evaluations in Epidemic Question Answering, Recognizing Ultra Fine-Grained Entities and Streaming Multimedia Knowledge Base Population.[32] Similarly, the CLEF Initiative (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum) is a self-organised body whose main mission is to promote research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information with various levels of structure.[33]

For instance, according to the SQuAD leaderboard (Rajpurkar et al., 2016), on January 3, 2018, Microsoft Research Asia submitted an end-to-end deep learning model that reached an EM score of 82.65 on the machine reading comprehension dataset made up of factoid questions about Wikipedia articles. While this score was better than human performance on the same set, it should not be taken to mean that machines read and comprehend documents as humans do. Today, SQuAD2.0, which aims to not only answer questions, but also to not answer them when they cannot be answered, presents an 89.452 F1-score of human performance, while the best submitted system reaches a 93.214 F1-score. More challenging benchmarks such as GLUE (Wang et al., 2019b), SuperGLUE (Wang et al., 2019a), SentEval (Conneau and Kiela, 2018) and DecaNLP (McCann et al., 2018) have been proposed to measure performance on multi-tasks, Natural Language Understanding (NLU) datasets such as Adversarial Natural Language Inference (NLI) (Nie et al., 2020) to address benchmark longevity and robustness, or even platforms for dynamic data collection and benchmarking to evaluate progress in NLP (Kiela et al., 2021).

Given their importance and success in advancing NLP models, it is unsurprising that benchmarks are wide-ranging and of multiple purpose. One recent survey presented an overview of close to fifty widely used benchmarks for NLI alone (Storks et al., 2019b). However, as they multiply and become ever more sophisticated in parallel to methodological advances and the onrush of data, benchmark tasks are expected to demand deeper understanding from models in addition to greater performance and accuracy. For although benchmarks demonstrate that performance is indeed rising, several areas for improving their capabilities have been identified. These include, for instance, the need for better evaluation of resources that include external knowledge or novel ways to automatically integrate a broad spectrum of reasoning. But surveys of benchmarking indicate that more attention must also be paid to “other qualities that the NLP community values in models, such as compactness, fairness, and energy efficiency” (Ethayarajh and Jurafsky, 2020). By way of example, many SOTA models do not perform well in terms of racial and gender biases. Part of the problem is that most leaderboards are not designed to measure performance with respect to bias. The same applies to other highly relevant factors such as carbon footprint (Henderson et al., 2020). Moreover, most of these evaluation datasets and benchmarks have been developed for English only. For instance, the Papers with code platform includes 1044 English datasets but only 142 for Chinese which appears in second position.[34] Also interesting are those few evaluation benchmarks that have been designed for low-resource scenarios (Goyal et al., 2021).

---

[31] https://semeval.github.io/
[32] https://tac.nist.gov/
[33] http://www.clef-initiative.eu/
[34] Last accessed in September 2021

3.4 Large Infrastructures for Language Technology

Regarding LT, the ESFRI Landmark CLARIN ERIC (Common Language Resources and Technology Infrastructure) offers inter operable access to language resources and technologies for researchers in the humanities and social sciences.[35] Unfortunately, not all EU Member States are official members of CLARIN (i.e., Belgium, Ireland, Luxembourg, Malta, Slovakia and Spain are not CLARIN members) and some of them just participate in CLARIN as observers (i.e., France). Moreover, as the research funding agencies are providing unbalanced resources to the different member states, the European languages are not equally supported by CLARIN (de Jong et al., 2020).

The European LT community has been demanding a dedicated LT platform for years. The European Language Grid (ELG)[36] (Rehm et al., 2020, 2021) with representatives from all European languages is targeted to evolve into the primary platform and marketplace for LT in Europe by providing one umbrella platform for the European LT landscape, including research and industry, enabling all stakeholders to upload, share and distribute their services, products and resources. ELG plans to provide access to approx. 1300 services for all European languages as well as thousands of datasets. ELG plans to establish a legal entity in 2022 with these assets.

Under the auspices of the successful Hugging Face platform (Wolf et al., 2019), the BigScience project took inspiration from scientific creation schemes such as CERN and the LHC, in which open scientific collaborations facilitate the creation of large-scale artefacts that are useful for the entire research community.[37] Hugging Face, at the origin of the project, develops open-source research tools that are widely used in the NLP language modeling community. The project also brings together more than thirty partners, in practice involving more than a hundred people, from academic laboratories, startups/SMEs, and large industrial groups and is now extending to a much wider international community of research laboratories.

In conclusion, 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 neural network components are increasingly common, including pretrained models that perform multilingual NLP tasks. In like manner, 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 surmounted. One problem is the question of the resources needed to design the best-performing neural language models, currently housed almost exclusively at the large US and Chinese technology companies. Another issue is the problem of stereotypes, prejudices and personal information within the text corpora used to train the models. The latter, as pointed out above, is an issue that also concerns benchmarks and leaderboards, the challenge of freeing both these and neural language models from such biases is daunting, but the problem of the predominance towards 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, including English. This capability would, in turn, enable Europe's languages to enjoy full and equal access to digital language technology.

[35] http://www.clarin.eu
[36] https://www.european-language-grid.eu
[37] https://bigscience.huggingface.co/

4 Research areas

Section 4 presents some of the most important research areas of LT. First of all, section 4.1 details the data, tools and services that are available. Then, section 4.2 highlights the main research topics for analysing textual data together with the use of newer approaches. Speech processing is approached in section 4.3, in which three central areas within the field (text to speech synthesis, automatic speech recognition and speaker recognition) are addressed. Section 4.4 concentrates on machine translation. Neural machine translation is compared to earlier statistical approaches and to more recent multilingual neural machine translation systems. The report turns to information extraction and information retrieval in section 4.5. Section 4.6 concentrates on language generation and summarisation, one of the most important yet challenging tasks in NLP. Finally, human-computer interaction is the focus of section 4.7 and it is focused on dialogue systems, and it is separated into conversational agents, interactive question answering systems and task-oriented systems.

4.1 LT Resources

Authors: Ainara Estarrona

The term ”Language Resource” (LR) refers to a set of speech or language data and descriptions in machine readable form. These are used for building, improving or evaluating natural language and speech algorithms or systems, or, as core resources for the software localisation and language services industries, for language studies, electronic publishing, international transactions, subject-area specialists and end users.

No widely used standard typology of language resources has been established. However, a general classification could be as follows:

- Data
  - Corpora (digital collections of natural language data)
  - Lexical/conceptual resources (machine-readable dictionaries, lexicons, ontologies)
- Tools/Services
  - Linguistic annotations
  - Tools for creating annotations
  - Search and retrieval applications (corpus management systems)
  - Applications for automatic annotation (part-of-speech tagging, syntactic parsing, semantic parsing, audio segmentation, speaker diarization)
- Metadata and vocabularies
  - Vocabularies or repositories of linguistic terminology
  - Language metadata

In this report we will focus on the first two categories: data and tools/services. A main objective of the language resource community is the development of infrastructures and platforms for presenting, discussing and disseminating language resources. There are numerous catalogues and repositories where the different resources for each language can be documented. Among the major catalogues at the European level are the following:

- ELRC-SHARE[38]

[38] http://www.elrc-share.eu)/

- European Language Grid (ELG)[39]
- European Language Resources Association (ELRA)[40]
- Common Language Resources and Technology Infrastructure (CLARIN)[41]
- META-SHARE[42]

The **European Language Resource Coordination** promotes **ELRC-SHARE**, a Language Resources repository used for documenting, storing, browsing and accessing language data and tools that are pertinent to MT and considered useful for feeding CEF eTranslation, the European Commission's Automated Translation platform. It currently hosts more than 2000 LRs, mainly bi- and multi-lingual corpora and terminological resources.

The **European Language Grid** (ELG) aspires to be Europe's leading language technology platform. It focuses on European languages and will eventually include all official languages of the European Union as well as other non-official or regional languages. The end result will be an online catalogue of LT and resources that can be browsed, searched and explored. Users will be able to filter and search by domains, regions, countries, languages, types of tools or services, datasets and much more. ELG will include more than 800 functional LT services and more than 3500 LT datasets, corpora, resources and models.

Among the **European Language Resources Association** (ELRA) missions are the promotion of language resources for the Human Language Technology (HLT) sector and the evaluation of language engineering technologies. Its main objectives within this context are to provide Language Resources through its repository, save researchers and developers the effort of rebuilding resources that already exist, and help them identify and access these resources. The ELRA catalogue contains resources for any language and does not differentiate between European and non-European languages. At the moment it has more than 1,300 resources that can be browsed using a search engine with different search criteria such as language, type of resource, licence, media type, etc.

Common Language Resources and Technology Infrastructure (CLARIN) is a digital infrastructure offering data, tools and services to support research based on language resources. CLARIN language resources are divided into data and tools. A distributed infrastructure provides access to digital language data that covers various dimensions (language, modality, time span, etc.). The CLARIN language resources can be accessed via the individual repositories or their unified catalogue, the Virtual Language Observatory (VLO).[43] The VLO provides a means to explore language resources and tools. Its easy to use interface allows for a uniform search and discovery process for a large number of resources from a wide variety of domains. A powerful query syntax makes it possible to carry out more targeted searches as well. As far as tools are concerned, the CLARIN centres offer a multitude of applications to discover, explore, exploit, annotate, analyse or combine language data. The Language Resource Switchboard[44] can assist with finding the right language processing tool for a researcher's data. If you upload a file, or enter a URL, the Switchboard provides a step-by-step guidance on how to process the data with a CLARIN tool.

Finally, it is worth mentioning the **META-SHARE** repository. META-SHARE is an open, integrated, secure and inter operable sharing and exchange facility for LRs (datasets and tools) for the Human Language Technologies domain and other relevant domains where language plays a critical role. This repository contains more than 2,800 LRs from all over the world that can be consulted through a web search engine.

39 https://live.european-language-grid.eu/
40 http://catalogue.elra.info/en-us/
41 https://www.clarin.eu/content/language-resources
42 http://metashare.ilsp.gr:8080/repository/search/
43 https://vlo.clarin.eu
44 https://switchboard.clarin.eu/

Outside European borders, the **Linguistic Data Consortium** should be highlighted. [45] LDC's primary role is as a repository and distribution point for language resources. With the help of its members, LDC has grown into an organisation that creates and distributes a wide array of language resources. LDC also supports sponsored research programs and language-based technology evaluations by providing resources and contributing ro organisational expertise. Data contained in the catalogue may be consulted through a web search engine[46] and different tools developed to support evolving annotation tasks are also available.[47]

In addition to these repositories, some relevant multilingual public domain initiatives also exist. To highlight a few, firstly the Common Voice Project, [48] specifically designed to encourage the development of ASR systems; the M-AILABS Speech Dataset, [49] for text to speech synthesis; the Ryerson Audio-Visual Database of Emotional Speech and Song, [50] to promote research on emotional multimedia content (available only for English); and LibriVox,[51] which is an audiobook repository that can be used in different research fields or applications.

A cursory glance at these catalogues and repositories not only gives us an idea of the amount of resources available for European languages, but also reveals the clear inequality between official and minority languages. Moreover, although the five European languages with the most resources are English, French, German, Spanish and Italian, English is by far ahead of the rest, with more than twice as many resources as the next language on the list. If we look at the ELG catalogue, for example, English has 2,372 resources, while the second language with the most resources, German, has only 784 resources. These five languages are followed by official languages with the fewest number of speakers: Bulgarian, Croatian, Czech, Danish, Dutch, Estonian, Finnish, Hungarian, Swedish, Portuguese, Polish, etc. It is worth mentioning the case of Estonian which, with around 1 million speakers, is in very good health in terms of resources. Languages of the European Union that do not have official status are far behind in terms of resource development: Norwegian, Basque, Catalan, Icelandic, Bosnian, Breton, Macedonian, etc. It is clear therefore that official status has an impact on the extent of available resources.

4.2 Text Analysis

Authors: Rodrigo Agerri

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 generate structured data out of free text content by identifying facts, relationships and entities that are buried in the textual data. TA employs a variety of methodologies to process text, one of the most important being NLP and, more specifically, Information Extraction.

The correct interpretation of a written text consists of correctly labeling actions or events and their participants, as well as capturing the relations that connect them. In order to achieve this, various types of analyses must be performed both at sentence and document level. This process should result not only in representing the explicit information denoted by the text, but also in discovering its implicit information. Moreover, in our increasingly multilingual world this information should be processed in multiple languages to allow for

[45] https://catalog.ldc.upenn.edu/search
[46] https://catalog.ldc.upenn.edu/
[47] https://www.ldc.upenn.edu/language-resources/tools
[48] https://commonvoice.mozilla.org/
[49] https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/
[50] https://zenodo.org/record/1188976
[51] https://librivox.org/

a cross-lingual and inter operable semantic interpretation. Ideally, this processing is robust enough to provide the same accurate results in multiple application domains and textual genres.

The best results for TA tasks are generally obtained by means of supervised, corpus-based approaches. This means that manually annotated data is used to train probabilistic models. This poses a major obstacle to train supervised models whenever there is not enough manually annotated data by linguists for a semantic task in a given language. In most cases, manually annotating text for every single specific need is generally extremely time-consuming and, in most cases, not affordable in terms of human resources and economic costs.

Even when manually annotated resources are available, a usual problem that researchers face is that texts need to be accurately analysed at many distinct levels for a full understanding. Furthermore, each of these levels are affected by ambiguous expressions that cannot be interpreted in isolation.

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 (see Section 3.1) include Part-of-Speech (POS) tagging, Lemmatization, 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, Aspect-based Sentiment Analysis (ABSA) and, more recently, Open Information Extraction (OIE).

The correct interpretation of a given text requires capturing the meaning of each word according to their context. WSD (Agirre and Edmonds, 2006) refers to the task of matching each word with its corresponding word sense in a lexical knowledge base, like WordNet (Fellbaum and Miller, 1998). This semantic analysis can be performed on any type of word, such as nouns, verbs or adjectives, as well as on named entities. For common words, POS tagging (disambiguating the morphosyntactic categories of words) is a first step that is usually performed before doing many of the other tasks mentioned above. Although this task is considered to be practically solved with current neural language models (Akbik et al., 2019; Devlin et al., 2019), POS tagger accuracy still degrades significantly when applied out of domain (Manning, 2011). Closely related to POS tagging is lemmatization (obtaining the canonical word or lemma from a given word form), because it has traditionally been considered that POS tagging (or fine-grained morphological information) is crucial in order to develop lemmatizers.

If we consider proper names, the NER (Tjong Kim Sang, 2002) task focuses on labeling entities with general semantic categories like person, organisation or place. However, the semantic interpretation of a sentence does not only depend on the meaning of the words. SRL (Carreras and Màrquez, 2004) tries to discover the predicates and their semantic roles in a sentence. In other words, who did what, when and where in a sentence. Like WSD, its aim is to label each element of the sentence with knowledge taken from a semantic source, such as FrameNet (Ruppenhofer et al., 2006), PropBank (Kingsbury and Palmer, 2002) or NomBank (Gerber and Chai, 2010), that describes predicate structures including roles as Agent, Patient or Location. Another more recent approach attempts to identify such semantic structures without depending on a particular semantic knowledge base, a task which is known as Open Information Extraction (OIE) (Stanovsky and Dagan, 2016).

For a text analysis system to be able to recognise, classify and link every mention of a specific named entity in a document, several tasks are considered, namely, NER, NED and Coreference Resolution. A named entity can appear in a great variety of surface forms. For instance, Barack Obama, President Obama, Mr. Obama, etc. could refer to the same person. Moreover, the same surface form can reference a variety of named entities. Therefore, to provide an adequate and comprehensive account of named entities in a text, a system must recognise a named entity, classify it as a type (e.g, person, location, organization, etc.), and recognise every form of the same entity even in multiple languages (Ratinov and Roth, 2009;

Turian et al., 2010; Agerri and Rigau, 2016; Lee et al., 2017; Akbik et al., 2019; Joshi et al., 2019; Cao et al., 2021).

SRL involves the recognition of semantic arguments of predicates. Conventional semantic roles include Agent, Patient, Instrument or Location. Many lexical databases currently contain complete descriptions of the predicate structure inclusive of its semantic roles and annotations in corpora (see, for example, FrameNet, PropBank, Predicate Matrix (Lopez de Lacalle et al., 2016), etc.). More recently, research is also focusing on Implicit SRL (ISRL), where the hope is to recover semantic roles beyond the syntactically close context of the predicates. Indeed, Gerber and Chai (2010) pointed out that solving implicit arguments can increase the coverage of role structures by 71%. Traditionally, tasks such as SRL or Coreference Resolution (Pradhan et al., 2012) required intermediate linguistic annotations provided by constituent (Collins, 2003) or dependency parsing (Straka, 2018), POS tagging and NER, among others.

Once the main events are identified, Temporal Processing aims to capture and structure Temporal Information. This consists of 1) identifying and normalising any temporal expression and event in the text and 2) establishing the temporal order in which the events occurred, as defined by the TempEval3 shared evaluation task (UzZaman et al., 2013).

To summarise, Text Analysis is crucial for establishing ”who did what, where and when”, a technology that has proved to be key for applications such as Information Extraction, Question Answering, Summarisation and nearly every linguistic processing task involving any level of semantic interpretation. Once the relevant information has been extracted, events can be annotated via Opinion Mining and ABSA, with the opinions and expressed polarity (positivity or negativity) referring to each event and its participants (Vossen et al., 2016). ABSA seeks to identify opinionated text content as well as obtain the sentiments (positive, neutral, negative) of the opinions, the opinion holders and targets (e.g. the particular aspect/feature of a product/event being evaluated) (Agerri et al., 2013; Pontiki et al., 2014).

Note that of all the Text Analysis tasks mentioned in this section, only POS tagging and, to a certain degree, NER, did not require intermediate linguistic information. Every other task usually depends on at least POS tags, constituent or dependency trees, NER and NED to obtain competitive systems. This is reflected in the traditional Text Analysis pipelines mentioned in Section 3.1.

Today, all these tasks are addressed in an end-to-end manner. This means that, 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 usually employ Recurrent Neural Network (LSTMs) and static word embeddings, such as Word2vec (Mikolov et al., 2013b), or on newer large pretrained Transformer language models such as BERT (Lee et al., 2017; Devlin et al., 2019; Joshi et al., 2019). Similarly, most current state-of-the-art Text Analysis toolkits including AllenNLP and Trankit, among others (Gardner et al., 2018; Nguyen et al., 2021) use a highly multilingual end-to-end approach. Avoiding intermediate tasks has aided in mitigating the common cascading errors problem that was pervasive in more traditional TA pipelines. As a consequence, the appearance of end-to-end systems has helped bring about a significant jump in performance across every TA task.[52]

4.3 Speech Processing

Authors: Inma Hernaez, Eva Navas, Jon Sanchez, Ibon Saratxaga

Speech processing aims at allowing humans to communicate with electronic devices through voice. This entails developing machines that understand and generate not only oral mes-

[52] https://nlpprogress.com/

sages, but also all the additional information that we can extract from the voice, like who is speaking, their age, their personality, their mood, their satisfaction with a service, 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 (Zen et al., 2009) generated 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) (Black et al., 2007), and more recently DNN (Deep Neural Networks) (Ze et al., 2013), have been used as statistical frameworks. Various network architectures have been tested, such as feed-forward networks (Qian et al., 2014), recurrent networks (Fan et al., 2014) and WaveNet (Oord et al., 2016). Among the criteria used for training, the most common is minimum generation error (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. A good review on possible strategies to utilise DNNs for the generation of speech acoustic parameters may be found in Ling et al. (2015).

Lately, the most favoured approach to speech systems is to substitute the whole chain in the TTS systems by DNNs (Ning et al., 2019). Deep Voice (Arik et al., 2017) was the first system where all the steps in the TTS system were implemented by means of DNNs. The quality of the generated voices was inferior to that obtained with WaveNet, so several improvements were proposed, such as Deep Voice 2 (Gibiansky et al., 2017) and 3 (Ping et al., 2018), where WaveNet could be used as a neural vocoder to analyse and synthesise the acoustic signal. Another approach that can be considered more end-to-end is Char2Wav (Sotelo et al., 2017), although it still concatenates two modules: the first predicts acoustic parameters from text and the second, a neural vocoder, generates a waveform from these parameters. Full end-to-end architectures have also been proposed, including Tacotron (Wang et al., 2017c), Tacotron2 (Shen et al., 2018), FastSpeech (Ren et al., 2019), FastSpeech 2 (Ren et al., 2020) and ClariNet (Ping et al., 2019). These systems are able to produce spectrograms from text, which are then converted to speech using the Griffin-Lim algorithm (Griffin and Lim, 1984), WaveNet or other neural vocoders such as WaveGlow (Prenger et al., 2019) and MelGAN (Kumar et al., 2019). The systems provide outstanding results in terms of the quality of the generated voices, but require large amounts of high quality recordings to be trained properly. Currently, efforts are being made to deploy these systems for low-resource languages by improving data efficiency (Chung et al., 2019), applying transfer learning (Chen et al., 2019b) or training multilingual models (Zhang et al., 2019c).

ASR the ability to produce a textual transcription from a computer's speech signal, has been long sought after in the speech processing field. The intrinsic difficulty of the task has required a step-by-step effort, with increasingly ambitious objectives: from discrete word, speaker dependent and reduced vocabulary systems to continuous speaker-independent recognition. Only in the last two decades has this technology jumped from the laboratory to industry. The first of these commercial systems were based on statistical models, namely the HMMs[53] In these systems, the speech signal is considered a short term stationary signal, and in this scale each stationary part is modelled by a hidden state of the Markov chain. Usually, each hidden state models a spectral representation of the sound wave by means of a Gaussian Mixture Model (GMM). Additionally, the actual language of the recogniser also has to be modelled and n-grams (Markowitz, 1995) are the usual choice among the statistical language models.

This technology was the standard during the first decade of the century. But in the 2010s,

[53] See Juang and Rabiner (2005) for a brief review and (Gales and Young, 2008) for a comprehensive description of this technology.

the increase of computing power and the ever-growing availability of training data allowed for the introduction of DNN techniques for ASR. The first attempts to adopt neural networks consisted in extracting parameters from speech using discriminatory trained feed-forward neural networks, also known as the tandem approach (Morgan, 2011). Other methods built on the traditional HMM-GMM architecture, replacing the GMMs by DNNs for acoustic modelling (Hinton et al., 2012). Open source tools like Kaldi (Povey et al., 2011) boosted the research and development of large vocabulary ASR systems.

More recently, end-to-end or fully differentiable architectures have appeared that aim to simplify a training process that is capable of exploiting the increasing available data. In these systems, a DNN maps the acoustic signal at the input directly to the textual output. Thus, the neural network models the acoustic information, the time evolution and some linguistic information, training everything jointly (Graves and Jaitly, 2014; Chorowski et al., 2015; Chan et al., 2016; Lu et al., 2016). Currently there are two general approaches for these systems. The first is the Connectionist Temporal Classification (CTC) (Graves and Jaitly, 2014; Miao et al., 2015; Chiu et al., 2018), which applies a Recurrent Neural Network (RNN) and a CTC output layer. The CTC is an objective function inspired by dynamic programming that avoids the need for any prior alignment between input and target sequences. The second approach is based on encoder-decoder neural networks with attention mechanisms. In these systems an input neural network, the encoder, models the acoustic input, generating an internal representation and another neural network, the decoder, generates the textual sequence from this internal representation (Chorowski et al., 2015; Chan et al., 2016). New architectures, in the form of transformers (Gulati et al., 2020; Huang et al., 2020; Wang et al., 2020) and teacher-student schemes (Zhang et al., 2020d; Liu et al., 2021a), have also been applied with great success to the ASR problem.

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 voice recordings. Two variants exist today: the speaker identification task, which identifies the speaker of an utterance from a known speaker set, and speaker verification, which determines if the speaker of an utterance matches the given enrolment. Classic speaker recognition techniques involve two steps: parameter extraction (first using mainly spectral magnitude parameters (Furui, 1981; De Leon et al., 2012) and later applying i-vectors (Dehak et al., 2010)) and classification itself (based on likelihood rates, primarily GMMs (Greenberg et al., 2014)). Nowadays, the classical systems have been outperformed by end-to-end neural networks based systems, which are being improved using widespread databases (Nagrani et al., 2017) and enforcing research (Nagrani et al., 2020), getting better recognition rates by means of new network architectures and techniques (Ding et al., 2020; Safari et al., 2020; Zhang et al., 2020c).

4.4 Machine Translation

Authors: Iakes Goenaga, Nora Aranberri, Gorka Labaka

Language can be considered the main means of communication for humans, a tool that allows us to present and express the ideas in our minds. There are over 6,500 languages in the world, which reflects the rich linguistic diversity of our cultures, but also points to the potential difficulty for people to understand one another, as such a large number of languages makes it impossible for an individual to master them all. To overcome this challenge, translation has long been used to convey meanings from one language to another.

MT is the automatic translation from one natural language into another using computers. Since its first implementation (Weaver, 1955) it has remained a key application in the field of Natural Language Processing (NLP).

While a number of approaches and architectures have been proposed and tested over the years, recently Neural Machine Translation (NMT) has become the most popular paradigm for MT development both within the research community (Bahdanau et al., 2015; Cho et al., 2014; Sutskever et al., 2014; Vaswani et al., 2018; Liu et al., 2020b; Zhu et al., 2020) and as large-scale production systems (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 (Cettolo et al., 2015; Ansari et al., 2020). NMT systems use distributed representations of the languages involved, which enables end-to-end training of the systems. And this is precisely one of the main reasons for their success. If we compare NMT systems with classical statistical machine translation models (Koehn et al., 2007; Callison-Burch et al., 2009b), 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; Yang et al., 2020; You et al., 2020; Zhang et al., 2020b).

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 systems (MNMT). The principal goal of an MNMT system is to translate between as many languages as possible by optimising the linguistic resources available. Multilingual NMT models (Aharoni et al., 2019; Bérard et al., 2020; Zhang et al., 2020a) are interesting for the research community for several reasons. On the one hand, they can address translations among all the different languages involved within a single model, which significantly reduces training time and facilitates deployment in 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., 2019a; Escolano et al., 2019; Hokamp et al., 2019). This phenomenon is known as translation knowledge transfer or transfer learning (Zoph et al., 2016; Nguyen and Chiang, 2017; Aji et al., 2020; Kocmi, 2020).

As mentioned, transfer learning has been regularly used for translation between low-resource languages that have few parallel corpora or other linguistic resources, but which can benefit from the linguistic resources of other languages. However, these systems have not as yet matched the results attained by bilingual models (Johnson et al., 2017b) because the model capacity must be split between different languages (Arivazhagan et al., 2019b). This challenge has been eased by increasing model capacity (Aharoni et al., 2019; Zhang et al., 2020a). Nevertheless, these models need to learn from even larger multilingual datasets, which are time-consuming and difficult to compile. To overcome this obstacle, or rather avoid it, most research thus far has focused primarily on English, the best resourced language, neglecting research for other language combinations.

A few efforts are emerging that aim to tackle the issue. To mention one, Fan et al. (2021) create 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 direct translation between 100 languages without using English as a pivot. Interestingly, it performs as well as bilingual models on many competitive benchmarks, including the WMT campaigns. In addition, they take advantage of backtranslation to improve the quality of their model on zero-shot and low-resource language pairs. Specifically, they create the first true many-to-many dataset by collecting 7.5B training sentences for 100 languages, facilitating direct training data for an extensive number of translation directions.

For resource-intensive language pairs, NMT systems have even claimed human parity in translation quality (Hassan et al., 2018; Toral et al., 2018). However, subsequent analyses have shown that this supposed parity was the consequence of certain features of the evaluation process, among them, that evaluators rated the translation quality at sentence level. In contrast, when the evaluation is done taking context into account, that is, by showing evaluators the whole document where the translated sentence belongs, machine translations lag behind human-generated translations (Läubli et al., 2018). This comes as no surprise given that NMT systems work at sentence level and, unlike human translators, do not consider linguistic phenomena that require a larger context when producing translations.

Therefore, in recent years, a new research line has attempted to extend the translation context to respond to this challenge. These newly proposed systems integrate the context directly into the model through different techniques, which can be divided into two main categories: those that modify the architecture of the neural network and those that only change the data that is input into the neural network.

Efforts that stem out of the first category modify the original neural network architecture by adding a context encoder. The context fed into the additional encoder usually consists of one or more sentences preceding the one to be translated (Voita et al., 2018). These architectures are limited in that they need parallel corpora with contextual information. To overcome this limitation, a two-step approach is applied, i.e. the new architecture is initialised with the parameters of a previously trained sentence-level system and then a fine-tuning step is performed using contextual information (Miculicich et al., 2018; Maruf et al., 2019; Yamagishi and Komachi, 2019). The modelling capabilities demonstrated by sentence-level attention mechanisms are also being explored for document-level translation. Jiang et al. (2019), for example, include Memory Networks in the NMT architecture to model inter-sentence attention, and thus, extract the most relevant discursive information in an extended context. This method has obtained significant improvements over robust Transformer-like systems.

The second modelling approach involves extending the information fed into the neural network without altering the neural network architecture (Tiedemann and Scherrer, 2017; Agrawal et al., 2018; Scherrer et al., 2019). This is mainly done by concatenating the sentence to be translated with the context. In Tiedemann and Scherrer (2017), although improvements in the automatic metrics are marginal, manual evaluation confirms that the system uses referential expressions between different sentences correctly.

4.5 Information Extraction and Information Retrieval

Authors: Aitor Soroa

Deep learning has had a tremendous impact on Information Retrieval (IR) and Information Extraction (IE). These are two of the oldest research topics in NLP, as early researchers realized the importance of retrieving and extracting structured information from textual sources.

The goal of IR is to meet the information needs of users by providing them with documents or text snippets that contain answers to a given query. IR is a mature technology that has allowed for the development of search engines worldwide. The area has been largely 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, 2021b; Yamada et al., 2021). Karpukhin et al. (2020) propose DPR (Dense Passage Retrieval), a method that relies on BERT (Devlin et al., 2019) to encode documents and queries into fixed-size representations, which are then queried using nearest neighbor techniques. One drawback of DPR is that it requires considerable memory due to the massive size of its passage index. To address this, Yamada et al. (2021) propose a method based on binary codes to represent the passage index in a compact way, which leads to a 97% reduction in the original size while maintaining good results. Dense representations are often combined with Question Answering techniques 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, 2021b,a; Yamada et al., 2021) or by generating the appropriate answers themselves (Lewis et al., 2021).

Information Extraction aims to derive structured information (often in the form of triplets) from text. Typically, IE systems recognize the main events described in a text, as well as the entities that participate in those events. Modern techniques on event extraction mostly focus on two central challenges: a) learning textual semantic representations for events in event extraction (both at sentence and document level) and b) acquiring or augmenting labeled instances for model training (Liu et al., 2020a). 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 various neural networks, including CNNs (Chen et al., 2015b), RNNs (Nguyen and Grishman, 2016) and Transformers (Yang et al., 2019). Data augmentation has been traditionally performed by using methods such as distant supervision or employing data from different languages to improve IE on the target language. The latter is especially useful when the target language does not have many resources. Deep learning techniques utilized in NMT (Wei et al., 2017; Liu et al., 2018) and pretrained multilingual LM models (Liu et al., 2019) have also helped in this task.

Another important task within IE is so-called Relation Extraction (RE), whose goal is to predict, if any, the semantic relationship between two entities. The best results to date on RE are obtained by fine-tuning large pretrained LMs, which are supplied with a classification head. Joshi et al. (2020) pretrain a LM by randomly masking contiguous spans of words, allowing it to learn to recognize span-boundaries and thus predict the masked spans. LUKE (Yamada et al., 2020) includes a pretraining phase to predict Wikipedia entities in text and uses entity information as an additional input. K-Adapter (Wang et al., 2021) freezes the parameters of a pretrained LM and utilizes Adapters to leverage factual knowledge from Wikipedia as well as syntactic information in the form of dependency parsing.

As with general IE, one of the most pressing problems in IE is the scarcity of manually annotated examples in real world applications, particularly when there is a domain and language shift. In such circumstances, the aforementioned methods perform poorly (Schick and Schütze, 2021a). In recent years, new methods have emerged that only require a few examples (few-shot) or no examples at all (zero-shot). *Prompt-based learning*, for instance, proposes to use task and label verbalizations that can be designed manually or learned automatically (Puri and Catanzaro, 2019; Schick and Schütze, 2021b,a) as an alternative to traditional fine-tuning (Gao et al., 2021; Le Scao and Rush, 2021). In these methods, the inputs are augmented with *prompts* and the LM objective is used in learning and inference. Brown et al. (2020) obtain good results by including the task descriptions along with input examples when pretraining a LM. In addition, (Schick and Schütze, 2021b,a; Tam et al., 2021) propose fine-tuning the prompt-based LMs on a variety of tasks.

4.6 Natural Language Generation and Summarization

Authors: German Rigau

Text generation, which is often formally referred as Natural Language Generation (NLG), has become one of the most important yet challenging tasks in NLP (Gehrmann et al., 2021). NLG is the task to automatically generate understandable texts, typically using a non-linguistic or textual representation of information as input (Reiter and Dale, 1997; Gatt and Krahmer, 2018; Li et al., 2021a). Example applications that generate new texts from existing (usually human-written) text include MT from one language to another (see subsection 4.4), fusion and summarization, simplification, text correction, paraphrases generation, question generation, etc. Often, however, it is necessary to generate texts which are not grounded in existing ones. With the recent resurgence of deep learning, various works have been proposed to solve text generation tasks based on different neural architectures (Li et al., 2021b). One of the advantages 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 of supervised text generation tasks are rather small (except MT). Therefore, researchers have proposed various methods to solve text generation tasks based on pretrained language models. Pretrained on large-scale corpus, these neural language models are able to encode massive linguistic and world knowledge accurately and express in human language fluently, both of which are critical abilities to fulfill text generation tasks. For text generation tasks, some of the pretrained language models utilize the standard Transformer architecture following the basic encoder-decoder framework, while others apply a decoder-only Transformer (see 3.2). Transformer models such T5 (Raffel et al., 2020) and BART (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.

With the rapid growth of enormous information generated each day on the internet, people are overwhelmed by this great amount of information (Gambhir and Gupta, 2017). Thus, summarizing techniques are becoming more and more popular and needed under the context of the information era for this task. A summary is the short version text produced from a single source or multiple sources while it conveys the main points of the original texts. The purpose of automatic text summarization is to create methods to produce this summary efficiently and precisely. Since the advent of text summarization in 1950s, researchers have been trying to improve techniques for generating summaries so that machine-generated summary matches with the human-made summary. Summaries can be generated through extractive as well as abstractive methods. An extractive method can be formulated as a sequence classification problem. Sequences are classified into two categories, summary sentence or non-summary sentence. This simple approach produces summaries in an extractive way. Several extractive approaches have been developed for automatic summary generation that implement a number of machine learning and optimization techniques (Xu and Durrett, 2019). Abstractive methods are more complex as they need natural language understanding capabilities. Abstractive sumarization produces an abstract summary which includes words and phrases different from the ones occurring in the source document. Therefore, an abstract is a summary that consists of ideas or concepts taken from the original document but are re-interpreted and shown in a different form (Du et al., 2021). Now, both approaches can be modeled using advanced Transformers (Liu and Lapata, 2019).

4.7 Human-Computer Interaction

Authors: Arantxa Otegi, Eneko Agirre

The demand for technologies that enable users to interact with machines at any time utilizing text and speech has grown, motivating the use of conversational systems known as Dialogue Systems. Such systems allow the user to converse with computers using natural language and include Siri,[54] Google Assistant,[55] and Amazon Alexa,[56] among others. Dialogue systems can be divided into three groups: task-oriented systems, conversational agents (better known as chatbots) and interactive question answering systems.

The distinguishing features of **task-oriented dialogue systems** are that they are oriented to perform a concrete task in a specific domain and their dialogue-flow is defined and struc-

[54] https://www.apple.com/es/siri/
[55] https://assistant.google.com/
[56] https://www.amazon.com

tured beforehand. For example, such systems are used to book a table at a restaurant, to call someone or check the weather forecast. The classical implementation of this type of system follows a pipeline architecture based on three modules.

The first one is the NLU module, whose aim is to identify the user intent and extract slots or concepts from the user utterance. The former objective is handled as a sentence classification task and employs different classification techniques, such as Support Vector Machines (Chelba et al., 2003) or neural network-based methods (Sarikaya et al., 2011). Slot extraction relies on sequence labelling approaches, exemplified by Conditional Random Field (Hahn et al., 2010; Wang et al., 2011) or RNN-based algorithms (biLSTM with CRF layer (Yao et al., 2014; Mesnil et al., 2015), for instance). More recently, different methods based on neural networks have been proposed to train a model for intent identification and slot extraction jointly (Mesnil et al., 2013; Xu and Sarikaya, 2013; Guo et al., 2014; Liu and Lane, 2016; Zhang and Wang, 2016; Goo et al., 2018). Schuster et al. (2019) present a multilingual dataset (English, Spanish and Thai) for slot extraction and use it to evaluate various cross-lingual transfer learning methods. Turning to MT and multilingual language models, López de Lacalle et al. (2020) propose two approaches for languages that have no training data for intent classification and slot extraction, through which they constructed a publicly available Basque dataset.

The next module, the **dialogue manager**, decides the dialogue policy, that is the next step to be taken by the agent (McTear et al., 2005; van Schooten et al., 2007). It analyzes whether the current information provided by the user is enough to finish the task, decides if additional information should be requested and offers the user several options. The dialogue manager relies on an ontology that describes the slots in the domain and on a set of dialogue acts that define the steps to be taken by the dialogue manager (Austin, 1962). Dialogue managers are generally implemented using manual rules or statistical approaches that learn from richly annotated dialogues (Levin et al., 1998; Young et al., 2013).

The third module performs the NLG and its objective is to generate the text of an answer for the user. Classical dialogue systems used rule-based methods or statistical language models based on phrases (Mairesse et al., 2010) or semantic trees (Dethlefs et al., 2013). Presently, algorithms based on neural networks are being proposed to focus on diverse issues in NLG: extend an LSTM to manage the semantics of an answer (Wen et al., 2015), extend a encoder-decoder architecture using a coarse-to-fine aligner to manage the content selection problem (Mei et al., 2016), exploit data counter fitting for the cases where there is insufficient in-domain training data available (Wen et al., 2016), train a variational autoencoder in an unsupervised way and use it to sample texts from the latent space (Bowman et al., 2016; Semeniuta et al., 2017), apply a sequence-to-sequence architecture with attention to generate deep syntax dependency trees in addition to text.

Classical dialogue systems used to train and evaluate these 3 modules separately. Alternatively, more recent systems rely on end-to-end trainable architectures based on neural networks (Zhao and Eskenazi, 2016; Bordes et al., 2017; Li et al., 2017; Wen et al., 2017).

The goal of **conversational agents** is to carry out engaging open-domain conversations, often by emulating the personality of a human (Zhang et al., 2018). The Alexa prize,[57] 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 such as encoder-decoders (Serban et al., 2017).

Interactive question answering systems try to respond to user questions by extracting answers from either documents (Rajpurkar et al., 2018) or knowledge bases (Yu et al., 2018b). In order to be able to have meaningful interactions, interactive question answering systems have a simple dialogue management procedure taking the previous questions and answers into account (Choi et al., 2018). The core technology is commonly based on pretrained lan-

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

guage models such as BERT (Devlin et al., 2019), where some mechanism is included to add context representation (Huang et al., 2019).

5 Domain Sectors

Natural language is the most common and versatile way for humans to convey information. We use language, our natural means of communication, to encode, store, transmit, share and manipulate information. In fact, most of the digital information available appears in the form of documents (written or spoken) in multiple languages, representing a challenge for any organization that wants to exploit and process its information. However, the language and background knowledge is different in the different domains of application. However, LT usually need of some kind of adaptation when they are used in specific domains such as the health, education, legal, finance, media, etc. (Ramponi and Plank, 2020) This section presents the state-of-the art in LT of some the most relevant domain sectors for LT namely, Health in section 5.1, Education in section 5.2 and Legal domain in section 5.3.

5.1 Health

Authors: Arantza Casillas, Aitziber Atutxa, Josu Goikoetxea, Koldo Gojenola, Maite Oronoz, Alicia Pérez, Olatz Perez de Viñaspre

In this section we introduce the resources available within the medical domain (corpora, embeddings and knowledge bases), followed by a review of relevant tasks and current trends in LT in the health domain.

We can distinguish between three main types of medical **corpora** depending on the original source they were obtained from: scientific reviews, clinical narratives and social media.

- **Scientific corpora:** many scientific articles and abstracts are freely available thanks to the PubMed portal of the National Library of Medicine (NLM). In addition, other initiatives such as HAL[58] and ISTEX,[59] HON,[60] CISMEF[61] and others provide generic portals for accessing medical and scientific publications. Some existing scientific corpora also offer annotations and categorizations, including PoS-tagging and negation. These are often built for the purposes of shared tasks.

- **Clinical corpora:** composed of Electronic Health Records (EHRs), these kinds of corpora are typically created in collaboration with Healthcare systems. For ethical reasons, even after a data anonymization process, it is rare to obtain permission to distribute this sort of medical data and it is therefore seldom freely available for research. The most well-known English corpus is perhaps the Medical Information Mart for Intensive Care (MIMIC) available from the Physionet portal.[62] Similarly, Informatics for Integrating Biology and the Bedside (i2b2)[63] is an NIH-funded initiative that promotes the development and testing of NLP tools for English-language documents. Several English corpora can be found at this portal. With respect to languages other than English, CLEF-eHEALTH challenges provide annotations for disorder detection and abbreviation normalization for various languages, including, English, French, Italian, German and others.

[58] https://hal.archives-ouvertes.fr/?lang=en
[59] https://www.istex.fr/
[60] https://www.hon.ch/en/
[61] https://www.cismef.org/cismef/
[62] https://www.i2b2.org/
[63] https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

- **Social media corpora:** one important source for biomedical and public health applications is the Social Media Mining for Health shared tasks (SMM4H). SMM4H task-related corpora are composed of different corpora depending on task and subtask. For example, Cadec, a corpus of adverse drug event annotations,[64] RuDREC corpus,[65] PsyTAR dataset, [66] and TwiMed corpus annotated medical Entities.[67]

The use of **embeddings** in the medical domain has escalated in recent years, leading to a wide variety of models. Distributional models, including word2vec (Mikolov et al., 2013b), Glove (Pennington et al., 2014) or FastText (Bojanowski et al., 2017), have been extensively utilized by the NLP community both for general purpose tasks and in the medical domain, improving state-of-the-art results. These are the so-called static embeddings.[68] Lately, dynamic embeddings such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018) have nearly replaced the former, enhancing the state of the art even further. Needless to say that in the biomedical domain these dynamic representations are currently widely used.

Regarding **static embeddings**, models that incorporate semantic and syntactic information from medical domain text corpora at the word level have been used to create embeddings at the word level (Moen and Ananiadou, 2013; Chiu et al., 2016; Liu et al., 2015; Zhao et al., 2018; Khattak et al., 2019) as well as at the subword level (Rei et al., 2016; Karmakar, 2018; Le et al., 2018; Zhang et al., 2019b). Note that the latter could be used to induce OOVs or rare words with very few appearances. Other authors have enriched the text embeddings with knowledge-based information (Yu et al., 2016b; Zhang et al., 2019b), thus combining the semantic information from the two mentioned sources into hybrid embeddings. Additionally, research has also focused on the conceptual level (CUIs), encoding the biomedical knowledge structure in a vector space (Vine et al., 2014; Choi et al., 2016; Beam et al., 2019). A final group of static embeddings are code embeddings, that is, distributional representations that encode medical codes such as diagnoses, procedures or drugs in a single vector space (Choi et al., 2016; Che et al., 2017; Cai et al., 2018).

Dynamic embeddings in the biomedical domain have also achieved state-of-the-art results with models such as BioBERT (Lee et al., 2019) and Bio_ClinicalBERT (Alsentzer et al., 2019) by focusing their pretraining process on domain-specific corpora (in English).

During language model pretraining, the representation of the words is learned from some given corpus. In this case, if a concept is missing from the corpus the language model will not be able to produce a meaningful representation. This problem is critical, especially in a low-resource setting like the clinical domain. Many **knowledge bases** are available today, each with a specific focus on medicine or tasks. The Unified Medical Language System[69] (UMLS) (Bodenreider, 2004) integrates more than 100 of those Knowledge Bases with mappings among them, so one can see UMLS as a whole. There are over 4.4M concepts and 16M concept names (or terms), most of them in English (70.88%). But another 25 languages, including Portuguese (2.64%), French (2.69%) and Spanish (9.94%), are also present with much smaller coverage. Of its over 150 sources, SNOMED CT (Donnelly et al., 2006) is UMLS's largest. The Standarized Nomenclature of Medicine - Clinical Terms[70] (SNOMED-CT) is a systematically organized computer processable collection of medical terms providing concepts, synonyms and relations between concepts. It is considered the most comprehensive multilingual clinical healthcare terminology in the world. The International Classification of Diseases[71] (ICD), also included in UMLS, is the reference term classification for death reasons,

64 https://github.com/gabrielStanovsky/CADEC-for-NLP
65 https://github.com/cimm-kzn/RuDReC
66 https://www.askapatient.com/
67 https://github.com/nestoralvaro/TwiMed
68 Each word has a unique distributional representation.
69 https://www.nlm.nih.gov/research/umls/index.html
70 https://www.snomed.org/
71 https://www.who.int/standards/classifications/classification-of-diseases

diagnostics and proceedings all over the world hospitals.

Having reviewed meaningful resources for the clinical domain, we shall now focus on relevant **tasks**. The basis of any further higher-level processing in medical NLP rests on NER and NER Classification (NERC). Even though these basic problems have been solved with scores over 90% for several languages (Lee et al., 2019; Kanakarajan et al., 2021), the mere recognition and classification of entities in a reduced set of classes (*diseases, drugs, symptoms, …*) is not enough, as medical texts can present a high orthographic variation, and therefore *Normalization* or *Entity Linking* is essential to accurately process medical texts. As an example, the disease "Type 2 diabetes mellitus" (standard name) can appear in multiple forms, including "Diabetes Mellitus 2", "Diab. Mel. II", "DM Type 2", "DM2", etc. In order to univocally refer to the same disease, concept identifiers from a medical ontology are needed. Indeed, in the aforementioned example, SNOMED-CT "C0011860" concept unifies all these variants into a single meaning. Given the considerable number of concepts (352,667 in SNOMED-CT), medical NER entails a significant challenge in practice, made all the more difficult by the scarcity of annotated data to automatically aid in the assignment of these identifiers.

Automatic coding classification tasks are next on our list of relevant tasks in the medical domain. Stanfill et al. (2010) carried out a systematic literature review of automated coding and **classification** clinical systems. Towards the classification of EHRs, a different shared task has been addressed for the codification of clinical documents with the International Classification of Diseases (ICD). Examples include: in 2018 CLEF (Névéol et al., 2018) for documents written in Italian, French and Hungarian; in 2019 for animal experiment (Kelly et al., 2019) summaries written in German; in 2020 the CodiEsp task at CLEF (Miranda-Escalada et al., 2020) for documents written in Spanish. Farkas and Szarvas (2008) presented an early state-of-the-art in ICD systems based on rules. The latest trends, in contrast, are generally based on language models (Silvestri et al., 2020; Velichkov et al., 2020) and integrate enhanced Machine Learning algorithms (Almagro et al., 2020; Blanco et al., 2020), although there are approaches that promote dictionary lookup as well (Cossin and Jouhet, 2020).

The best performing NLU systems are based on deep neural methods that have been criticized for their opaque nature. In this context, a new research stream is arising that encourages **explainable AI** (Nguyen-Duc et al., 2021). Explainable AI is opening new and relevant horizons, particularly in the development of clinical decision support systems. Medical professionals must be able to understand how and why a machine decision has been made (Holzinger et al., 2017). As London (2019) has indicated, "to the extent that deep learning systems cannot explain their findings, some have questioned whether medical systems should avoid such approache". To overcome this limit of successful "black box" neural architectures, efficient attention mechanisms have been used for continuous data monitoring (Xu et al., 2018b). The above-mentioned medical semantic lexicons (e.g., SNOMED-CT and UMLS) are excellent sources of knowledge. Faruqui et al. (2015) refine vector space representations using relational information from semantic lexicons (e.g., WordNet and FrameNet), an idea that is applied to the medical domain in Yu et al. (2016b). Holzinger et al. (2017) believe that a more promising approach in the medical domain "is the use of hybrid distributional models that combine sparse graph-based representations (Biemann and Riedl, 2013) with dense vector representations (Mikolov et al., 2013a) and link them to lexical resources and knowledge bases (Faralli et al., 2016)". Tjoa and Guan (2019) provide a review on interpretabilities suggested by various authors and categorize them in a section devoted to the medical domain. Similarly, in a separate review written by Mueller et al. (2019), DARPA asks, "what makes for a good explanation?". In the clinical domain, keeping the human in the loop tends to be a better choice than fully automated systems because there is a balance between facilitating the job of the practitioners and optimizing difficult decisions.

5.2 Education

Authors: Mikel Iruskieta, Jose Mari Arriola

Educational Data Mining and the analysis of the educational ecosystem in all European languages is necessary in order to develop a roadmap for achieving digital language equality in European education. Often, education is the first domain to foster an endangered language. This was the case for Irish, Basque, Catalan and Galician, among others. Many of these languages' speakers learn them in their educational environments and use them in their daily lives.

Conversely, these language learners generally encounter NLP language resources outside the educational system, especially when dealing with ICTs, ICALL systems and any digital broadcasting medium or social media. Unfortunately, many of these under-resourced languages do not possess sufficient resources to learn or monitor the language. They tend to be under-resourced in terms of technology and the data needed for AI. Indeed, under-resourced languages suffer from a chronic lack of available resources (human-, financial-, time-, data- and technology-wise), and from the fragmentation of efforts in resource development (Sayers et al., 2021). Their scarce resources are only usable for limited purposes, or are developed in isolation, without much connection with other resources and initiatives. The benefits of reusability, accessibility and data sustainability are often out of reach for such languages. Another challenging setting for technology is its use by minority languages communities. From a machine learning perspective, the shortage of digital infrastructure to support these languages may hamper development of appropriate technologies. Speakers of less widely-used languages may lag in access to the exciting resources that are coming. The consequences of this can be far-reaching, well beyond the technological domain: unavailability of a certain technology may lead speakers of a language to use another one, hastening the disappearance of their language altogether.

Moreover, under-resourced language curricula are rarely oriented towards the use of ICTs or the digital skills needed in language learning. As a result, teachers and students tend to be poorly prepared to employ digital technology when learning or teaching in such languages.[72] Developing these resources with an adequate pedagogical approach in bilingual communities where one language is not spoken by all citizens or used in every work environment would help foment language vitality and revitalization of all European languages, as well as help achieve digital language equality.

Under-resourced European languages must create better starting conditions for research as well as basic NLP-oriented toolkits. One example is Krauwer (2003), who presented the Basic Language Resource Kit (BLARK), part of the First Milestone for the Language Resources Roadmap, to do just this for "research, education and development in language and speech technology".

When we compare human to machine-based feedback, we find that the former is more accurate (Golke et al., 2015). However, educational LT-based tools are key to giving appropriate feedback when needed (Hattie and Timperley, 2007) and they are effective when the feedback is more explicit, since this can lead to a more successful uptake (Heift, 2004) and to more resubmissions that improve a learner's work (Heift, 2010). That is, uptake is even more effective when the feedback interacts with prior knowledge (Fyfe and Rittle-Johnson, 2016).

There are three trends at the moment in Educational Data Mining: 1) tools that provide statistics and visualization, 2) tools that provide feedback to teachers (diagnostic and prescriptive tools) 3) tools that provide recommendations to learners.

[72] UPgrading the SKIlls of Linguistics and Language Students: "The central goal of the UPSKILLS project is to identify and tackle the gaps in digital skills through the development of a new curriculum component and supporting the embedding of adequate materials in existing programmes."

Although there are many sources (also pedagogically oriented) that describe tools and their usage, it is difficult to find a top-rated list that indicates if they have been developed for a specific language or if they might be easily adaptable to others. In order to classify such tools, we will consider four types from the following three research fields: 1) Second Language Acquisition, 2) Tutoring Systems, and 3) Learning Analytics and Educational Data Mining.

1) Language learning environments: Moodle, Duolinguo, ICALL systems...
2) Corpus based tools: SpinTX, Korp, Ant, Sketch Engine.
3) ICT tools for language learning (Strobl et al., 2019) to help in writing skills: Academic Vocabulary, Article Writing Tool, AWSuM, C-SAW (Computer-Supported Argumentative Writing), Calliope, Carnegie Mellon prose style tool, CohVis, Corpuscript, Correct English (Vantage Learning), Criterion, De-Jargonizer, Deutsch-uni online, DicSci (Dictionary of Verbs in Science), Editor (Serenity Software), escribo, Essay Jack, Essay Map, Gingko, Grammark, Klinkende Taal, Lärka, Marking Mate (standard version), My Access!, Open Essay, Paper rater, PEG Writing, Rationale, RedacText, Research Writing Tutor, Right WriterSWAN (Scientific Writing Assistant), Scribo - Research Question and Literature Search Tool, StyleWriter, Thesis Writer, Turnitin (Revision Assistant), White Smoke, Write&Improve, WriteCheck, Writefull, WriteLab, Writer's Workbench, WriteToLearn, Writing Aid English, Writing Pal.
4) Language Analysis based tools: Markin, View, Complexity Schole, Grammarly, Text inspector, Readable, Reverso Speller and Feedbook, among others.

While it is a challenge for today's under-resourced languages, the use of tools that employ AI techniques and NLP-based systems Hernández-Blanco et al. (2019) are the basis of ICALLs and of the systems that will help students in the future. One limitation of these systems to take into account is that they are trained with texts that do not belong to the school environment. The type of text most frequently favoured is journalistic in nature due to its homogeneous characteristics: numerous texts are accessible, textual quality is acceptable and errors are minimal.

Although many additional tools and projects could be mentioned, Table 1 presents an overview of some that might be used to respond to previously diagnosed demands.

Another significant source for language learning are corpora. We can classify these in several ways:

- Learner raw written or spoken corpora.
- Learner analysed and findable corpora.
- Native multimedia corpora.
- Native multimedia corpora and findable corpora.[73]
- Interactive native multimedia corpora and findable corpora.[74]

Although there are many resources for many European languages, they are often difficult to find and do not always follow the same protocols. There are some interesting portals such as CLARIN resource families.[75] The CLARIN infrastructure, for example, provides access to 74 L2 learner corpora. Some of these are multilingual (11 corpora), while the rest are monolingual L2 data in 13 respective languages: Arabic, Czech, English, Finnish, French,

[73] https://www.clarin.eu/resource-families/L2-corpora and https://www.talkbank.org/
[74] https://www.coerll.utexas.edu/spintx/
[75] https://www.clarin.eu/resource-families/L2-corpora

Table 1: Projects for teaching language and writing, including language technologies and linguistic levels

Project	**Kind of feedback**	**Language level**	**Less NLP and less complexity**
Markin	Text correction: Manual system without NLP	All levels: manual	
VIEW	Language structure detection: Automatic with NLP	Morphosyntactic: automatic	
ReadLang			
FLAIR	Detector of textual complexity: (detection of level and linguistic structures) and recommender of more complex readings.	Characters, morphological, lexical and syntactic	
Grammarly	Detection of linguistic structures and automatic recommendations for writing: Use NLP for English and dependent on genre, level and textual genre	Selection of the linguistic level and type of text manually. Written text with personalized and automatic feedback	
Tagarela	Detection and partial correction of exercises: Portuguese ICALL with NLP Written text with personalized and automatic feedback		
Duolinguo	Personalized learning, immediate feedback,	Lexical and syntactic	
Feedbook	An Intelligent Language Tutoring System that was fully integrated as a homework platform into 14 regular 7th grade English classes in German secondary schools during a full-year study		
Project	**Kind of feedback**	**Language level**	**More NLP and more complexity**

German, Hungarian, Icelandic, Italian, Mandarin, Norwegian, Spanish, and Swedish. Despite the fact that many of these corpora are available through public licences, there are other European languages that do not have these data: Irish, Basque, Catalan, Galician... The TalkBank CLARIN-K centre[76] offers numerous additional languages obtained from researchers. Data in TalkBank use a consistent XML-compatible representation obtained from NLP analysis (not available for all languages) for automatic analysis and searching.

From the perspective of language ecology, almost all European resources, including corpora, LMS, ICALL systems and knowledge centres, are necessary for digital language equality. Data-driven research in this area can have a significant impact on society and help make the European Language Equality a reality. As for the future we also should have in mind the impact of AI on learning, teaching, and education (Tuomi, 2018). This policy foresight report suggests that in the next years AI will change learning, teaching, and education.

[76] https://www.talkbank.org/

5.3 Legal domain

Authors: German Rigau

Legal LT mainly focuses on applying LT to help legal tasks. The majority of the resources in this field are presented in text forms, such as judgment documents, contracts, and legal opinions (Zhong et al., 2020). Legal LT plays a significant role in the legal domain, as they can reduce heavy and redundant work for legal professionals. Many tasks in the legal domain require the expertise of legal practitioners and a thorough understanding of various legal documents. Retrieving and understanding legal documents take lots of time, even for legal professionals. Therefore, a qualified LT system should reduce the time consumption of these tedious jobs and benefit the legal system. Besides, LT can also provide a reliable reference to those who are not familiar with the legal domain, serving as an affordable form of legal aid.

In order to promote the development of legal LT, many researchers have devoted considerable efforts over the past few decades.[77] Early works (Kort, 1957; Ulmer, 1963; Segal, 1984), always use hand-crafted rules or features due to computational limitations at the time. In recent years, with rapid developments in deep learning, researchers begin to apply deep learning techniques to legal LT. Several new datasets have been proposed, which can serve as benchmarks for research in the field (Kano et al., 2018). Based on these datasets, researchers began exploring NLP-based solutions to a variety of legal tasks, such as Legal Judgment Prediction (Aletras et al., 2016; Luo et al., 2017; Chen et al., 2019a), Court View Generation (Ye et al., 2018), Legal Entity Recognition and Classification (Angelidis et al., 2018; Fernandes et al., 2020), Legal Question Answering (Kim and Goebel, 2017) or Legal Summarization (Bhattacharya et al., 2019). Lately, considerable efforts have been devoted to employing powerful pre-trained language models to promote the development of legal LT (Shaghaghian et al., 2020; Shao et al., 2020; Chalkidis et al., 2020; Xiao et al., 2021).

6 LT beyond Language

Authors: Gorka Azkune, Oier Lopez de Lacalle

Language is grounded in our physical world, as well as our societal and cultural context. Knowledge about our surrounding world is required to properly understand natural language utterances (Bender and Koller, 2020). That knowledge is known as commonsense knowledge and many authors argue that it is one of the key ingredients to achieve human-level NLU (Storks et al., 2019a). Following the irruption of deep learning methods (Salakhutdinov, 2014), new paradigms have been adopted and the field of NLU has advanced significantly in the last few years. However, many researchers in different application domains of deep learning have shown that those systems learn to find shortcuts to the correct answers through dataset-specific input-output correlations, essentially solving the dataset but not the underlying task. Examples of such can be found for dialogue generation (Li et al., 2016) and reading comprehension (Jia and Liang, 2017). One of the ways to acquire the necessary world knowledge to improve NLU is to explore the visual world together with the textual world (Elu et al., 2021).

With respect to multimodal and unimodal representation learning, CNNs have become the standard architecture for generating representations for images (LeCun et al., 1995). Most of these models learn transferable general image features in tasks such as image classification, detection, semantic segmentation and action recognition. The most utilized transferable global image representations are learned with deep CNN architectures such as AlexNet

[77] https://github.com/thunlp/CLAIM

(Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2015), Inception-v3 (Szegedy et al., 2016), and ResNet (He et al., 2016) using large datasets that include ImageNet (Deng et al., 2009), MSCOCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2017). Graph Convolution Networks (GCNs) appeared to be a promising way to distill multiple input types multimodal representations (Zhang et al., 2019a). 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 other approaches, Transformers allow multiple modalities to be processed (e.g., images, videos, text and speech) using similar processing blocks and demonstrate excellent scalability properties in sizable datasets. Language is mostly represented with pretrained word embeddings like Glove (Pennington et al., 2014) and sequence learning techniques such as RNNs (Hochreiter and Schmidhuber, 1997). Of late, Transformers have provided transferable models (Devlin et al., 2019; Radford et al., 2019) that significantly improve many state-of-the-art tasks in NLP. **Caption generation** is a typical visio-linguistic task, where given an image, a textual description of that image must be generated. The first approaches to solve this problem combined CNNs with RNNs in a encoder-decoder architecture (Vinyals et al., 2015). The CNN encoded the image, providing a high-level representation in the form of a dense vector. The RNN used that representation to generate the textual description. These architectures were trained end-to-end with paired images and captions, available in datasets such as MSCOCO (Lin et al., 2014) or Flickr30K (Plummer et al., 2015). Further improvements were achieved when attention was included in the encoder-decoder architecture (Xu et al., 2015). Some researchers proposed utilizing object-based attention instead of spatial attention (Anderson et al., 2018), paving the way for current multimodal transformers (Li et al., 2020), which also use object-based attention to feed a multimodal transformer that generates text for a given image. The quality of the text generated by these models is already high, as measured by automatic metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005).

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; Goyal et al., 2017; Johnson et al., 2017a). Some VQA datasets demand leveraging external knowledge to infer an answer and, thus, they are known as knowledge-based VQA tasks. Good examples are KB-VQA (Wang et al., 2017b), KVQA (Shah et al., 2019), FVQA (Wang et al., 2017a) and OK-VQA (Marino et al., 2019). All 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. Currently, multimodal transformers are the most successful systems for VQA and can be broadly classified into two types: single-stream and double-stream transformers. An example of the former is VisualBERT (Li et al., 2019a). In this case, the BERT architecture (Devlin et al., 2019) is utilized, adding visual features obtained by an object detector as input and using visio-linguistic pretraining tasks, such as image-text matching. OSCAR (Li et al., 2020) also follows a similar philosophy, applying object tags to the input and proposing different pretraining strategies. Among double-stream transformers, VilBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019) employ a dedicated transformer for each modality (text and image) to fuse them with a cross-modal transformer. Their differences lie mainly on some architectural choices and pretraining task selection.

Visual Referring Expressions are one of the multimodal tasks that may be considered an extension of a text only NLP task. More concretely, they are an extension of referring expressions (Krahmer and van Deemter, 2012) in natural language generation systems. The objective of the task is to ground a natural language expression (e.g., a noun phrase or a longer piece of text) to objects in a visual input.

Several methods have been proposed for 1) referring expression *generation*, in which an algorithm generates a referring expression for a given target object that is present in a visual scene, (Golland et al., 2010; Mitchell et al., 2013); 2) referring expression *comprehension*, where the referred object must be found in the image (Kazemzadeh et al., 2014); or 3) some combination of both (Mao et al., 2016; Yu et al., 2016a). Recent approaches use attention mechanisms to merge the textual and visual modalities (Yu et al., 2018a), as well as a combination of Gated Graph Convolutional Networks and Cross-Modal Relationship Extractors to highlight objects and relationships that have connections with a given referring expression through a multimodal structured relation graph (Yang et al., 2019).

In textual entailment, given a textual premise and a textual hypothesis, systems need to decide whether the first entails the second, they are in contradiction, or neither (Dagan et al., 2006; Bowman et al., 2015). As a natural extension of textual entailment, **Visual Entailment** is an inference task for predicting whether the image semantically entails the text. Vu et al. (2018) initially proposed a visually-grounded version of the textual entailment task, where an image is augmented to 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, Semantic Textual Similarity datasets (Cer et al., 2017) comprise pairs of sentences that have been annotated with similarity scores. Lopez de Lacalle et al. (2020) presented **Visual Semantic Textual Similarity** (vSTS), a task and dataset which allows to study whether better sentence representations can be built when having access to the corresponding images, in contrast to the text alone. Experiments using simple multimodal representations show that the addition of image representations produces better inference compared to text-only representations.

Presented as the opposite of caption generation, **visual 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. Inception score (Salimans et al., 2017) and Fréchet Inception Distance (Heusel et al., 2017) are frequently utilized, but as they have several problems, human evaluation is always included for assessing text-to-image systems. Recent studies have proposed a variety of models to generate an image given a sentence. Reed et al. (2016b) used a GAN (Goodfellow et al., 2014) that is conditioned on a text encoding for generating images of flowers and birds. Xu et al. (2018a) introduced a GAN-based image generation framework, where the image is progressively generated in two stages at increasing resolutions. Reed et al. (2016a) performed image generation with sentence input along with additional information in the form of keypoints or bounding boxes. Some (Hong et al., 2018; Li et al., 2019b) break down the process of generating an image from a sentence into multiple stages. The input sentence is first used to predict the entities that are presenting the scene, followed by the prediction of bounding boxes, then semantic segmentation masks, and finally the image. X-LXMERT (Cho et al., 2020) demonstrates that multimodal transformers can also generate state-of-the-art images from textual input. For that purpose, researchers sampled visual features for masked inputs and added an image generator to transform those sampled visual features into images. Following this trend, OpenAI recently presented DALL·E, a multimodal transformer decoder of over 1 billion parameters that achieves highly realistic results (Ramesh et al., 2021).

Multimodal Machine Translation (MMT), another popular task, aims 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 (Specia et al., 2016; Hitschler et al., 2016; Calixto et al., 2017c,b; 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., 2017a). Some view MMT as a two subtask problem of learning to translate and learning visually grounded representations combined in a multi-task learning framework (Elliott and Kádár, 2017). In a similar manner, a compact bilinear pooling method is proposed in (Delbrouck and Dupont, 2017), where the outer product of two vectors combines the attention features of the two modalities. Alternatively, Zhou et al. (2018) employed a shared visual-language embedding and a translator for learning a visual attention grounding mechanism that links the visual semantics with the corresponding textual semantics. Due to the recent success of unsupervised machine translation (Lample et al., 2018), there is also a growing interest in extending it for unsupervised MMT (Su et al., 2019).

7 Discussion

Language tools and resources have increased and improved since the end of the last century, a process further catalyzed by the advent of deep learning and neural networks 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 emphasized in this report, a good 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 worrisome. The unfortunate truth is that European digital language equality 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. A similar language discrepancy is also found in several of the domain sectors covered above: medical corpora, models and knowledge bases suffer from this disparity, as do users of under-resourced languages in education, where access to language-related tools is limited for most smaller language communities.

Yet, we believe this time of technological transition represents an opportunity to right the ship; that now is the moment to seek balance between European languages in the digital realm. There are ample reasons for optimism. Although there is more work that can and must be done, Europe's leading language resource repositories, platforms, libraries, models and benchmarks have begun to make inroads in this regard. 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.

8 Summary and Conclusions

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 (Wei et al., 2021). 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 of the LT tasks, NLU 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 pretrained language models 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, computing 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 businesses to the citizens.

References

Rodrigo Agerri and German Rigau. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63–82, 2016.

Rodrigo Agerri, Montse Cuadros, Seán Gaines, and German Rigau. Opener: Open polarity enhanced named entity recognition. *Procesamiento del Lenguaje Natural*, 51(0):215–218, 2013. ISSN 1989-7553. URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4891.

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. Give your text representation models some love: the case for Basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.588.

Eneko Agirre and Philip Edmonds, editors. *Word Sense Disambiguation: Algorithms and Applications*. Springer, 1 edition, 2006. ISBN 1402048084.

Ruchit Rajeshkumar Agrawal, Marco Turchi, and Matteo Negri. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *21st Annual Conference of the European Association for Machine Translation*, pages 11–20, 2018.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 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)*, pages 3874–3884, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388. URL https://aclanthology.org/N19-1388.

Nur Ahmed and Muntasir Wahed. The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*, 2020. URL https://arxiv.org/abs/2010.15581.

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

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.688. URL https://aclanthology.org/2020.acl-main.688.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-4010. URL https://aclanthology.org/N19-4010.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93, 2016.

Mario Almagro, Raquel Martínez Unanue, Víctor Fresno, and Soto Montalvo. Icd-10 coding of spanish electronic discharge summaries: An extreme classification problem. *IEEE Access*, 8:100073–100083, 2020. doi: 10.1109/ACCESS.2020.2997241.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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*, pages 6077–6086. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00636. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html.

Iosif Angelidis, Ilias Chalkidis, and Manolis Koubarakis. Named entity recognition, linking and generation for greek legislation. In *JURIX*, pages 1–10, 2018.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.1. URL https://aclanthology.org/2020.iwslt-1.1.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.279. URL https://doi.org/10.1109/ICCV.2015.279.

Sercan Ömer Arik, Mike Chrzanowski, Adam Coates, Gregory Frederick Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Y. Ng, Jonathan Raiman, Shubho Sengupta, and Mohammad Shoeybi. Deep voice: Real-time neural text-to-speech. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 195–204. PMLR, 2017. URL http://proceedings.mlr.press/v70/arik17a.html.

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

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*, 2019b. URL https://arxiv.org/abs/1907.05019.

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi: 10.1162/tacl_a_00288. URL https://aclanthology.org/Q19-1038.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1019. URL https://aclanthology.org/P19-1019.

John Langshaw Austin. *How to do things with words*. William James Lectures. Oxford University Press, 1962.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6402. URL https://aclanthology.org/W18-6402.

Andrew L Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. Clinical concept embeddings learned from massive sources of multimodal medical data. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*, pages 295–306. World Scientific, 2019.

Emily M. Bender and Alexander Koller. 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*, pages 5185–5198, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://aclanthology.org/2020.acl-main.463.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

Alexandre Bérard, Zae Myung Kim, Vassilina Nikoulina, Eunjeong Lucy Park, and Matthias Gallé. A multilingual neural machine translation model for biomedical data. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcovid19-2.16. URL https://aclanthology.org/2020.nlpcovid19-2.16.

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. A comparative study of summarization algorithms applied to legal case judgments. In *European Conference on Information Retrieval*, pages 413–428. Springer, 2019.

Chris Biemann and Martin Riedl. Text: Now in 2d! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95, 2013.

Alan W Black, Heiga Zen, and Keiichi Tokuda. Statistical parametric speech synthesis. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–1229. IEEE, 2007.

Alberto Blanco, Olatz Perez de Viñaspre, Alicia Pérez, and Arantza Casillas. Boosting icd multi-label classification of health records with contextual embeddings and label-granularity. *Computer Methods and Programs in Biomedicine*, 188:105264, 2020.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051. URL https://aclanthology.org/Q17-1010.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Kohd, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2021. URL https://arxiv.org/abs/2108.07258.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 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, 2017. URL https://openreview.net/forum?id=S1Bb3D5gg.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1002. URL https://aclanthology.org/K16-1002.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Xiangrui Cai, Jinyang Gao, Kee Yuan Ngiam, Beng Chin Ooi, Ying Zhang, and Xiaojie Yuan. Medical concept embedding with time-aware attention. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 3984–3990. ijcai.org, 2018. doi: 10.24963/ijcai.2018/554. URL https://doi.org/10.24963/ijcai.2018/554.

Iacer Calixto, Qun Liu, and Nick Campbell. 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)*, pages 1913–1924, Vancouver, Canada, 2017a. Association for Computational Linguistics. doi: 10.18653/v1/P17-1175. URL https://aclanthology.org/P17-1175.

Iacer Calixto, Daniel Stein, Evgeny Matusov, Sheila Castilho, and Andy Way. Human evaluation of multi-modal neural machine translation: a case study on e-commerce listing titles. In *The 6th Workshop on Vision and Language*, page 31, 2017b.

Iacer Calixto, Daniel Stein, Evgeny Matusov, Pintu Lohar, Sheila Castilho, and Andy Way. 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*, pages 637–643, 2017c.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, 2009a. Association for Computational Linguistics. URL https://aclanthology.org/W09-0401.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, 2009b. Association for Computational Linguistics. URL https://aclanthology.org/W09-0401.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=5k8F6UU39V.

Xavier Carreras and Lluís Màrquez. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 89–97, Boston, Massachusetts, USA, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-2412.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv:2103.12028*, 2021. URL https://arxiv.org/abs/2103.12028.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL https://aclanthology.org/S17-2001.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. The IWSLT 2015 evaluation campaign. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, Da Nang, Vietnam, 2015. URL https://aclanthology.org/2015.iwslt-evaluation.1.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert:"preparing the muppets for court". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2898–2904, 2020.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 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 2016, Shanghai, China, March 20-25, 2016*, pages 4960–4964. IEEE, 2016. doi: 10.1109/ICASSP.2016.7472621. URL https://doi.org/10.1109/ICASSP.2016.7472621.

Zhengping Che, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu. Boosting deep learning risk prediction with generative adversarial networks for electronic health records. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 787–792. IEEE, 2017.

C. Chelba, M. Mahajan, and A. Acero. Speech utterance classification. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I, 2003. doi: 10.1109/ICASSP.2003.1198772.

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. Charge-based prison term prediction with deep gating network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6362–6367, Hong Kong, China, 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1667. URL https://aclanthology.org/D19-1667.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015a. URL https://arxiv.org/abs/1504.00325.

Yuan-Jui Chen, Tao Tu, Cheng chieh Yeh, and Hung-Yi Lee. End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning. In *Proc. Interspeech 2019*, pages 2075–2079, 2019b.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 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)*, pages 167–176, Beijing, China, 2015b. Association for Computational Linguistics. doi: 10.3115/v1/P15-1017. URL https://aclanthology.org/P15-1017.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2922. URL https://aclanthology.org/W16-2922.

Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 4774–4778. IEEE, 2018. doi: 10.1109/ICASSP.2018.8462105. URL https://doi.org/10.1109/ICASSP.2018.8462105.

Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. 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)*, pages 8785–8805, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.707. URL https://aclanthology.org/2020.emnlp-main.707.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 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*, pages 103–111, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL https://aclanthology.org/W14-4012.

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

Youngduck Choi, Chill Yi-I Chiu, and David Sontag. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41, 2016.

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

Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 577–585, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/1068c6e4c8051cfd4e9ea8072e3189e2-Abstract.html.

Yu-An Chung, Yuxuan Wang, Wei-Ning Hsu, Yu Zhang, and R. J. Skerry-Ryan. Semi-supervised training for improving data efficiency in end-to-end speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 6940–6944. IEEE, 2019. doi: 10.1109/ICASSP.2019.8683862. URL https://doi.org/10.1109/ICASSP.2019.8683862.

Michael Collins. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637, 2003. doi: 10.1162/089120103322753356. URL https://aclanthology.org/J03-4003.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

Alexis Conneau and Douwe Kiela. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1269.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Sébastien Cossin and Vianney Jouhet. Iam at clef ehealth 2020: Concept annotation in spanish electronic health records. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, 2020.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190, Berlin, Heidelberg, 2006. Springer. ISBN 978-3-540-33428-6.

Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3079–3087, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/7137debd45ae4d0ab9aa953017286b20-Abstract.html.

Robert Dale. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118, 2021.

Franciska de Jong, Bente Maegaard, Darja Fišer, Dieter van Uytvanck, and Andreas Witt. Interoperability in an infrastructure enabling multidisciplinary research: The case of CLARIN. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3406–3413, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.417.

Phillip L De Leon, Michael Pucher, Junichi Yamagishi, Inma Hernaez, and Ibon Saratxaga. Evaluation of speaker verification security and detection of hmm-based synthetic speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2280–2290, 2012.

Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19 (4):788–798, 2010.

Jean-Benoit Delbrouck and Stéphane Dupont. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 910–919, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1095. URL https://aclanthology.org/D17-1095.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.292. URL https://aclanthology.org/2020.findings-emnlp.292.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Nina Dethlefs, Helen Hastie, Heriberto Cuayáhuitl, and Oliver Lemon. Conditional random fields for responsive surface realisation using global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1254–1263, Sofia, Bulgaria, 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1123.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Shaojin Ding, Tianlong Chen, Xinyu Gong, Weiwei Zha, and Zhangyang Wang. Autospeech: Neural architecture search for speaker recognition. In *INTERSPEECH*, pages 916–920, 2020.

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

Kevin Donnelly et al. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279, 2006.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. All nlp tasks are generation tasks: A general pretraining framework. *arXiv preprint arXiv:2103.10360*, 2021. URL https://arxiv.org/abs/2103.10360.

Desmond Elliott and Ákos Kádár. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141, Taipei, Taiwan, 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1014.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4718. URL https://aclanthology.org/W17-4718.

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

Carlos Escolano, Marta R Costa-jussà, and José AR Fonollosa. Towards interlingua neural machine translation. *arXiv preprint arXiv:1905.06831*, 2019. URL https://arxiv.org/abs/1905.06831.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of nlp leaderboard design. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, 2020.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48, 2021.

Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K Soong. Tts synthesis with bidirectional lstm based recurrent neural networks. In *Fifteenth annual conference of the international speech communication association*, 2014.

Stefano Faralli, Alexander Panchenko, Chris Biemann, and Simone P Ponzetto. Linked disambiguated distributional semantic networks. In *International Semantic Web Conference*, pages 56–64. Springer, 2016.

Richárd Farkas and György Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*, 9(3):1–9, 2008.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado, 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1184. URL https://aclanthology.org/N15-1184.

C. Fellbaum and G. Miller, editors. *Wordnet: An Electronic Lexical Database*. MIT Press, Cambridge (MA), 1998.

William Paulo Ducca Fernandes, Luiz José Schirmer Silva, Isabella Zalcberg Frajhof, Guilherme da Franca Couto Fernandes de Almeida, Carlos Nelson Konder, Rafael Barbosa Nasser, Gustavo Robichez de Carvalho, Simone Diniz Junqueira Barbosa, Hélio Côrtes Vieira Lopes, et al. Appellate court modifications extraction for portuguese. *Artificial Intelligence and Law*, 28(3):327–360, 2020.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.

Sadaoki Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):254–272, 1981.

Emily R Fyfe and Bethany Rittle-Johnson. Feedback both helps and hinders learning: The causal role of prior knowledge. *Journal of Educational Psychology*, 108(1):82, 2016.

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

Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66, 2017.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.295. URL https://aclanthology.org/2021.acl-long.295.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2501. URL https://aclanthology.org/W18-2501.

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

Sebastian Gehrmann, 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. 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)*, pages 96–120, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gem-1.10. URL https://aclanthology.org/2021.gem-1.10.

Matthew Gerber and Joyce Chai. Beyond NomBank: A study of implicit arguments for nominal predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1583–1592, Uppsala, Sweden, 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-1160.

Andrew Gibiansky, Sercan Ömer Arik, Gregory Frederick Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2962–2970, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/c59b469d724f7919b7d35514184fdc0f-Abstract.html.

Stefanie Golke, Tobias Dörfler, and Cordula Artelt. The impact of elaborated feedback on text comprehension within a computer-based assessment. *Learning and instruction*, 39:123–136, 2015.

Dave Golland, Percy Liang, and Dan Klein. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 410–419, Cambridge, MA, 2010. Association for Computational Linguistics. URL https://aclanthology.org/D10-1040.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2118. URL https://aclanthology.org/N18-2118.

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

Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*, 2021. URL https://arxiv.org/abs/2106.03193.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL https://doi.org/10.1109/CVPR.2017.670.

Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1764–1772. JMLR.org, 2014. URL http://proceedings.mlr.press/v32/graves14.html.

Craig S Greenberg, Désiré Bansé, George R Doddington, Daniel Garcia-Romero, John J Godfrey, Tomi Kinnunen, Alvin F Martin, Alan McCree, Mark Przybocki, and Douglas A Reynolds. The nist 2014 speaker recognition i-vector machine learning challenge. In *Odyssey: The Speaker and Language Recognition Workshop*, pages 224–230, 2014.

Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015. URL http://dx.doi.org/10.21437/Interspeech.2020-3015.

Daniel Guo, Gokhan Tur, Wen-tau Yih, and Geoffrey Zweig. Joint semantic utterance classification and slot filling with recursive neural networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, IEEE 2014, pages 554–559, South Lake Tahoe, California, USA, 2014. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2014/12/SLT2014-daniel.pdf.

Stefan Hahn, Marco Dinarelli, Christian Raymond, Fabrice Lefèvre, Patrick Lehen, Renato De Mori, Alessandro Moschitti, Hermann Ney, and Giuseppe Riccardi. Comparing Stochastic Approaches to Spoken Language Understanding in Multiple Languages. *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, 16:1569–1583, 2010. ISSN 1558-7916. URL https://hal.archives-ouvertes.fr/file/index/docid/746965/filename/plugin-05639034.pdf.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Liang Zhang, Wentao Han, Minlie Huang, et al. Pre-trained models: Past, present and future. *AI Open*, 2021.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*, 2018. URL https://arxiv.org/abs/1803.05567.

John Hattie and Helen Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Trude Heift. Corrective feedback and learner uptake in call. *ReCALL*, 16(2):416–431, 2004.

Trude Heift. Prompting in call: A longitudinal study of learner uptake. *The Modern Language Journal*, 94(2):198–216, 2010.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020.

Antonio Hernández-Blanco, Boris Herrera-Flores, David Tomás, and Borja Navarro-Colorado. A systematic review of deep learning approaches to educational data mining. *Complexity*, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

Lynette Hirschman. The evolution of evaluation: Lessons from the message understanding conferences. *Computer Speech & Language*, 12(4):281–305, 1998.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2399–2409, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1227. URL https://aclanthology.org/P16-1227.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Chris Hokamp, John Glover, and Demian Gholipour Ghalandari. Evaluating the supervised and zero-shot performance of multi-lingual translation models. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 209–217, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5319. URL https://aclanthology.org/W19-5319.

Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable AI systems for the medical domain? *CoRR*, abs/1712.09923, 2017. URL http://arxiv.org/abs/1712.09923.

Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7986–7994. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00833. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Hong_Inferring_Semantic_Layout_CVPR_2018_paper.html.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL https://aclanthology.org/P18-1031.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. Flowqa: Grasping flow in history for conversational machine comprehension. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=ByftGnR9KX.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2360. URL https://aclanthology.org/W16-2360.

Wenyong Huang, Wenchao Hu, Yu Ting Yeung, and Xiao Chen. Conv-Transformer Transducer: Low Latency, Low Frame Rate, Streamable End-to-End Speech Recognition. In *Proc. Interspeech 2020*, pages 5001–5005, 2020. doi: 10.21437/Interspeech.2020-2361. URL http://dx.doi.org/10.21437/Interspeech.2020-2361.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online, 2021a. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.74.

Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=NTEz-6wysdb.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL https://aclanthology.org/D17-1215.

Shu Jiang, Rui Wang, Zuchao Li, Masao Utiyama, Kehai Chen, Eiichiro Sumita, Hai Zhao, and Baoliang Lu. Document-level neural machine translation with inter-sentence attention. *arXiv preprint arXiv:1910.14528*, 2019. URL https://arxiv.org/abs/1910.14528.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 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*, pages 1988–1997. IEEE Computer Society, 2017a. doi: 10.1109/CVPR.2017.215. URL https://doi.org/10.1109/CVPR.2017.215.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017b. doi: 10.1162/tacl_a_00065. URL https://aclanthology.org/Q17-1024.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1588. URL https://aclanthology.org/D19-1588.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020. doi: 10.1162/tacl_a_00300. URL https://aclanthology.org/2020.tacl-1.5.

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

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. BioELECTRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.bionlp-1.16. URL https://aclanthology.org/2021.bionlp-1.16.

Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In *JSAI International Symposium on Artificial Intelligence*, pages 177–192. Springer, 2018.

Amitabha Karmakar. Classifying medical notes into standard disease codes using machine learning. *arXiv preprint arXiv:1802.00382*, 2018. URL https://arxiv.org/abs/1802.00382.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. URL https://aclanthology.org/D14-1086.

Liadh Kelly, Hanna Suominen, Lorraine Goeuriot, Mariana Neves, Evangelos Kanoulas, Dan Li, Leif Azzopardi, Rene Spijker, Guido Zuccon, Harrisen Scells, et al. Overview of the clef ehealth evaluation lab 2019. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 322–339. Springer, 2019.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey, 2021. URL https://arxiv.org/abs/2101.01169.

Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics: X*, 4:100057, 2019.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324.

Mi-Young Kim and Randy Goebel. Two-step cascaded textual entailment for legal bar exam question answering. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 283–290, 2017.

P. Kingsbury and M. Palmer. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain, May 2002.

Veysel Kocaman and David Talby. Spark nlp: Natural language understanding at scale. *Software Impacts*, 8:100058, 2021.

Tom Kocmi. Exploring benefits of transfer learning in neural machine translation. *arXiv preprint arXiv:2001.01622*, 2020. URL https://arxiv.org/abs/2001.01622.

Philipp Koehn, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, et al. Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. In *CLSP Summer Workshop Final Report WS-2006, Johns Hopkins University*, 2007.

Dan Kondratyuk and Milan Straka. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1279. URL https://aclanthology.org/D19-1279.

Fred Kort. Predicting supreme court decisions mathematically: A quantitative analysis of the "right to counsel" cases. *American Political Science Review*, 51(1):1–12, 1957.

Emiel Krahmer and Kees van Deemter. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218, 2012. doi: 10.1162/COLI_a_00088. URL https://aclanthology.org/J12-1006.

Steven Krauwer. The basic language resource kit (blark) as the first milestone for the language resources roadmap. In *Proceedings of SPECOM*, volume 2003, pages 8–15, 2003.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73, 2017. ISSN 0920-5691. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C. Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14881–14892, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/6804c9bca0a615bdb9374d00a9fcba59-Abstract.html.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1549. URL https://aclanthology.org/D18-1549.

Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1512. URL https://aclanthology.org/D18-1512.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.302.

Hoang-Quynh Le, Duy-Cat Can, Sinh T. Vu, Thanh Hai Dang, Mohammad Taher Pilehvar, and Nigel Collier. Large-scale exploration of neural relation classification architectures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2266–2277, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1250. URL https://aclanthology.org/D18-1250.

Teven Le Scao and Alexander Rush. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.208. URL https://aclanthology.org/2021.naacl-main.208.

Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1018. URL https://aclanthology.org/D17-1018.

E. Levin, R. Pieraccini, and W. Eckert. Using Markov decision process for learning dialogue strategies. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1 of *ICASSP*, pages 201–204, Seattle, WA, USA, 1998. doi: 10.1109/ICASSP.1998.674402.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. Paq: 65 million probably-asked questions and what you can do with them, 2021.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1127. URL https://aclanthology.org/D16-1127.

Junyi Li, Tianyi Tang, Gaole He, Jinhao Jiang, Xiaoxuan Hu, Puzhao Xie, Zhipeng Chen, Zhuohao Yu, Wayne Xin Zhao, and Ji-Rong Wen. 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*, pages 30–39, Online, 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.4. URL https://aclanthology.org/2021.acl-demo.4.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. Pretrained language model for text generation: A survey. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4492–4499. International Joint Conferences on Artificial Intelligence Organization, 2021b. doi: 10.24963/ijcai.2021/612. URL https://doi.org/10.24963/ijcai.2021/612. Survey Track.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019a. URL http://arxiv.org/abs/1908.03557.

Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12174–12182. Computer Vision Foundation / IEEE, 2019b. doi: 10.1109/CVPR.2019.01245. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Li_Object-Driven_Text-To-Image_Synthesis_via_Adversarial_Training_CVPR_2019_paper.html.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. End-to-end task-completion neural dialogue systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 733–743, Taipei, Taiwan, 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1074.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Zhen-Hua Ling, Shi-Yin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiao-Jun Qian, Helen M Meng, and Li Deng. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine*, 32(3):35–52, 2015.

Bing Liu and I. Lane. Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling. In *INTERSPEECH*, 2016.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. Event detection via gated multilingual attention mechanism. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *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*, pages 4865–4872. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16371.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 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)*, pages 738–748, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1068. URL https://aclanthology.org/D19-1068.

Jing Liu, Rupak Vignesh Swaminathan, Sree Hari Krishnan Parthasarathi, Chunchuan Lyu, Athanasios Mouchtaris, and Siegfried Kunzmann. Exploiting large-scale teacher-student training for on-device acoustic models. In *Proc. International Conference on Text, Speech and Dialogue (TSD)*, 2021a.

Kang Liu, Yubo Chen, Jian Liu, Xinyu Zuo, and Jun Zhao. Extracting events and their relations from texts: A survey on recent research progress and challenges. *AI Open*, 1:22–39, 2020a. ISSN 2666-6510. doi: https://doi.org/10.1016/j.aiopen.2021.02.004. URL https://www.sciencedirect.com/science/article/pii/S266665102100005X.

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

Yang Liu and Mirella Lapata. 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)*, pages 3730–3740, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1387. URL https://aclanthology.org/D19-1387.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020b. doi: 10.1162/tacl_a_00343. URL https://aclanthology.org/2020.tacl-1.47.

Yue Liu, Tao Ge, Kusum Mathews, Heng Ji, and Deborah McGuinness. Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. In *Proceedings of BioNLP 15*, pages 92–97, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3810. URL https://aclanthology.org/W15-3810.

Alex John London. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report*, 49(1):15–21, 2019.

Maddalen Lopez de Lacalle, Egoitz Laparra, Itziar Aldabe, and German Rigau. A multilingual predicate matrix. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2662–2668, Portorož, Slovenia, 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1423.

Maddalen López de Lacalle, Xabier Saralegi, and Iñaki San Vicente. Building a task-oriented dialog system for languages with no training data: the case for Basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2796–2802, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.340.

Oier Lopez de Lacalle, Ander Salaberria, Aitor Soroa, Gorka Azkune, and Eneko Agirre. 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*, 2020.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.

Zhiyun Lu, Dong Guo, Alireza Bagheri Garakani, Kuan Liu, Avner May, Aurélien Bellet, Linxi Fan, Michael Collins, Brian Kingsbury, Michael Picheny, and Fei Sha. A comparison between deep neural nets and kernel acoustic models for speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pages 5070–5074. IEEE, 2016. doi: 10.1109/ICASSP.2016.7472643. URL https://doi.org/10.1109/ICASSP.2016.7472643.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1289. URL https://aclanthology.org/D17-1289.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561, Uppsala, Sweden, 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-1157.

Christopher D. Manning. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 171–189, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-19400-9.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 11–20. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.9. URL https://doi.org/10.1109/CVPR.2016.9.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 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*, pages 3195–3204. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00331. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Marino_OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.html.

Judith A Markowitz. *Using speech recognition.* Prentice-Hall, Inc., 1995.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.645. URL https://aclanthology.org/2020.acl-main.645.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. Selective attention for context-aware 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)*, pages 3092–3102, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1313. URL https://aclanthology.org/N19-1313.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018. URL https://arxiv.org/abs/1806.08730.

Michael McTear, Ian O'Neill, Philip Hanna, and Xingkun Liu. Handling errors and determining confirmation strategies—an object-based approach. *Speech Communication*, 45(3):249–269, 2005. URL https://www.aclweb.org/anthology/N16-1086.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. What to talk about and how? selective generation using LSTMs with coarse-to-fine alignment. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 720–730, San Diego, California, 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1086. URL https://aclanthology.org/N16-1086.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Byj72udxe.

G. Mesnil, X. He, L. Deng, and Yoshua Bengio. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *INTERSPEECH*, 2013.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, and Geoffrey Zweig. Using Recurrent Neural Networks for Slot Filling in Spoken Language Understanding. *IEEE/ACM Trans. Audio, Speech and Language Processing*, 23(3):530–539, 2015. ISSN 2329-9290. URL http://www.iro.umontreal.ca/~lisa/pointeurs/taslp_RNNSLU_final_doubleColumn.pdf.

Yajie Miao, Mohammad Gowayyed, and Florian Metze. Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 167–174. IEEE, 2015.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1325. URL https://aclanthology.org/D18-1325.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a. URL https://arxiv.org/abs/1301.3781.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *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*, pages 3111–3119, 2013b. URL https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1008.

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

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40, 2021.

Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, 2020.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. Generating expressions that refer to visible objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1174–1184, Atlanta, Georgia, 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1137.

SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. Distributional semantics resources for biomedical text processing. *Proceedings of LBM*, pages 39–44, 2013.

Nelson Morgan. Deep and wide: Multiple layers in automatic speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):7–13, 2011.

Shane T. Mueller, Robert R. Hoffman, William J. Clancey, Abigail Emrey, and Gary Klein. Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *CoRR*, abs/1902.01876, 2019. URL http://arxiv.org/abs/1902.01876.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. In *Proc. Interspeech 2017*, pages 2616–2620, 2017.

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

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gůlçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1028. URL https://aclanthology.org/K16-1028.

Aurélie Névéol, Aude Robert, Francesco Grippo, Claire Morgand, Chiara Orsi, Laszlo Pelikan, Lionel Ramadier, Grégoire Rey, and Pierre Zweigenbaum. CLEF eHealth 2018 Multilingual Information Extraction Task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian. In *CLEF (Working Notes)*, pages 1–18, 2018.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 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*, pages 80–90, Online, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-demos.10.

Thien Huu Nguyen and Ralph Grishman. Modeling skip-grams for event detection with convolutional neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 886–891, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1085. URL https://aclanthology.org/D16-1085.

Toan Q. Nguyen and David Chiang. 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)*, pages 296–301, Taipei, Taiwan, 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-2050.

Thanh Nguyen-Duc, Natasha Mulligan, Gurdeep S Mannu, and Joao H Bettencourt-Silva. Deep ehr spotlight: a framework and mechanism to highlight events in electronic health records for explainable predictions. *arXiv preprint arXiv:2103.14161*, 2021. URL https://arxiv.org/abs/2103.14161.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441.

Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. A review of deep learning based speech synthesis. *Applied Sciences*, 9(19):4050, 2019.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks, 2021.

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

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. IEEE, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://doi.org/10.1109/ICASSP.2015.7178964.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

John R Pierce and John B Carroll. *Language and machines: Computers in translation and linguistics*. National Academy of Sciences/National Research Council, 1966.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: 2000-speaker neural text-to-speech. In *Proceedings of ICLR*, pages 214–217, 2018.

Wei Ping, Kainan Peng, and Jitong Chen. Clarinet: Parallel wave generation in end-to-end text-to-speech. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HklY120cYm.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.303. URL https://doi.org/10.1109/ICCV.2015.303.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, 2014. Association for Computational Linguistics. doi: 10.3115/v1/S14-2004. URL https://aclanthology.org/S14-2004.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea, 2012. Association for Computational Linguistics. URL https://aclanthology.org/W12-4501.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 3617–3621. IEEE, 2019. doi: 10.1109/ICASSP.2019.8683143. URL https://doi.org/10.1109/ICASSP.2019.8683143.

Raul Puri and Bryan Catanzaro. Zero-shot text classification with generative language models, 2019. URL https://arxiv.org/abs/1912.10165.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.14. URL https://aclanthology.org/2020.acl-demos.14.

Yao Qian, Yuchen Fan, Wenping Hu, and Frank K Soong. On the training aspects of deep neural network (dnn) for parametric tts synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3829–3833. IEEE, 2014.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26, 2020.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *Technical Report. Open AI.*, 2018.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 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)*, pages 784–789, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. URL https://arxiv.org/abs/2102.12092.

Alan Ramponi and Barbara Plank. Neural unsupervised domain adaptation in nlp—a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, 2020.

Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado, 2009. Association for Computational Linguistics. URL https://aclanthology.org/W09-1119.

Scott E. Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 217–225, 2016a. URL https://proceedings.neurips.cc/paper/2016/hash/a8f15eda80c50adb0e71943adc8015cf-Abstract.html.

Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1060–1069. JMLR.org, 2016b. URL http://proceedings.mlr.press/v48/reed16.html.

Georg Rehm, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajič, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiļjevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez-Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch. European language grid: An overview. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3366–3380, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.413.

Georg Rehm, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, Jose Manuel Gomez-Perez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlova, Dusan Varis, Lukas Kacena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Julija Melnika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals. European language grid: A joint platform for the European language technology community. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 221–230, Online, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-demos.26.

Marek Rei, Gamal Crichton, and Sampo Pyysalo. Attending to characters in neural sequence labeling models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 309–318, Osaka, Japan, 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1030.

Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3165–3174, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f63f65b503e22cb970527f23c9ad7db1-Abstract.html.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020. URL https://arxiv.org/abs/2006.04558.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1079. URL https://aclanthology.org/P18-1079.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1621. URL https://aclanthology.org/P19-1621.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54.

Rudolf Rosa, Ondřej Dušek, Tom Kocmi, David Mareček, Tomáš Musil, Patrícia Schmidtová, Dominik Jurko, Ondřej Bojar, Daniel Hrbek, David Košťák, Martina Kinská, Josef Doležal, and Klára Vosecká. Theaitre: Artificial intelligence to write a theatre play. In *Proceedings of AI4Narratives2020 workshop at IJCAI2020*, 2020.

J. Ruppenhofer, M. Ellsworth, M.R. Petruck, C.R. Johnson, and J. Sheffczyk. Framenet ii: Extended theory and practice. http://framenet.icsi.berkeley.edu/book/book.html, 2006.

Pooyan Safari, Miquel India, and Javier Hernando. Self-attention encoding and pooling for speaker recognition. In *INTERSPEECH*, pages 941–945, 2020.

Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. Statistical parametric speech synthesis incorporating generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1):84–96, 2017.

Ruslan Salakhutdinov. Deep learning. In Sofus A. Macskassy, Claudia Perlich, Jure Leskovec, Wei Wang, and Rayid Ghani, editors, *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, page 1973. ACM, 2014. doi: 10.1145/2623330.2630809. URL https://doi.org/10.1145/2623330.2630809.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=BJrFC6ceg.

R. Sarikaya, Geoffrey E. Hinton, and B. Ramabhadran. Deep belief nets for natural language call-routing. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5680–5683, 2011.

Dave Sayers, Rui Sousa-Silva, Sviatlana Höhn, Lule Ahmedi, Kais Allkivi-Metsoja, Dimitra Anastasiou, Lynne Beňuš, Štefan; Bowker, Eliot Bytyçi, Alejandro Catala, Anila Çepani, Sami Chacón-Beltrán, Rubén; Dadi, Fisnik Dalipi, Vladimir Despotovic, Agnieszka Doczekalska, Sebastian Drude, Robert Fort, Karën; Fuchs, Christian Galinski, Christian Galinski, Christian Galinski, Federico Gobbo, Tunga Gungor, Siwen Guo, Klaus Höckner, PetraLea Láncos, Tomer Libal, Tommi Jantunen, Dewi Jones, Blanka Klimova, EminErkan Korkmaz, Mirjam Sepesy Maučec, Miguel Melo, Fanny Meunier, Bettina Migge, Verginica Barbu Mititelu, Arianna Névéol, Aurélie; Rossi, Antonio Pareja-Lora, Aysel Sanchez-Stockhammer, C.; Şahin, Angela Soltan, Claudia Soria, Sarang Shaikh, Marco Turchi, Sule Yildirim Yayilgan, Maximino Bessa, Luciana Cabral, Matt Coler, Chaya Liebeskind, Ilan Kernerman, Rebekah Rousi, and Cynog Prys. The dawn of the human-machine era : A forecast of new and emerging language technologies. Technical report, LITHME project, 2021. URL http://urn.fi/URN:NBN:fi:jyu-202105183003.

Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. Analysing concatenation approaches to document-level NMT in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6506. URL https://aclanthology.org/D19-6506.

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online, 2021a. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.20.

Timo Schick and Hinrich Schütze. 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*, pages 2339–2352, Online, 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.185. URL https://aclanthology.org/2021.naacl-main.185.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 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)*, pages 3795–3805, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1380. URL https://aclanthology.org/N19-1380.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 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)*, pages 6490–6500, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.507. URL https://aclanthology.org/2021.acl-long.507.

Jeffrey A Segal. Predicting supreme court cases probabilistically: The search and seizure cases, 1962-1981. *American Political Science Review*, 78(4):891–900, 1984.

Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. A hybrid convolutional variational autoencoder for text generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 627–637, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1066. URL https://aclanthology.org/D17-1066.

Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron C. Courville. Multiresolution recurrent neural networks: An application to dialogue response generation. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3288–3294. AAAI Press, 2017. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14571.

Shohreh Shaghaghian, Luna Yue Feng, Borna Jafarpour, and Nicolai Pogrebnyakov. Customizing contextualized language models for legal document reviews. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 2139–2148. IEEE, 2020.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. KVQA: knowledge-aware visual question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8876–8884. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33018876. URL https://doi.org/10.1609/aaai.v33i01.33018876.

Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. BERT-PLI: modeling paragraph-level interactions for legal case retrieval. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3501–3507. ijcai.org, 2020. doi: 10.24963/ijcai.2020/484. URL https://doi.org/10.24963/ijcai.2020/484.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ-Skerrv Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 4779–4783. IEEE, 2018. doi: 10.1109/ICASSP.2018.8461368. URL https://doi.org/10.1109/ICASSP.2018.8461368.

Stefano Silvestri, Francesco Gargiulo, Mario Ciampi, and Giuseppe De Pietro. Exploit multilingual language model at scale for icd-10 clinical text classification. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–7, 2020. doi: 10.1109/ISCC50000.2020.9219640.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2wav: End-to-end speech synthesis. In *Proceedings of 5th International Conference on Learning Representations*, pages 1–6, 2017.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2346. URL https://aclanthology.org/W16-2346.

Mary H Stanfill, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6):646–651, 2010.

Gabriel Stanovsky and Ido Dagan. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2300–2305, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1252. URL https://aclanthology.org/D16-1252.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, pages 1–60, 2019a.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, 2019b. URL https://arxiv.org/abs/1904.01172.

Milan Straka. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-2020. URL https://aclanthology.org/K18-2020.

Carola Strobl, Emilie Ailhaud, Kalliopi Benetos, Ann Devitt, Otto Kruse, Antje Proske, and Christian Rapp. Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education*, 131:33–48, 2019.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://aclanthology.org/P19-1355.

Yuanhang Su, Kai Fan, Nguyen Bach, C.-C. Jay Kuo, and Fei Huang. Unsupervised multi-modal neural machine translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10482–10491. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.01073. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Su_Unsupervised_Multi-Modal_Neural_Machine_Translation_CVPR_2019_paper.html.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.308. URL https://doi.org/10.1109/CVPR.2016.308.

Derek Tam, Rakesh R Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. Improving and simplifying pattern exploiting training, 2021. URL https://arxiv.org/abs/2103.11955.

Hao Tan and Mohit Bansal. 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)*, pages 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514.

Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4811. URL https://aclanthology.org/W17-4811.

Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (XAI): towards medical XAI. *CoRR*, abs/1907.07374, 2019. URL http://arxiv.org/abs/1907.07374.

Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002. URL https://aclanthology.org/W02-2024.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, 2018.

Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavvaf, and Edward A Fox. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*, 2020. URL https://arxiv.org/abs/2003.01200.

Ilkka Tuomi. The impact of artificial intelligence on learning, teaching, and education. *Luxembourg: Publications Office of the European Union*, 2018. URL https://core.ac.uk/reader/162257140.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433.

S Sidney Ulmer. Quantitative analysis of judicial processes: Some practical and theoretical applications. *Law and Contemporary Problems*, 28(1):164–184, 1963.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics. URL https://aclanthology.org/S13-2001.

Boris van Schooten, Sophie Rosset, Olivier Galibert, Aurélien Max, Rieks op den Akker, and Gabriel Illouz. Handling speech input in the Ritel QA dialogue system. In *8th Annual Conference of the International Speech Communication Association*, INTERSPEECH 2007, pages 126–129, Antwerp, Belgium, 2007. URL https://www.isca-speech.org/archive/interspeech_2007/i07_0126.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Ashish Vaswani, 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. Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 193–199, Boston, MA, 2018. Association for Machine Translation in the Americas. URL https://aclanthology.org/W18-1819.

Boris Velichkov, Simeon Gerginov, Panayot Panayotov, Sylvia Vassileva, Gerasim Velchev, Ivan Koychev, and Svetla Boytcheva. Automatic icd-10 codes association to diagnosis: Bulgarian case. In *CSBio'20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics*, pages 46–53, 2020.

Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. Medical semantic similarity with a neural language model. In Jianzhong Li, Xiaoyang Sean Wang, Minos N. Garofalakis, Ian Soboroff, Torsten Suel, and Min Wang, editors, *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 1819–1822. ACM, 2014. doi: 10.1145/2661829.2661974. URL https://doi.org/10.1145/2661829.2661974.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 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*, pages 3156–3164. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298935. URL https://doi.org/10.1109/CVPR.2015.7298935.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1117. URL https://aclanthology.org/P18-1117.

Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, Marco Rospocher, and Roxane Segers. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, 110:60–85, 2016. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2016.07.013. URL https://www.sciencedirect.com/science/article/pii/S0950705116302271.

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

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275, 2019a. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019b. URL https://openreview.net/forum?id=rJ4km2R5t7.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017a.

Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. Explicit knowledge-based reasoning for visual question answering. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1290–1296. ijcai.org, 2017b. doi: 10.24963/ijcai.2017/179. URL https://doi.org/10.24963/ijcai.2017/179.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.121. URL https://aclanthology.org/2021.findings-acl.121.

Ye-Yi Wang, Li Deng, and Alex Acero. *Semantic Frame-Based Spoken Language Understanding*, chapter 3, pages 41–91. John Wiley and Sons, Ltd, 2011. ISBN 9781119992691. doi: https://doi.org/10.1002/9781119992691.ch3. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119992691.ch3.

Yongqiang Wang, Abdelrahman Mohamed, Duc Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, Christian Fuegen, Geoffrey Zweig, and Michael L. Seltzer. Transformer-based acoustic modeling for hybrid speech recognition. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 6874–6878. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9054345. URL https://doi.org/10.1109/ICASSP40776.2020.9054345.

Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech 2017*, pages 4006–4010, 2017c.

Warren Weaver. Translation. *Machine translation of languages*, 14(15-23):10, 1955.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2021. URL https://arxiv.org/abs/2109.01652.

Sam Wei, Igor Korostil, Joel Nothman, and Ben Hachey. English event detection with translated language features. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 293–298, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2046. URL https://aclanthology.org/P17-2046.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1199. URL https://aclanthology.org/D15-1199.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. Multi-domain neural network language generation for spoken dialogue systems. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 120–129, San Diego, California, 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1015. URL https://aclanthology.org/N16-1015.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain, 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1042.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. URL https://arxiv.org/abs/1910.03771.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. URL https://arxiv.org/abs/1609.08144.

Zhizheng Wu and Simon King. Improving trajectory modelling for dnn-based speech synthesis by using stacked bottleneck features and minimum generation error training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(7):1255–1265, 2016.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2021.

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

Jiacheng Xu and Greg Durrett. 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)*, pages 3292–3303, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1324. URL https://aclanthology.org/D19-1324.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/xuc15.html.

Puyang Xu and Ruhi Sarikaya. Convolutional neural network based triangular CRF for joint intent detection and slot filling. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 78–83, 2013. doi: 10.1109/ASRU.2013.6707709.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1316–1324. IEEE Computer Society, 2018a. doi: 10.1109/CVPR.2018.00143. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Xu_AttnGAN_Fine-Grained_Text_CVPR_2018_paper.html.

Yanbo Xu, Siddharth Biswal, Shriprasad R. Deshpande, Kevin O. Maher, and Jimeng Sun. RAIM: recurrent attentive and intensive model of multimodal patient monitoring data. In Yike Guo and Faisal Farooq, editors, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 2565–2573. ACM, 2018b. doi: 10.1145/3219819.3220051. URL https://doi.org/10.1145/3219819.3220051.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.523. URL https://aclanthology.org/2020.emnlp-main.523.

Ikuya Yamada, Akari Asai, and Hannaneh Hajishirzi. Efficient passage retrieval with hashing for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 979–986, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.123. URL https://aclanthology.org/2021.acl-short.123.

Hayahide Yamagishi and Mamoru Komachi. Improving context-aware neural machine translation with target-side context. In *International Conference of the Pacific Association for Computational Linguistics*, pages 112–122. Springer, 2019.

Mingming Yang, Min Zhang, Kehai Chen, Rui Wang, and Tiejun Zhao. Neural machine translation with target-attention model. *IEICE TRANSACTIONS on Information and Systems*, 103(3):684–694, 2020.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1522. URL https://aclanthology.org/P19-1522.

Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, Geoffrey Zweig, and Yangyang Shi. Spoken language understanding using long short-term memory neural networks. In *Spoken Language Technology Workshop (SLT)*, IEEE 2014, pages 189–194, South Lake Tahoe, NV, USA, 2014. IEEE. doi: 10.1109/SLT.2014.7078572. URL https://ieeexplore.ieee.org/document/7078572.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1168. URL https://aclanthology.org/N18-1168.

Weiqiu You, Simeng Sun, and Mohit Iyyer. Hard-coded Gaussian attention for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7689–7700, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.687. URL https://aclanthology.org/2020.acl-main.687.

S. Young, M. Gašić, B. Thomson, and J. D. Williams. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013. ISSN 0018-9219. doi: 10.1109/JPROC.2012.2225812.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4584–4593. IEEE Computer Society, 2016a. doi: 10.1109/CVPR.2016.496. URL https://doi.org/10.1109/CVPR.2016.496.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. 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*, pages 1307–1315. IEEE Computer Society, 2018a. doi: 10.1109/CVPR.2018.00142. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Yu_MAttNet_Modular_Attention_CVPR_2018_paper.html.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 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*, pages 3911–3921, Brussels, Belgium, 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1425. URL https://aclanthology.org/D18-1425.

Zhiguo Yu, Trevor Cohen, Byron Wallace, Elmer Bernstam, and Todd Johnson. Retrofitting word vectors of MeSH terms to improve semantic similarity measures. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 43–51, Auxtin, TX, 2016b. Association for Computational Linguistics. doi: 10.18653/v1/W16-6106. URL https://aclanthology.org/W16-6106.

Heiga Ze, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing*, pages 7962–7966. IEEE, 2013.

Heiga Zen, Keiichi Tokuda, and Alan W Black. Statistical parametric speech synthesis. *speech communication*, 51(11):1039–1064, 2009.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online, 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.148. URL https://aclanthology.org/2020.acl-main.148.

Biao Zhang, Deyi Xiong, Jun Xie, and Jinsong Su. Neural machine translation with gru-gated attention model. *IEEE transactions on neural networks and learning systems*, 31(11):4688–4698, 2020b.

Hanyi Zhang, Longbiao Wang, Yunchun Zhang, Meng Liu, Kong Aik Lee, and Jianguo Wei. Adversarial separation network for speaker recognition. In *INTERSPEECH*, pages 951–955, 2020c.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 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)*, pages 2204–2213, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205.

Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):11, 2019a.

Xiaodong Zhang and Houfeng Wang. A joint model of intent determination and slot filling for spoken language understanding. In Subbarao Kambhampati, editor, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2993–2999. IJCAI/AAAI Press, 2016. URL http://www.ijcai.org/Abstract/16/425.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9, 2019b.

Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, R.J. Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. In *Proc. Interspeech 2019*, pages 2080–2084, 2019c.

Ziqiang Zhang, Yan Song, Jian shu Zhang, Ian McLoughlin, and Li-Rong Dai. Semi-Supervised End-to-End ASR via Teacher-Student Learning with Conditional Posterior Distribution. In *Proc. Interspeech 2020*, pages 3580–3584, 2020d. doi: 10.21437/Interspeech.2020-1574. URL http://dx.doi.org/10.21437/Interspeech.2020-1574.

Mengnan Zhao, Aaron J. Masino, and Christopher C. Yang. A framework for developing and evaluating word embeddings of drug-named entity. In *Proceedings of the BioNLP 2018 workshop*, pages 156–160, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2319. URL https://aclanthology.org/W18-2319.

Tiancheng Zhao and Maxine Eskenazi. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–10, Los Angeles, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3601. URL https://aclanthology.org/W16-3601.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.466. URL https://aclanthology.org/2020.acl-main.466.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1400. URL https://aclanthology.org/D18-1400.

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

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1163. URL https://aclanthology.org/D16-1163.